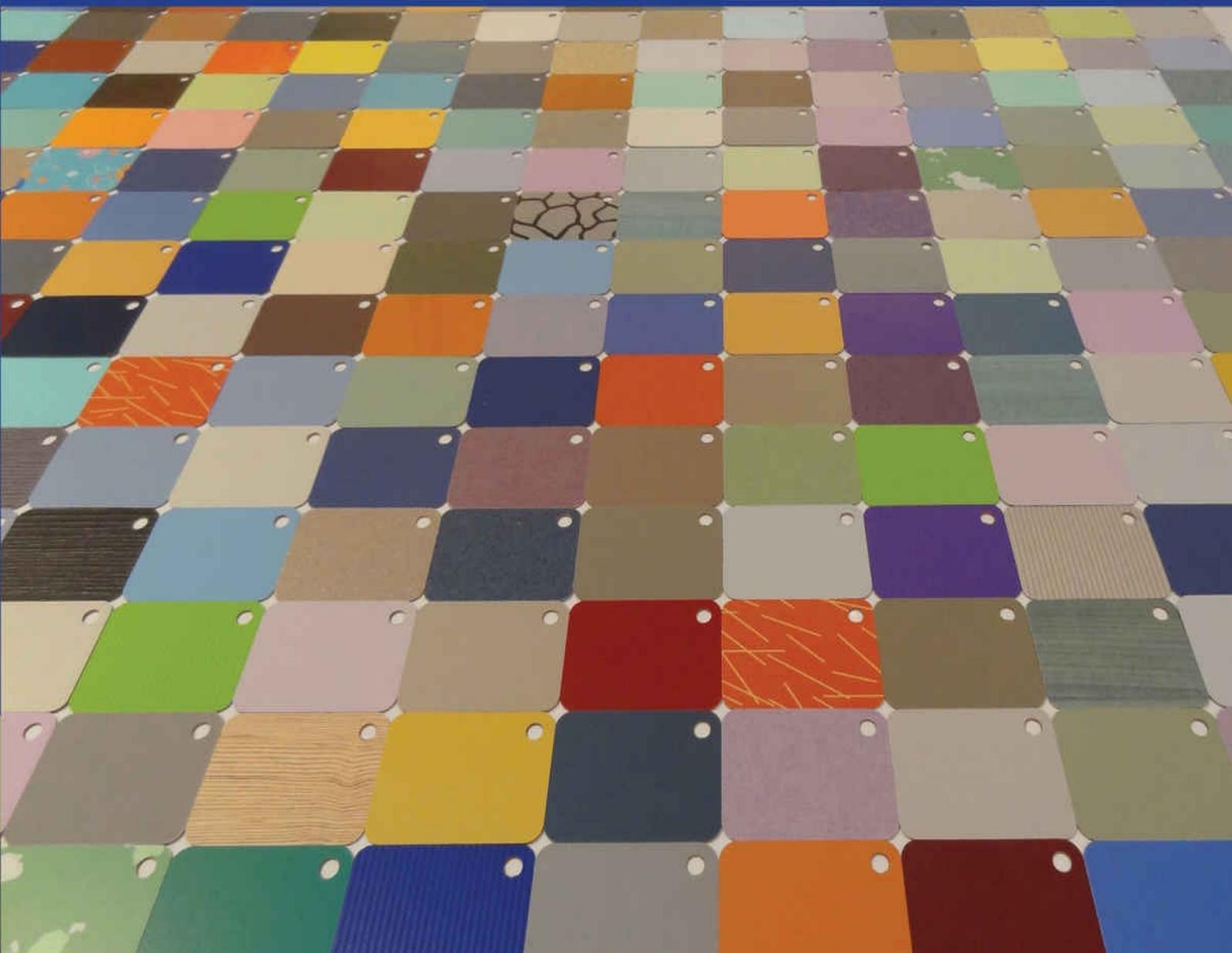


# INTRODUCTION to the **PRACTICE** of **STATISTICS**

EIGHTH EDITION

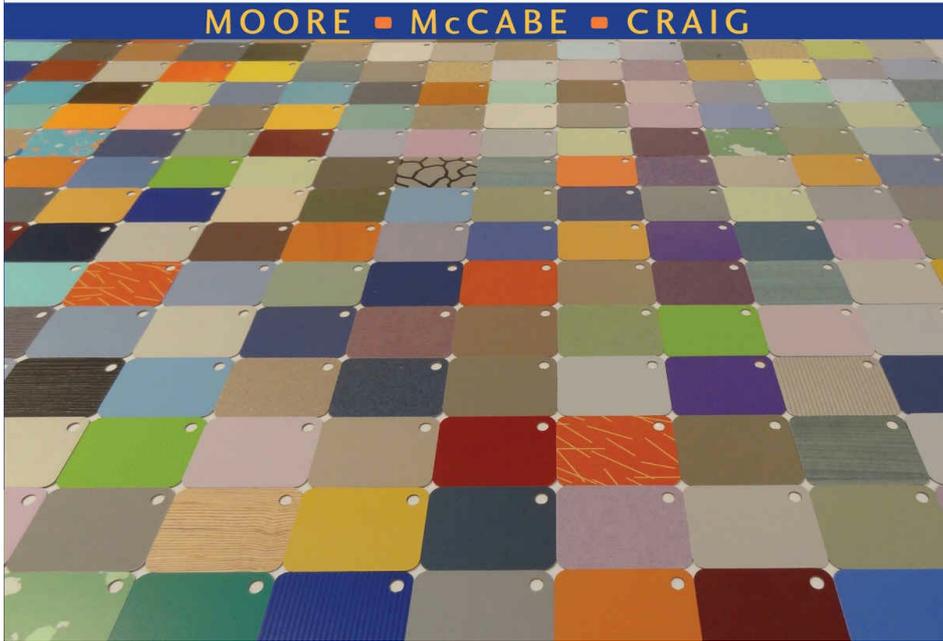
MOORE ■ McCABE ■ CRAIG



INTRODUCTION  
to the **PRACTICE** of  
**STATISTICS**

EIGHTH EDITION

MOORE ■ McCABE ■ CRAIG



**INTRODUCTION**  
the **PRACTICE** of  
**STATISTICS**

EIGHTH EDITION

**David S. Moore**  
**George P. McCabe**  
**Bruce A. Craig**  
*Purdue University*



W. H. Freeman and Company  
A Macmillan Higher Education Company

---

**SENIOR PUBLISHER: Ruth Baruth**  
**ACQUISITIONS EDITOR: Karen Carson**  
**MARKETING MANAGER: Steve Thomas**  
**DEVELOPMENTAL EDITOR: Katrina Wilhelm**  
**SENIOR MEDIA EDITOR: Laura Judge**  
**MEDIA EDITOR: Catriona Kaplan**  
**ASSOCIATE EDITOR: Jorge Amaral**  
**ASSISTANT MEDIA EDITOR: Liam Ferguson**  
**PHOTO EDITOR: Cecilia Varas**  
**PHOTO RESEARCHER: Dena Digilio Betz**  
**COVER DESIGNER: Victoria Tomaselli**  
**Text Designer: Patrice Sheridan**  
**PROJECT EDITOR: Elizabeth Geller**  
**ILLUSTRATIONS: Aptara<sup>®</sup>, Inc.**  
**PRODUCTION COORDINATOR: Lawrence Guerra**  
**COMPOSITION: Aptara<sup>®</sup>, Inc.**  
**PRINTING AND BINDING: QuadGraphics**

Library of Congress Control Number: 2013953337

Student Edition Hardcover (packaged with EESEE/CrunchIt! access card):

ISBN-13: 978-1-4641-5893-3

ISBN-10: 1-4641-5893-2

Student Edition Looseleaf (packaged with EESEE/CrunchIt! access card):

ISBN-13: 978-1-4641-5897-1

ISBN-10: 1-4641-5897-5

Instructor Complimentary Copy:

ISBN-13: 978-1-4641-3338-1

ISBN-10: 1-4641-3338-7

© 2014, 2012, 2009, 2006 by W. H. Freeman and Company

All rights reserved

Printed in the United States of America

First printing

W. H. Freeman and Company  
41 Madison Avenue  
New York, NY 10010  
Houndmills, Basingstoke RG21 6XS, England  
[www.whfreeman.com](http://www.whfreeman.com)

## BRIEF CONTENTS

---

To Teachers: About This Book

To Students: What Is Statistics?

About the Authors

Data Table Index

Beyond the Basics Index

**PART I**

**Looking at Data**

CHAPTER 1 **Looking at Data—Distributions**

CHAPTER 2 **Looking at Data—Relationships**

CHAPTER 3 **Producing Data**

**PART II** **Probability and Inference**

CHAPTER 4 **Probability: The Study of Randomness**

CHAPTER 5 **Sampling Distributions**

CHAPTER 6 **Introduction to Inference**

CHAPTER 7 **Inference for Distributions**

CHAPTER 8 **Inference for Proportions**

**PART III** **Topics in Inference**

CHAPTER 9 **Analysis of Two-Way Tables**

CHAPTER 10 **Inference for Regression**

CHAPTER 11 **Multiple Regression**

CHAPTER 12 **One-Way Analysis of Variance**

CHAPTER 13 **Two-Way Analysis of Variance**

**Companion Chapters** (on the IPS website  
[www.whfreeman.com/ips8e](http://www.whfreeman.com/ips8e))

CHAPTER 14 **Logistic Regression**

CHAPTER 15 **Nonparametric Tests**

CHAPTER 16 **Bootstrap Methods and Permutation Tests**

CHAPTER 17 **Statistics for Quality: Control and  
Capability**

Tables

Answers to Odd-Numbered Exercises

Notes and Data Sources

Photo Credits

Index

# CONTENTS

---

To Teachers: About This Book  
To Students: What Is Statistics?  
About the Authors  
Data Table Index  
Beyond the Basics Index

## **PART I** Looking at Data

CHAPTER 1  
**Looking at Data—Distributions**

## Introduction

### 1.1 Data

Key characteristics of a data set

Section 1.1 Summary

Section 1.1 Exercises

### 1.2 Displaying Distributions with Graphs

Categorical variables: bar graphs and pie charts

Quantitative variables: stemplots

Histograms

Data analysis in action: Don't hang up on me

Examining distributions

Dealing with outliers

Time plots

Section 1.2 Summary

Section 1.2 Exercises

### 1.3 Describing Distributions with Numbers

Measuring center: the mean

Measuring center: the median

Mean versus median

Measuring spread: the quartiles

The five-number summary and boxplots

The  $1.5 \times IQR$  rule for suspected outliers

Measuring spread: the standard deviation

Properties of the standard deviation

Choosing measures of center and spread

Changing the unit of measurement

Section 1.3 Summary

Section 1.3 Exercises

### 1.4 Density Curves and Normal Distributions

Density curves

Measuring center and spread for density curves

Normal distributions

The 68–95–99.7 rule

Standardizing observations

Normal distribution calculations

Using the standard Normal table

Inverse Normal calculations

Normal quantile plots

Beyond the Basics: Density estimation

Section 1.4 Summary

Section 1.4 Exercises

Chapter 1 Exercises

CHAPTER 2

**Looking at Data—Relationships**

## 2.1 Relationships

Examining relationships

Section 2.1 Summary

Section 2.1 Exercises

## 2.2 Scatterplots

Interpreting scatterplots

The log transformation

Adding categorical variables to scatterplots

Beyond the Basics: Scatterplot smoothers

Categorical explanatory variables

Section 2.2 Summary

Section 2.2 Exercises

## 2.3 Correlation

The correlation  $r$

Properties of correlation

Section 2.3 Summary

Section 2.3 Exercises

## 2.4 Least-Squares Regression

Fitting a Line to Data

Prediction

Least-squares regression

Interpreting the regression line

Facts about least-squares regression

Correlation and regression

Another view of  $r^2$

Section 2.4 Summary

Section 2.4 Exercises

## 2.5 Cautions about Correlation and Regression

Residuals

Outliers and influential observations

Beware of the lurking variable

Beware of correlations based on averaged data

Beware of restricted ranges

Beyond the Basics: Data mining

Section 2.5 Summary

Section 2.5 Exercises

## **2.6 Data Analysis for Two-Way Tables**

The two-way table

Joint distribution

Marginal distributions

Describing relations in two-way tables

Conditional distributions

Simpson's paradox

Section 2.6 Summary

Section 2.6 Exercises

## **2.7 The Question of Causation**

Explaining association

Establishing causation

Section 2.7 Summary

Section 2.7 Exercises

Chapter 2 Exercises

## **CHAPTER 3**

### **Producing Data**

## **Introduction**

### **3.1 Sources of Data**

- Anecdotal data
- Available data
- Sample surveys and experiments

[Section 3.1 Summary](#)

[Section 3.1 Exercises](#)

### **3.2 Design of Experiments**

- Comparative experiments
- Randomization
- Randomized comparative experiments
- How to randomize
- Cautions about experimentation
- Matched pairs designs
- Block designs

[Section 3.2 Summary](#)

[Section 3.2 Exercises](#)

### **3.3 Sampling Design**

- Simple random samples
- Stratified random samples
- Multistage random samples
- Cautions about sample surveys

[Section 3.3 Summary](#)

[Section 3.3 Exercises](#)

### **3.4 Toward Statistical Inference**

- Sampling variability
- Sampling distributions
- Bias and variability
- Sampling from large populations
- Why randomize?
- Beyond the Basics: Capture-recapture sampling

[Section 3.4 Summary](#)

[Section 3.4 Exercises](#)

### **3.5 Ethics**

- Institutional review boards
- Informed consent

Confidentiality

Clinical trials

Behavioral and social science experiments

Section 3.5 Summary

Section 3.5 Exercises

Chapter 3 Exercises

## **PART II** **Probability and Inference**

CHAPTER 4

**Probability: The Study of Randomness**

## **Introduction**

### **4.1 Randomness**

- The language of probability
- Thinking about randomness
- The uses of probability

Section 4.1 Summary

Section 4.1 Exercises

### **4.2 Probability Models**

- Sample spaces
- Probability rules
- Assigning probabilities: finite number of outcomes
- Assigning probabilities: equally likely outcomes
- Independence and the multiplication rule
- Applying the probability rules

Section 4.2 Summary

Section 4.2 Exercises

### **4.3 Random Variables**

- Discrete random variables
- Continuous random variables
- Normal distributions as probability distributions

Section 4.3 Summary

Section 4.3 Exercises

### **4.4 Means and Variances of Random Variables**

- The mean of a random variable
- Statistical estimation and the law of large numbers
- Thinking about the law of large numbers

Beyond the Basics: More laws of large numbers

- Rules for means
- The variance of a random variable
- Rules for variances and standard deviations

Section 4.4 Summary

Section 4.4 Exercises

### **4.5 General Probability Rules**

- General addition rules
- Conditional probability
- General multiplication rules

Tree diagrams

Bayes's rule

Independence again

Section 4.5 Summary

Section 4.5 Exercises

Chapter 4 Exercises

CHAPTER 5

**Sampling Distributions**

## **5.1 The Sampling Distribution of a Sample Mean**

The mean and standard deviation of  $\bar{x}$

The central limit theorem

A few more facts

Beyond the Basics: Weibull distributions

Section 5.1 Summary

Section 5.1 Exercises

## **5.2 Sampling Distributions for Counts and Proportions**

The binomial distributions for sample counts

Binomial distributions in statistical sampling

Finding binomial probabilities

Binomial mean and standard deviation

Sample proportions

Normal approximation for counts and proportions

The continuity correction

Binomial formula

The Poisson distributions

Section 5.2 Summary

Section 5.2 Exercises

Chapter 5 Exercises

## **CHAPTER 6**

### **Introduction to Inference**

## Introduction

Overview of inference

### 6.1 Estimating with Confidence

Statistical confidence

Confidence intervals

Confidence interval for a population mean

How confidence intervals behave

Choosing the sample size

Some cautions

Beyond the Basics: The bootstrap

[Section 6.1 Summary](#)

[Section 6.1 Exercises](#)

### 6.2 Tests of Significance

The reasoning of significance tests

Stating hypotheses

Test statistics

$P$ -values

Statistical significance

Tests for a population mean

Two-sided significance tests and confidence intervals

The  $P$ -value versus a statement of significance

[Section 6.2 Summary](#)

[Section 6.2 Exercises](#)

### 6.3 Use and Abuse of Tests

Choosing a level of significance

What statistical significance does not mean

Don't ignore lack of significance

Statistical inference is not valid for all sets of data

Beware of searching for significance

[Section 6.3 Summary](#)

[Section 6.3 Exercises](#)

### 6.4 Power and Inference as a Decision

Power

Increasing the power

Inference as decision

Two types of error

Error probabilities

The common practice of testing hypotheses

Section 6.4 Summary

Section 6.4 Exercises

Chapter 6 Exercises

CHAPTER 7

**Inference for Distributions**

## Introduction

### 7.1 Inference for the Mean of a Population

- The  $t$  distributions
- The one-sample  $t$  confidence interval
- The one-sample  $t$  test
- Matched pairs  $t$  procedures
- Robustness of the  $t$  procedures
- The power of the  $t$  test
- Inference for non-Normal populations

Section 7.1 Summary

Section 7.1 Exercises

### 7.2 Comparing Two Means

- The two-sample  $z$  statistic
- The two-sample  $t$  procedures
- The two-sample  $t$  confidence interval
- The two-sample  $t$  significance test
- Robustness of the two-sample procedures
- Inference for small samples
- Software approximation for the degrees of freedom
- The pooled two-sample  $t$  procedures

Section 7.2 Summary

Section 7.2 Exercises

### 7.3 Other Topics in Comparing Distributions

- Inference for population spread
- The  $F$  test for equality of spread
- Robustness of Normal inference procedures
- The power of the two-sample  $t$  test

Section 7.3 Summary

Section 7.3 Exercises

Chapter 7 Exercises

## CHAPTER 8

### Inference for Proportions

## Introduction

### 8.1 Inference for a Single Proportion

Large-sample confidence interval for a single proportion

Beyond the Basics: The plus four confidence interval for a single proportion

Significance test for a single proportion

Choosing a sample size

Section 8.1 Summary

Section 8.1 Exercises

### 8.2 Comparing Two Proportions

Large-sample confidence interval for a difference in proportions

Beyond the Basics: Plus four confidence interval for a difference in proportions

Significance test for a difference in proportions

Beyond the Basics: Relative risk

Section 8.2 Summary

Section 8.2 Exercises

Chapter 8 Exercises

## **PART III** Topics in Inference

### CHAPTER 9

### Analysis of Two-Way Tables

## **Introduction**

### **9.1 Inference for Two-Way Tables**

The hypothesis: no association Expected cell counts The chi-square test

Computations

Computing conditional distributions

The chi-square test and the  $z$  test

Models for two-way tables

Beyond the Basics: Meta-analysis

Section 9.1 Summary

### **9.2 Goodness of Fit**

Section 9.2 Summary

Chapter 9 Exercises

## **CHAPTER 10**

### **Inference for Regression**

## Introduction

### 10.1 Simple Linear Regression

- Statistical model for linear regression
- Data for simple linear regression
- Estimating the regression parameters
- Confidence intervals and significance tests
- Confidence intervals for mean response
- Prediction intervals
- Transforming variables
- Beyond the Basics: Nonlinear regression

#### Section 10.1 Summary

### 10.2 More Detail about Simple Linear Regression

- Analysis of variance for regression
- The ANOVA  $F$  test
- Calculations for regression inference
- Inference for correlation

#### Section 10.2 Summary

#### Chapter 10 Exercises

## CHAPTER 11

### Multiple Regression

## Introduction

### 11.1 Inference for Multiple Regression

- Population multiple regression equation
- Data for multiple regression
- Multiple linear regression model
- Estimation of the multiple regression parameters
- Confidence intervals and significance tests for regression coefficients
- ANOVA table for multiple regression
- Squared multiple correlation  $R^2$

### 11.2 A Case Study

- Preliminary analysis
- Relationships between pairs of variables
- Regression on high school grades
- Interpretation of results
- Residuals
- Refining the model
- Regression on SAT scores
- Regression using all variables
- Test for a collection of regression coefficients
- Beyond the Basics: Multiple logistic regression

Chapter 11 Summary

Chapter 11 Exercises

## CHAPTER 12

### One-Way Analysis of Variance

## Introduction

### 12.1 Inference for One-Way Analysis of Variance

- Data for one-way ANOVA
- Comparing means
- The two-sample  $t$  statistic
- An overview of ANOVA
- The ANOVA model
- Estimates of population parameters
- Testing hypotheses in one-way ANOVA
- The ANOVA table
- The  $F$  test

### 12.2 Comparing the Means

- Contrasts
- Multiple comparisons
- Software
- Power

Chapter 12 Summary

Chapter 12 Exercises

## CHAPTER 13

### Two-Way Analysis of Variance

## **Introduction**

### **13.1 The Two-Way ANOVA Model**

Advantages of two-way ANOVA

The two-way ANOVA model

Main effects and interactions

### **13.2 Inference for Two-Way ANOVA**

The ANOVA table for two-way ANOVA

Chapter 13 Summary

Chapter 13 Exercises



## **Companion Chapters**

(on the IPS website [www.whfreeman.com/ips8e](http://www.whfreeman.com/ips8e))

CHAPTER 14

**Logistic Regression**

## **Introduction**

### **14.1 The Logistic Regression Model**

Binomial distributions and odds

Odds for two groups

Model for logistic regression

Fitting and interpreting the logistic regression model

### **14.2 Inference for Logistic Regression**

Confidence intervals and significance tests

Multiple logistic regression

Chapter 14 Summary

Chapter 14 Exercises

Chapter 14 Notes and Data Sources

## **CHAPTER 15**

### **Nonparametric Tests**

## **Introduction**

### **15.1 The Wilcoxon Rank Sum Test**

- The rank transformation
- The Wilcoxon rank sum test
- The Normal approximation
- What hypotheses does Wilcoxon test?
- Ties
- Rank,  $t$ , and permutation tests

Section 15.1 Summary

Section 15.1 Exercises

### **15.2 The Wilcoxon Signed Rank Test**

- The Normal approximation
- Ties
- Testing a hypothesis about the median of a distribution

Section 15.2 Summary

Section 15.2 Exercises

### **15.3 The Kruskal-Wallis Test**

- Hypotheses and assumptions
- The Kruskal-Wallis test

Section 15.3 Summary

Section 15.3 Exercises

Chapter 15 Exercises

Chapter 15 Notes and Data Sources

## **CHAPTER 16**

### **Bootstrap Methods and Permutation Tests**

## **Introduction**

Software

### **16.1 The Bootstrap Idea**

The big idea: resampling and the bootstrap distribution

Thinking about the bootstrap idea

Using software

Section 16.1 Summary

Section 16.1 Exercises

### **16.2 First Steps in Using the Bootstrap**

Bootstrap  $t$  confidence intervals

Bootstrapping to compare two groups

Beyond the Basics: The bootstrap for a scatterplot smoother

Section 16.2 Summary

Section 16.2 Exercises

### **16.3 How Accurate Is a Bootstrap Distribution?**

Bootstrapping small samples

Bootstrapping a sample median

Section 16.3 Summary

Section 16.3 Exercises

### **16.4 Bootstrap Confidence Intervals**

Bootstrap percentile confidence intervals

A more accurate bootstrap confidence interval: BCa

Confidence intervals for the correlation

Section 16.4 Summary

Section 16.4 Exercises

### **16.5 Significance Testing Using Permutation Tests**

Using software

Permutation tests in practice

Permutation tests in other settings

Section 16.5 Summary

Section 16.5 Exercises

Chapter 16 Exercises

Chapter 16 Notes and Data Sources

CHAPTER 17

## **Statistics for Quality: Control and Capability**

## Introduction

Use of data to assess quality

## 17.1 Processes and Statistical Process Control

Describing processes

Statistical process control

$\bar{x}$  charts for process monitoring

$s$  charts for process monitoring

Section 17.1 Summary

Section 17.1 Exercises

## 17.2 Using Control Charts

$\bar{x}$  and  $R$  charts

Additional out-of-control rules

Setting up control charts

Comments on statistical control

Don't confuse control with capability!

Section 17.2 Summary

Section 17.2 Exercises

## 17.3 Process Capability Indexes

The capability indexes  $C_p$  and  $C_{pk}$

Cautions about capability indexes

Section 17.3 Summary

Section 17.3 Exercises

## 17.4 Control Charts for Sample Proportions

Control limits for  $p$  charts

Section 17.4 Summary

Section 17.4 Exercises

Chapter 17 Exercises

Chapter 17 Notes and Data Sources

Tables

Answers to Odd-Numbered Exercises

Notes and Data Sources

Photo Credits

## Index

## TO TEACHERS About This Book

Statistics is the science of data. *Introduction to the Practice of Statistics (IPS)* is an introductory text based on this principle. We present methods of basic statistics in a way that emphasizes working with data and mastering statistical reasoning. *IPS* is elementary in mathematical level but conceptually rich in statistical ideas. After completing a course based on our text, we would like students to be able to think objectively about conclusions drawn from data and use statistical methods in their own work.

In *IPS* we combine attention to basic statistical concepts with a comprehensive presentation of the elementary statistical methods that students will find useful in their work. *IPS* has been successful for several reasons:

1. *IPS* examines the nature of modern statistical practice at a level suitable for beginners. We focus on the production and analysis of data as well as the traditional topics of probability and inference.
2. *IPS* has a logical overall progression, so data production and data analysis are a major focus, while inference is treated as a tool that helps us draw conclusions from data in an appropriate way.
3. *IPS* presents data analysis as more than a collection of techniques for exploring data. We emphasize systematic ways of thinking about data. Simple principles guide the analysis: always plot your data; look for overall patterns and deviations from them; when looking at the overall pattern of a distribution for one variable, consider shape, center, and spread; for relations between two variables, consider form, direction, and strength; always ask whether a relationship between variables is influenced by other variables lurking in the background. We warn students about pitfalls in clear cautionary discussions.
4. *IPS* uses real examples to drive the exposition. Students learn the technique of least-squares regression and how to interpret the regression slope. But they also learn the conceptual ties between regression and correlation and the importance of looking for influential observations.
5. *IPS* is aware of current developments both in statistical science and in teaching statistics. Brief optional Beyond the Basics sections give quick overviews of topics such as density estimation, scatterplot smoothers, data mining, nonlinear regression, and meta-analysis. Chapter 16 gives an elementary introduction to the bootstrap and other computer-intensive statistical methods.

The title of the book expresses our intent to introduce readers to statistics as it is used in practice. Statistics in practice is concerned with drawing conclusions from data. We focus on problem solving rather than on methods that may be useful in specific settings.

**GAISE** The College Report of the Guidelines for Assessment and Instruction in Statistics Education (GAISE) Project (<http://www.amstat.org/education/gaise/>) was funded by the American Statistical Association to make recommendations for how introductory statistics courses should be taught. This report contains many interesting teaching suggestions and we strongly recommend that you read it. The philosophy and approach of *IPS* closely reflect the GAISE recommendations. Let's examine each of the recommendations in the context of *IPS*.

1. **Emphasize statistical literacy and develop statistical thinking.** Through our experiences as applied statisticians, we are very familiar with the components that are needed for the appropriate use of statistical methods. We focus on collecting and finding data, evaluating the quality of data, performing statistical analyses, and drawing conclusions. In examples and exercises throughout the text, we emphasize putting the analysis in the proper context and translating numerical and graphical summaries into conclusions.
2. **Use real data.** Many of the examples and exercises in *IPS* include data that we have obtained from collaborators or consulting clients. Other data sets have come from research related to these activities. We have also used the Internet as a data source, particularly for data related to social media and other topics of interest to undergraduates. With our emphasis on real data, rather than artificial data chosen to illustrate a calculation, we frequently encounter interesting issues that we explore. These include outliers and nonlinear relationships. All data sets are available from the text website.
3. **Stress conceptual understanding rather than mere knowledge of procedures.** With the software available today, it is very easy for almost anyone to apply a wide variety of statistical procedures, both simple and complex, to a set of data. Without a firm grasp of the concepts, such applications are frequently meaningless. By using the methods that we present on real sets of data, we believe that students will gain an excellent understanding of these concepts. Our emphasis is on the input (questions of interest, collecting or finding data, examining data) and the output (conclusions) for a statistical analysis. Formulas are given only where they will provide some insight into concepts.
4. **Foster active learning in the classroom.** As we mentioned above, we believe that statistics is exciting as something to do rather than something to talk about. Throughout the text we provide exercises in Use Your Knowledge sections that ask the students to perform some relatively simple tasks that reinforce the material just presented. Other exercises are particularly suited to being worked and discussed within a classroom setting.
5. **Use technology for developing concepts and analyzing data.** Technology has altered statistical practice in a fundamental way. In the past, some of the calculations that we performed were particularly difficult and tedious. In other words, they were not fun. Today, freed from the burden of computation by software, we can concentrate our efforts on the big picture: what questions are

we trying to address with a study and what can we conclude from our analysis?

5. **Use assessments to improve and evaluate student learning.** Our goal for students who complete a course based on *IPS* is that they are able to design and carry out a statistical study for a project in their capstone course or other setting. Our exercises are oriented toward this goal. Many ask about the design of a statistical study and the collection of data. Others ask for a paragraph summarizing the results of an analysis. This recommendation includes the use of projects, oral presentations, article critiques, and written reports. We believe that students using this text will be well prepared to undertake these kinds of activities. Furthermore, we view these activities not only as assessments but also as valuable tools for learning statistics.

**Teaching Recommendations** We have used *IPS* in courses taught to a variety of student audiences. For general undergraduates from mixed disciplines, we recommend covering Chapters 1 to 8 and Chapters 9, 10, or 12. For a quantitatively strong audience—sophomores planning to major in actuarial science or statistics—we recommend moving more quickly. Add Chapters 10 and 11 to the core material in Chapters 1 to 8. In general, we recommend de-emphasizing the material on probability because these students will take a probability course later in their program. For beginning graduate students in such fields as education, family studies, and retailing, we recommend that the students read the entire text (Chapters 11 and 13 lightly), again with reduced emphasis on Chapter 4 and some parts of Chapter 5. In all cases, beginning with data analysis and data production (Part I) helps students overcome their fear of statistics and builds a sound base for studying inference. We believe that *IPS* can easily be adapted to a wide variety of audiences.

## The Eighth Edition: What's New?

- **Text Organization** Each section now begins with the phrase “When you complete this section, you will be able to” followed by a bulleted list of behavioral objectives that the students should be able to master. Exercises that focus on these objectives appear at the beginning of the section exercises. The long introduction to **Chapter 1** has been replaced by a short introduction and a new section titled “Data,” which gives an overview of the basic ideas on the key characteristics of a set of data. The same approach has been taken with **Chapters 2 and 3**, which now have new sections titled “Relationships” and “Sources of Data,” respectively. A short introduction to the Poisson distribution has been added to **Section 5.2**. **Sections 9.1 and 9.2** have been combined with a more concise presentation of the material on computation and models from Section 5.2 of the seventh edition. In **Chapter 16**, the use of S-PLUS software has been replaced by R. Sections previously marked as **optional** are no longer given this designation. We have found that instructors make a variety of choices regarding what to include in their courses. General guidelines for different types of students are given in the Teaching Recommendations paragraph above.

- **Design** A new design incorporates colorful, revised figures throughout to aid the students’ understanding of text material. Photographs related to chapter examples and exercises make connections to real-life applications and provide a visual context for topics. More figures with software output have been included.

- **Exercises and Examples** Over 50% of the exercises are new or revised. There are more than 1700 exercises, a slight increase over the total in the seventh edition. To maintain the attractiveness of the examples to students, we have replaced or updated a large number of them. Over 35% of the 422 examples are new or revised. A list of exercises and examples categorized by application area is provided on the inside of the front cover.

In addition to the new eighth edition enhancements, *IPS* has retained the successful pedagogical features from previous editions:

- **Look Back** At key points in the text, Look Back margin notes direct the reader to the first explanation of a topic, providing page numbers for easy reference.



- **Caution** Warnings in the text, signaled by a caution icon, help students avoid common errors and misconceptions.



- **Challenge Exercises** More challenging exercises are signaled with an icon. Challenge exercises are varied: some are mathematical, some require open-ended investigation, and others require deeper thought about the basic concepts.



- **Applets** Applet icons are used throughout the text to signal where related interactive statistical applets can be found on the *IPS* website.



- **Use Your Knowledge Exercises** We have found these to be a very useful learning tool. Therefore, we have increased the number and variety of these exercises. These exercises are listed, with page numbers, before the section-ending exercises.



## Acknowledgments

We are pleased that the first seven editions of *Introduction to the Practice of Statistics* have helped to move the teaching of introductory statistics in a direction supported by most statisticians. We are grateful to the many colleagues and students who have provided helpful comments, and we hope that they will find this new edition another step forward. In particular, we would like to thank the following colleagues who offered specific comments on the new edition:

Ali Arab *Georgetown University*  
Sanjib Basu *Northern Illinois University*  
Mary Ellen Bock *Purdue University*  
Max Buot *Xavier University*  
Jerry J. Chen *Suffolk Community College*  
Pinyuen Chen *Syracuse University*  
Scott Crawford *University of Wyoming*  
Carolyn K. Cuff *Westminster College*  
K. L. D. Gunawardena *University of Wisconsin–Oshkosh*  
C. Clinton Harshaw *Presbyterian College*  
James Helmreich *Marist College*  
Ulrich Hoensch *Rocky Mountain College*  
Jeff Hovermill *Northern Arizona University*  
Debra L. Hydorn *University of Mary Washington*  
Monica Jackson *American University*  
Tiffany Kolba *Valparaiso University*  
Sharon Navard *College of New Jersey*  
Ronald C. Neath *Hunter College of CUNY*  
Esther M. Pearson *Lasell College*  
Thomas Pfaff *Ithaca College*  
Kathryn Prewitt *Arizona State University*  
Robert L. Sims *George Mason University*  
Thomas M. Songer *Portland Community College*  
Haiyan Su *Montclair State University*  
Anatoliy Swishchuk *University of Calgary*  
Frederick C. Tinsley *Colorado College*  
Terri Torres *Oregon Institute of Technology*

The professionals at W. H. Freeman and Company, in particular Ruth Baruth, Karen Carson, Katrina Wilhelm, Liam Ferguson, Elizabeth Geller, Vicki Tomaselli, and Lawrence Guerra, have contributed greatly to the success of *IPS*. In addition, we would like to thank Pamela Bruton, Jackie Miller, and Patricia Humphrey for their valuable contributions to the eighth edition. Most of all, we are grateful to the many friends and collaborators whose data and research questions have enabled us to gain a deeper understanding of the science of data. Finally, we would like to acknowledge the contributions of John W. Tukey, whose contributions to data analysis have had such a great influence on us as well as a whole

generation of applied statisticians.

## MEDIA AND SUPPLEMENTS

W. H. Freeman's new online homework system, **LaunchPad**, offers our quality content curated and organized for easy assignability in a simple but powerful interface. We've taken what we've learned from thousands of instructors and hundreds of thousands of students to create a new generation of W. H. Freeman/Macmillan technology.



**Curated Units.** Combining a curated collection of videos, homework sets, tutorials, applets, and e-Book content, LaunchPad's interactive units give you a building block to use as is or as a starting point for your own learning units. Thousands of exercises from the text can be assigned as online homework, including many algorithmic exercises. An entire unit's worth of work can be assigned in seconds, drastically reducing the amount of time it takes for you to have your course up and running.

**Easily customizable.** You can customize the LaunchPad Units by adding quizzes and other activities from our vast wealth of resources. You can also add a discussion board, a dropbox, and RSS feed, with a few clicks. LaunchPad allows you to customize your students' experience as much or as little as you like.

**Useful analytics.** The gradebook quickly and easily allows you to look up performance metrics for classes, individual students, and individual assignments.

**Intuitive interface and design.** The student experience is simplified. Students' navigation options and expectations are clearly laid out at all times, ensuring they can never get lost in the system.

### **Assets integrated into LaunchPad include:**

**Interactive e-Book.** Every LaunchPad e-Book comes with powerful study tools for students, video and multimedia content, and easy customization for instructors. Students can search, highlight, and bookmark, making it easier to study and access key content. And teachers can ensure that their classes get just the book they want to deliver: customize and rearrange chapters, add and share notes and discussions, and link to quizzes, activities, and other resources.



**LearningCurve** provides students and instructors with powerful adaptive quizzing, a game-like format, direct links to the e-Book, and instant feedback. The quizzing system features questions tailored specifically to the text and adapts to

students' responses, providing material at different difficulty levels and topics based on student performance.



**SolutionMaster** offers an easy-to-use web-based version of the instructor's solutions, allowing instructors to generate a solution file for any set of homework exercises.

**New Stepped Tutorials** are centered on algorithmically generated quizzing with step-by-step feedback to help students work their way toward the correct solution. These new exercise tutorials (two to three per chapter) are easily assignable and assessable.

**Statistical Video Series** consists of StatClips, StatClips Examples, and Statistically Speaking "Snapshots." View animated lecture videos, whiteboard lessons, and documentary-style footage that illustrate key statistical concepts and help students visualize statistics in real-world scenarios.

**New Video Technology Manuals** available for TI-83/84 calculators, Minitab, Excel, JMP, SPSS, R, Rcmdr, and CrunchIT!® provide brief instructions for using specific statistical software.

**Updated StatTutor Tutorials** offer multimedia tutorials that explore important concepts and procedures in a presentation that combines video, audio, and interactive features. The newly revised format includes built-in, assignable assessments and a bright new interface.



**Updated Statistical Applets** give students hands-on opportunities to familiarize themselves with important statistical concepts and procedures, in an interactive setting that allows them to manipulate variables and see the results graphically. Icons in the textbook indicate when an applet is available for the material being covered.

**CrunchIT!®** is a web-based statistical program that allows users to perform all the statistical operations and graphing needed for an introductory statistics course and more. It saves users time by automatically loading data from *IPS* 8e, and it provides the flexibility to edit and import additional data.

**Stats@Work Simulations** put students in the role of the statistical consultant, helping them better understand statistics interactively within the context of real-life scenarios.

**EESEE Case Studies** (Electronic Encyclopedia of Statistical Examples and Exercises), developed by The Ohio State University Statistics Department, teach

students to apply their statistical skills by exploring actual case studies using real data.

**Data files** are available in ASCII, Excel, TI, Minitab, SPSS (an IBM Company),\* and JMP formats.

\*SPSS was acquired by IBM in October 2009.

**Student Solutions Manual** provides solutions to the odd-numbered exercises in the text. Available electronically within LaunchPad, as well as in print form.

**Interactive Table Reader** allows students to use statistical tables interactively to seek the information they need.

**Instructor's Guide with Full Solutions** includes teaching suggestions, chapter comments, and detailed solutions to all exercises. Available electronically within LaunchPad, as well as on the IRCD and in print form.

**Test Bank** offers hundreds of multiple-choice questions. Also available on CD-ROM (for Windows and Mac), where questions can be downloaded, edited, and resequenced to suit each instructor's needs.

**Lecture PowerPoint Slides** offer a detailed lecture presentation of statistical concepts covered in each chapter of *IPS*.

## **Additional Resources Available with IPS 8e**

**Companion Website [www.whfreeman.com/ips8e](http://www.whfreeman.com/ips8e)** This open-access website includes statistical applets, data files, supplementary exercises, and self-quizzes. The website also offers four optional companion chapters covering logistic regression, nonparametric tests, bootstrap methods and permutation tests, and statistics for quality control and capability.

Instructor access to the Companion Website requires user registration as an instructor and features all of the open-access student web materials, plus:

- Instructor version of **EESEE** with solutions to the exercises in the student version.
- **PowerPoint Slides** containing all textbook figures and tables.
- **Lecture PowerPoint Slides**

**Special Software Packages** Student versions of JMP and Minitab are available for packaging with the text. Contact your W. H. Freeman representative for information or visit [www.whfreeman.com](http://www.whfreeman.com).

**Enhanced Instructor's Resource CD-ROM, ISBN: 1-4641-3360-3** Allows instructors to **search** and **export** (by key term or chapter) all the resources available on the student companion website plus the following:

- All text images and tables
- Instructor's Guide with Full Solutions
- PowerPoint files and lecture slides

- Test Bank files

**Course Management Systems** W. H. Freeman and Company provides courses for Blackboard, Angel, Desire2Learn, Canvas, Moodle, and Sakai course management systems. These are completely integrated solutions that you can easily customize and adapt to meet your teaching goals and course objectives. Visit [macmillanhighered.com/Catalog/other/Coursepack](http://macmillanhighered.com/Catalog/other/Coursepack) for more information.



**iClicker** is a two-way radio-frequency classroom response solution developed by educators for educators. Each step of i-clicker's development has been informed by teaching and learning. To learn more about packaging i-clicker with this textbook, please contact your local sales rep or visit [www.iclicker.com](http://www.iclicker.com).

## TO STUDENTS What Is Statistics?

Statistics is the science of collecting, organizing, and interpreting numerical facts, which we call *data*. We are bombarded by data in our everyday lives. The news mentions movie box-office sales, the latest poll of the president's popularity, and the average high temperature for today's date. Advertisements claim that data show the superiority of the advertiser's product. All sides in public debates about economics, education, and social policy argue from data. A knowledge of statistics helps separate sense from nonsense in this flood of data.

The study and collection of data are also important in the work of many professions, so training in the science of statistics is valuable preparation for a variety of careers. Each month, for example, government statistical offices release the latest numerical information on unemployment and inflation. Economists and financial advisers, as well as policy makers in government and business, study these data in order to make informed decisions. Doctors must understand the origin and trustworthiness of the data that appear in medical journals. Politicians rely on data from polls of public opinion. Business decisions are based on market research data that reveal consumer tastes and preferences. Engineers gather data on the quality and reliability of manufactured products. Most areas of academic study make use of numbers and, therefore, also make use of the methods of statistics. This means it is extremely likely that your undergraduate research projects will involve, at some level, the use of statistics.

## Learning from Data

*The goal of statistics is to learn from data.* To learn, we often perform calculations or make graphs based on a set of numbers. But to learn from data, we must do more than calculate and plot, because data are not just numbers; they are numbers that have some context that helps us learn from them.

Two-thirds of Americans are overweight or obese according to the Centers for Disease Control and Prevention (CDC) website ([www.cdc.gov/nchs/nhanes.htm](http://www.cdc.gov/nchs/nhanes.htm)). What does it mean to be obese or to be overweight? To answer this question we need to talk about body mass index (BMI). Your weight in kilograms divided by the square of your height in meters is your BMI. A man who is 6 feet tall (1.83 meters) and weighs 180 pounds (81.65 kilograms) will have a BMI of  $81.65/(1.83)^2 = 24.4 \text{ kg/m}^2$ . How do we interpret this number? According to the CDC, a person is classified as overweight if his or her BMI is between 25 and 29  $\text{kg/m}^2$  and as obese if his or her BMI is 30  $\text{kg/m}^2$  or more. Therefore, two-thirds of Americans have a BMI of 25  $\text{kg/m}^2$  or more. The man who weighs 180 pounds and is 6 feet tall is not overweight or obese, but if he gains 5 pounds, his BMI would increase to 25.1, and he would be classified as overweight.

When you do statistical problems, even straightforward textbook problems, don't just graph or calculate. Think about the context and state your conclusions in the specific setting of the problem. As you are learning how to do statistical calculations and graphs, remember that the goal of statistics is not calculation for its own sake but gaining understanding from numbers. The calculations and graphs can be automated by a calculator or software, but you must supply the understanding. This book presents only the most common specific procedures for statistical analysis. A thorough grasp of the principles of statistics will enable you to quickly learn more advanced methods as needed. On the other hand, a fancy computer analysis carried out without attention to basic principles will often produce elaborate nonsense. As you read, seek to understand the principles as well as the necessary details of methods and recipes.

## The Rise of Statistics

Historically, the ideas and methods of statistics developed gradually as society grew interested in collecting and using data for a variety of applications. The earliest origins of statistics lie in the desire of rulers to count the number of inhabitants or measure the value of taxable land in their domains. As the physical sciences developed in the seventeenth and eighteenth centuries, the importance of careful measurements of weights, distances, and other physical quantities grew. Astronomers and surveyors striving for exactness had to deal with variation in their measurements. Many measurements should be better than a single measurement, even though they vary among themselves. How can we best combine many varying observations? Statistical methods that are still important were invented in order to analyze scientific measurements.

By the nineteenth century, the agricultural, life, and behavioral sciences also began to rely on data to answer fundamental questions. How are the heights of parents and children related? Does a new variety of wheat produce higher yields than the old, and under what conditions of rainfall and fertilizer? Can a person's mental ability and behavior be measured just as we measure height and reaction time? Effective methods for dealing with such questions developed slowly and with much debate.

As methods for producing and understanding data grew in number and sophistication, the new discipline of statistics took shape in the twentieth century. Ideas and techniques that originated in the collection of government data, in the study of astronomical or biological measurements, and in the attempt to understand heredity or intelligence came together to form a unified "science of data." That science of data—statistics—is the topic of this text.

## The Organization of This Book

Part I of this book, called simply “Looking at Data,” concerns data analysis and data production. The first two chapters deal with statistical methods for organizing and describing data. These chapters progress from simpler to more complex data. Chapter 1 examines data on a single variable, Chapter 2 is devoted to relationships among two or more variables. You will learn both how to examine data produced by others and how to organize and summarize your own data. These summaries will first be graphical, then numerical, and then, when appropriate, in the form of a mathematical model that gives a compact description of the overall pattern of the data. Chapter 3 outlines arrangements (called designs) for producing data that answer specific questions. The principles presented in this chapter will help you to design proper samples and experiments for your research projects and to evaluate other such investigations in your field of study.

Part II, consisting of Chapters 4 to 8, introduces statistical inference— formal methods for drawing conclusions from properly produced data. Statistical inference uses the language of probability to describe how reliable its conclusions are, so some basic facts about probability are needed to understand inference. Probability is the subject of Chapters 4 and 5. Chapter 6, perhaps the most important chapter in the text, introduces the reasoning of statistical inference. Effective inference is based on good procedures for producing data (Chapter 3), careful examination of the data (Chapters 1 and 2), and an understanding of the nature of statistical inference as discussed in Chapter 6. Chapters 7 and 8 describe some of the most common specific methods of inference, for drawing conclusions about means and proportions from one and two samples.

The five shorter chapters in Part III introduce somewhat more advanced methods of inference, dealing with relations in categorical data, regression and correlation, and analysis of variance. Four supplementary chapters, available from the text website, present additional statistical topics.

### What Lies Ahead

*Introduction to the Practice of Statistics* is full of data from many different areas of life and study. Many exercises ask you to express briefly some understanding gained from the data. In practice, you would know much more about the background of the data you work with and about the questions you hope the data will answer. No textbook can be fully realistic. But it is important to form the habit of asking, “What do the data tell me?” rather than just concentrating on making graphs and doing calculations.

You should have some help in automating many of the graphs and calculations. You should certainly have a calculator with basic statistical functions. Look for keywords such as “two-variable statistics” or “regression” when you shop for a calculator. More advanced (and more expensive) calculators will do much more,

including some statistical graphs. You may be asked to use software as well. There are many kinds of statistical software, from spreadsheets to large programs for advanced users of statistics. The kind of computing available to learners varies a great deal from place to place—but the big ideas of statistics don't depend on any particular level of access to computing.

Because graphing and calculating are automated in statistical practice, the most important assets you can gain from the study of statistics are an understanding of the big ideas and the beginnings of good judgment in working with data. Ideas and judgment can't (at least yet) be automated. They guide you in telling the computer what to do and in interpreting its output. This book tries to explain the most important ideas of statistics, not just teach methods. Some examples of big ideas that you will meet are “always plot your data,” “randomized comparative experiments,” and “statistical significance.”

You learn statistics by doing statistical problems. “Practice, practice, practice.” Be prepared to work problems. The basic principle of learning is persistence. Being organized and persistent is more helpful in reading this book than knowing lots of math. The main ideas of statistics, like the main ideas of any important subject, took a long time to discover and take some time to master. The gain will be worth the pain.

## ABOUT THE AUTHORS

**David S. Moore** is Shanti S. Gupta Distinguished Professor of Statistics, Emeritus, at Purdue University and was 1998 president of the American Statistical Association. He received his AB from Princeton and his PhD from Cornell, both in mathematics. He has written many research papers in statistical theory and served on the editorial boards of several major journals. Professor Moore is an elected fellow of the American Statistical Association and of the Institute of Mathematical Statistics and an elected member of the International Statistical Institute. He has served as program director for statistics and probability at the National Science Foundation.

In recent years, Professor Moore has devoted his attention to the teaching of statistics. He was the content developer for the Annenberg/Corporation for Public Broadcasting college-level telecourse *Against All Odds: Inside Statistics* and for the series of video modules *Statistics: Decisions through Data*, intended to aid the teaching of statistics in schools. He is the author of influential articles on statistics education and of several leading texts. Professor Moore has served as president of the International Association for Statistical Education and has received the Mathematical Association of America's national award for distinguished college or university teaching of mathematics.

**George P. McCabe** is Associate Dean for Academic Affairs in the College of Science and Professor of Statistics at Purdue University. In 1966, he received a BS degree in mathematics from Providence College and in 1970 a PhD in mathematical statistics from Columbia University. His entire professional career has been spent at Purdue, with sabbaticals at Princeton University, the Commonwealth Scientific and Industrial Research Organization (CSIRO) in Melbourne (Australia), the University of Berne (Switzerland), the National Institute of Standards and Technology (NIST) in Boulder, Colorado, and the National University of Ireland in Galway. Professor McCabe is an elected fellow of the American Association for the Advancement of Science and of the American Statistical Association; he was 1998 Chair of its section on Statistical Consulting. In 2008–2010, he served on the Institute of Medicine Committee on Nutrition Standards for the National School Lunch and Breakfast Programs. He has served on the editorial boards of several statistics journals. He has consulted with many major corporations and has testified as an expert witness on the use of statistics in several cases.

Professor McCabe's research interests have focused on applications of statistics. Much of his recent work has focused on problems in nutrition, including nutrient requirements, calcium metabolism, and bone health. He is the author or coauthor of over 170 publications in many different journals.

**Bruce A. Craig** is Professor of Statistics and Director of the Statistical Consulting Service at Purdue University. He received his BS in mathematics and economics from Washington University in St. Louis and his PhD in statistics from the University of Wisconsin–Madison. He is an elected fellow of the American Statistical Association and was chair of its section on Statistical Consulting in 2009. He is also an active member of the Eastern North American Region of the International Biometrics Society and was elected by the voting membership to the Regional Committee between 2003 and 2006. Professor Craig has served on the editorial board of several statistical journals and has been a member of several data and safety monitoring boards, including Purdue’s institutional review board.

Professor Craig’s research interests focus on the development of novel statistical methodology to address research questions in the life sciences. Areas of current interest are protein structure determination, diagnostic testing, and animal abundance estimation. In 2005, he was named Purdue University Faculty Scholar.

## DATA TABLE INDEX

**TABLE 1.1** IQ test scores for 60 randomly chosen fifth-grade students

**TABLE 1.2** Service times (seconds) for calls to a customer service center

**TABLE 1.3** Educational data for 78 seventh-grade students

**TABLE 2.1** World record times for the 10,000-meter run

**TABLE 2.2** Four data sets for exploring correlation and regression

**TABLE 2.3** Two measures of glucose level in diabetics

**TABLE 2.4** Dwelling permits, sales, and production for 21 European countries

**TABLE 2.5** Fruit and vegetable consumption and smoking

**TABLE 7.1** Monthly rates of return on a portfolio (%)

**TABLE 7.2** Aggressive behaviors of dementia patients

**TABLE 7.3** Length (in seconds) of audio files sampled from an iPod

**TABLE 7.4** DRP scores for third-graders

**TABLE 7.5** Seated systolic blood pressure (mm Hg)

**TABLE 10.1** In-state tuition and fees (in dollars) for 33 public universities

**TABLE 10.2** Sales price and assessed value (in \$ thousands) of 30 homes in a midwestern city

**TABLE 10.3** Annual number of tornadoes in the United States between 1953 and 2012

**TABLE 10.4** Watershed area (km<sup>2</sup>), percent forest, and index of biotic integrity

**TABLE 12.1** Age at death for North American women writers

**TABLE 13.1** Safety behaviors of abused women

**TABLE 13.2** Iron content (mg/100 g) of food cooked in different pots

**TABLE 13.3** Tool diameter data

<b>16.1</b>	Degree of Reading Power scores for third-graders
<b>TABLE 16.2</b>	Aggressive behaviors of dementia patients
<b>TABLE 16.3</b>	Serum retinol levels in two groups of children
<b>TABLE 17.1</b>	Twenty control chart samples of water resistance
<b>TABLE 17.2</b>	Control chart constants
<b>TABLE 17.3</b>	Twenty samples of size 3, with $\bar{x}$ and $s$
<b>TABLE 17.4</b>	Three sets of $\bar{x}$ 's from 20 samples of size
<b>TABLE 17.5</b>	Twenty samples of size 4, with $\bar{x}$ and $s$
<b>TABLE 17.6</b>	$\bar{x}$ and $s$ for 24 samples of elastomer viscosity
<b>TABLE 17.7</b>	$\bar{x}$ and $s$ for 24 samples of label placement
<b>TABLE 17.8</b>	$\bar{x}$ and $s$ for 24 samples of label placement
<b>TABLE 17.9</b>	Hospital losses for 15 samples of DRG 209 patients
<b>TABLE 17.10</b>	Daily calibration samples for a Lunar bone densitometer
<b>TABLE 17.11</b>	$\bar{x}$ and $s$ for samples of bore diameter
<b>TABLE 17.12</b>	Fifty control chart samples of call center response times
<b>TABLE 17.13</b>	Proportions of workers absent during four weeks
<b>TABLE 17.14</b>	$\bar{x}$ and $s$ for samples of film thickness

## **BEYOND THE BASICS INDEX**

- CHAPTER 1** Density estimation
- CHAPTER 2** Scatterplot smoothers
- CHAPTER 2** Data mining
- CHAPTER 3** Capture-recapture sampling
- CHAPTER 4** More laws of large numbers
- CHAPTER 5** Weibull distributions
- CHAPTER 6** The bootstrap
- CHAPTER 8** The plus four confidence interval for a single proportion
- CHAPTER 8** The plus four confidence interval for a difference in proportions
- CHAPTER 8** Relative risk
- CHAPTER 9** Meta-analysis
- CHAPTER 10** Nonlinear regression
- CHAPTER 11** Multiple logistic regression
- CHAPTER 16** The bootstrap for a scatterplot smoother

# 1 Looking at Data—Distributions CHAPTER



- 1.1 Data
- 1.2 Displaying Distributions with Graphs
- 1.3 Describing Distributions with Numbers
- 1.4 Density Curves and Normal Distributions

# Introduction

*Statistics is the science of learning from data.* Data are numerical or qualitative descriptions of the objects that we want to study. In this chapter, we will master the art of examining data.

We begin in Section 1.1 with some basic ideas about data. We will learn about the different types of data that are collected and how data sets are organized.

Section 1.2 starts our process of learning from data by looking at graphs. These visual displays give us a picture of the overall patterns in a set of data. We have excellent software tools that help us make these graphs. However, it takes a little experience and a lot of judgment to study the graphs carefully and to explain what they tell us about our data.

Section 1.3 continues our process of learning from data by computing numerical summaries. These sets of numbers describe key characteristics of the patterns that we saw in our graphical summaries.

A final section in this chapter helps us make the transition from data summaries to statistical models. We learn about using density curves to describe a set of data. The Normal distributions are also introduced in this section. These distributions can be used to describe many sets of data that we will encounter. They also play a fundamental role in the methods that we will use to draw conclusions from many sets of data.

## 1.1 Data

When you complete this section, you will be able to

- Give examples of cases in a data set.
- Identify the variables in a data set.
- Demonstrate how a label can be used as a variable in a data set.
- Identify the values of a variable.
- Classify variables as categorical or quantitative.
- Describe the key characteristics of a set of data.
- Explain how a rate is the result of adjusting one variable to create another.

A statistical analysis starts with a set of data. We construct a set of data by first deciding what *cases*, or units, we want to study. For each case, we record information about characteristics that we call *variables*.

### CASES, LABELS, VARIABLES, AND VALUES

**Cases** are the objects described by a set of data. Cases may be customers, companies, subjects in a study, units in an experiment, or other objects.

A **label** is a special variable used in some data sets to distinguish the different cases.

A **variable** is a characteristic of a case.

Different cases can have different **values** of the variables.

### EXAMPLE

#### 1.1 Over 12 billion sold.

Apple's music-related products and services generated \$1.8 billion in the third quarter of 2012. Since Apple started marketing iTunes in 2003, they have sold

over 12 billion songs. Let's take a look at this remarkable product. Figure 1.1 is part of an iTunes playlist named IPS. The six songs shown are *cases*. They are numbered from 1 to 6 in the first column. These numbers are the *labels* that distinguish the six songs. The following five columns give name (of the song), time (the length of time it takes to play the song), artist, album, and genre.

Some variables, like the name of a song and the artist simply place cases into categories. Others, like the length of a song, take numerical values for which we can do arithmetic. It makes sense to give an average length of time for a collection of songs, but it does not make sense to give an “average” album. We can, however, count the numbers of songs on different albums, and we can do arithmetic with these counts.

	Name	Time	Artist	Album	Genre
1	Die Young	3:32	Ke\$ha	Warrior (Deluxe Version)	Pop
2	One More Night	3:39	Maroon 5	Overexposed (Deluxe Version)	Pop
3	The Morning of the Rain (feat. Jonathan Jackson)	3:15	Nashville Cast	The Morning of the Rain (feat. Jonathan Jackson) - Single	Country
4	Diamonds	3:45	Rihanna	Diamonds - Single	Pop
5	Love and War	4:01	Tamar Braxton	Love and War - Single	R&B/Soul
6	Locked Out of Heaven	3:53	Bruno Mars	Locked Out of Heaven - Single	Pop

**FIGURE 1.1**  
Part of an iTunes playlist, for Example 1.1.

## CATEGORICAL AND QUANTITATIVE VARIABLES

A **categorical variable** places a case into one of several groups or categories. A **quantitative variable** takes numerical values for which arithmetic operations such as adding and averaging make sense. The **distribution** of a variable tells us what values it takes and how often it takes these values.

### EXAMPLE

#### 1.2 Categorical and quantitative variables in iTunes playlist.

The IPS iTunes playlist contains five variables. These are the name, time, artist, album, and genre. The time is a quantitative variable. Name, artist, album, and genre are categorical variables.

*An appropriate label for your cases should be chosen carefully.* In our iTunes example, a natural choice of a label would be the name of the song. However, if you have more than one artist performing the same song, or the same artist performing the same song on different albums, then the name of the song would not uniquely label each of the songs in your playlist.



*A quantitative variable such as the time in the iTunes playlist requires some special attention before we can do arithmetic with its values.* The first song in the playlist has time equal to 3:32—that is, 3 minutes and 32 seconds. To do arithmetic with this variable, we should first convert all the values so that they have a single unit. We could convert to seconds; 3 minutes is 180 seconds, so the total time is  $180 + 32$ , or 212 seconds. An alternative would be to convert to minutes; 32 seconds is 0.533 minute, so time written in this way is 3.533 minutes.



## USE YOUR KNOWLEDGE

### 1.1 Time in the iTunes playlist.

In the iTunes playlist, do you prefer to convert the time to seconds or minutes? Give a reason for your answer.

We use the term *units of measurement* to refer to the seconds or minutes that tell us how the variable time is measured. If we were measuring heights of children, we might choose to use either inches or centimeters. The units of measurement are an important part of the description of a quantitative variable.

units of measurement

## Key characteristics of a data set

In practice, any set of data is accompanied by background information that helps us understand the data. When you plan a statistical study or explore data from someone else's work, ask yourself the following questions:

1. **Who?** What **cases** do the data describe? **How many** cases does the data set contain?
2. **What?** How many **variables** do the data contain? What are the **exact definitions** of these variables? What are the units of measurement for each quantitative variable?
3. **Why? What purpose** do the data have? Do we hope to answer some specific questions? Do we want to draw conclusions about cases other than the ones we actually have data for? Are the variables that are recorded suitable for the intended purpose?

### EXAMPLE

#### 1.3 Data for students in a statistics class.



Figure 1.2 shows part of a data set for students enrolled in an introductory statistics class. Each row gives the data on one student. The values for the different variables are in the columns. This data set has eight variables. ID is a label for each student. Exam1, Exam2, Homework, Final, and Project give the

points earned, out of a total of 100 possible, for each of these course requirements. Final grades are based on a possible 200 points for each exam and the Final, 300 points for Homework, and 100 points for Project. TotalPoints is the variable that gives the composite score. It is computed by adding 2 times Exam1, Exam2, and Final, 3 times Homework, and 1 times Project. Grade is the grade earned in the course. This instructor used cutoffs of 900, 800, 700, etc. for the letter grades.

## USE YOUR KNOWLEDGE

### 1.2 Who, what, and why for the statistics class data.

Answer the who, what, and why questions for the statistics class data set.

### 1.3 Read the spreadsheet.

Refer to Figure 1.2. Give the values of the variables Exam1, Exam2, and Final for the student with ID equal to 104.

	A	B	C	D	E	F	G	H
1	ID	Exam1	Exam2	Homework	Final	Project	TotalPoints	Grade
2	101	89	94	88	87	95	899	B
3	102	78	84	90	89	94	866	B
4	103	71	80	75	79	95	780	C
5	104	95	98	97	96	93	962	A
6	105	79	88	85	88	96	861	B

**FIGURE 1.2**  
Spreadsheet for Example 1.3.

### 1.4 Calculate the grade.

A student whose data do not appear on the spreadsheet scored 83 on Exam1, 82 on Exam2, 77 for Homework, 90 on the Final, and 80 on the Project. Find TotalPoints for this student and give the grade earned.

The display in Figure 1.2 is from an Excel *spreadsheet*. Spreadsheets are very useful for doing the kind of simple computations that you did in Exercise 1.4. You can type in a formula and have the same computation performed for each row.

### spreadsheet

Note that the names we have chosen for the variables in our spreadsheet do not have spaces. For example, we could have used the name “Exam 1” for the first-exam score rather than Exam1. In some statistical software packages, however, spaces are not allowed in variable names. *For this reason, when creating spreadsheets for eventual use with statistical software, it is best to avoid spaces in variable names.* Another convention is to use an underscore ( `_` ) where you would normally use a space. For our data set, we could use Exam\_1, Exam\_2, and Final\_Exam.



---

## EXAMPLE

### 1.4 Cases and variables for the statistics class data.

The data set in Figure 1.2 was constructed to keep track of the grades for students in an introductory statistics course. The cases are the students in the class. There are eight variables in this data set. These include a label for each student and scores for the various course requirements. There are no units for ID and grade. The other variables all have “points” as the unit.

---

## EXAMPLE

### 1.5 Statistics class data for a different purpose.

	A	B	C	D	E	F
1	ID	TotalPoints	Grade	Gender	PrevStat	Year
2	101	899	B	F	Yes	4
3	102	866	B	M	Yes	3
4	103	780	C	M	No	3
5	104	962	A	M	No	1
6	105	861	B	F	No	4

**FIGURE 1.3**  
Spreadsheet for Example 1.5.

Suppose that the data for the students in the introductory statistics class were also to be used to study relationships between student characteristics and success in the course. For this purpose, we might want to use a data set like the spreadsheet in Figure 1.3. Here, we have decided to focus on the TotalPoints and Grade as the outcomes of interest. Other variables of interest have been included: Gender, PrevStat (whether or not the student has taken a statistics course previously), and Year (student classification as first, second, third, or fourth year). ID is a categorical variable, TotalPoints is a quantitative variable, and the remaining variables are all categorical.

In our example, the possible values for the grade variable are A, B, C, D, and F. When computing grade point averages, many colleges and universities translate these letter grades into numbers using  $A = 4$ ,  $B = 3$ ,  $C = 2$ ,  $D = 1$ , and  $F = 0$ . The transformed variable with numeric values is considered to be quantitative because we can average the numerical values across different courses to obtain a grade point average.

Sometimes, experts argue about numerical scales such as this. They ask whether or not the difference between an A and a B is the same as the difference between a D and an F. Similarly, many questionnaires ask people to respond on a 1 to 5 scale with 1 representing strongly agree, 2 representing agree, etc. Again, we could ask whether or not the five possible values for this scale are equally spaced in some sense. From a practical point of view, however, the averages that can be computed when we convert categorical scales such as these to numerical values frequently provide a very useful way to summarize data.

## USE YOUR KNOWLEDGE

### 1.5 Apartment rentals.

A data set lists apartments available for students to rent. Information provided includes the monthly rent, whether or not cable is included free of charge, whether or not pets are allowed, the number of bedrooms, and the distance to the campus. Describe the cases in the data set, give the number of variables, and specify whether each variable is categorical or quantitative.

Often the variables in a statistical study are easy to understand: height in centimeters, study time in minutes, and so on. But each area of work also has its own special variables. A psychologist uses the Minnesota Multiphasic Personality Inventory (MMPI), and a physical fitness expert measures “VO2 max,” the volume of oxygen consumed per minute while exercising at your maximum capacity. Both of these variables are measured with special *instruments*. VO2 max is measured by exercising while breathing into a mouthpiece connected to an apparatus that measures oxygen consumed. Scores on the MMPI are based on a long questionnaire, which is also an instrument.

#### instrument

Part of mastering your field of work is learning what variables are important and how they are best measured. Because details of particular measurements usually require knowledge of the particular field of study, we will say little about them.

*Be sure that each variable really does measure what you want it to. A poor choice of variables can lead to misleading conclusions.* Often, for example, the **rate** at which something occurs is a more meaningful measure than a simple count of occurrences.

#### rate



## EXAMPLE

### 1.6 Comparing colleges based on graduates.

Think about comparing colleges based on the numbers of graduates. This view tells you something about the relative sizes of different colleges. However, if you are interested in how well colleges succeed at graduating students whom they admit, it would be better to use a rate. For example, you can find data on the Internet on the six-year graduation rates of different colleges. These rates are computed by examining the progress of first-year students who enroll in a given year. Suppose that at College A there were 1000 first-year students in a particular year, and 800 graduated within six years. The graduation rate is

$$\frac{800}{1000}=0.80$$

or 80%. College B has 2000 students who entered in the same year, and 1200 graduated within six years. The graduation rate is

$$\frac{1200}{2000}=0.60$$

or 60%. How do we compare these two colleges? College B has more graduates, but College A has a better graduation rate.

## USE YOUR KNOWLEDGE

### 1.6 How should you express the change?

Between the first exam and the second exam in your statistics course you increased the amount of time that you spent working exercises. Which of the following three ways would you choose to express the results of your increased work: (a) give the grades on the two exams, (b) give the ratio of the grade on the second exam divided by the grade on the first exam, or (c) take the difference between the grade on the second exam and the grade on the first exam, and express this as a percent of the grade on the first exam. Give reasons for your answer.

### 1.7 Which variable would you choose.

Refer to Example 1-6, on colleges and their graduates.

- (a) Give a setting where you would prefer to evaluate the colleges based on the numbers of graduates. Give a reason for your choice.
- (b) Give a setting where you would prefer to evaluate the colleges based on the graduation rates. Give a reason for your choice.

In Example 1.6, when we computed the graduation rate, we used the total number of students to adjust the number of graduates. We constructed a new variable by dividing the number of graduates by the total number of students. Computing a rate is just one of several ways of *adjusting one variable to create another*. We often divide one variable by another to compute a more meaningful variable to study. Example 1.20 (page 22) is another type of adjustment.

### adjusting one variable to create another

Exercises 1.6 and 1.7 illustrate an important point about presenting the results of your statistical calculations. *Always consider how to best communicate your results to a general audience.* For example, the numbers produced by your calculator or by statistical software frequently contain more digits than are needed. Be sure that you do not include extra information generated by software that will distract from a clear explanation of what you have found.



## SECTION 1.1 Summary

A data set contains information on a number of **cases**. Cases may be customers, companies, subjects in a study, units in an experiment, or other objects.

For each case, the data give values for one or more **variables**. A variable describes some characteristic of a case, such as a person's height, gender, or salary. Variables can have different **values** for different cases.

A **label** is a special variable used to identify cases in a data set.

Some variables are **categorical** and others are **quantitative**. A categorical variable places each individual into a category, such as male or female. A quantitative variable has numerical values that measure some characteristic of each case, such as height in centimeters or annual salary in dollars.

The **key characteristics** of a data set answer the questions Who?, What?, and Why?

## SECTION 1.1 Exercises

*For Exercise 1.1, see page 3; for Exercises 1.2 to 1.4, see pages 4–5; for Exercise 1.5, see page 6; and for Exercises 1.6 and 1.7, see page 7.*

## 1.8 Summer jobs.

You are collecting information about summer jobs that are available for college students in your area. Describe a data set that you could use to organize the information that you collect.

- (a) What are the cases?
- (b) Identify the variables and their possible values.
- (c) Classify each variable as categorical or quantitative. Be sure to include at least one of each.
- (d) Use a label and explain how you chose it.
- (e) Summarize the key characteristics of your data set.

## 1.9 Employee application data.

The personnel department keeps records on all employees in a company. Here is the information that they keep in one of their data files: employee identification number, last name, first name, middle initial, department, number of years with the company, salary, education (coded as high school, some college, or college degree), and age.

- (a) What are the cases for this data set?
- (b) Describe each type of information as a label, a quantitative variable, or a categorical variable.
- (c) Set up a spreadsheet that could be used to record the data. Give appropriate column headings and five sample cases.

## 1.10 How would you rank cities?

Various organizations rank cities and produce lists of the 10 or the 100 best based on various measures. Create a list of criteria that you would use to rank cities. Include at least eight variables and give reasons for your choices. Say whether each variable is quantitative or categorical.

## 1.11 Survey of students.

A survey of students in an introductory statistics class asked the following questions: (1) age; (2) do you like to sing? (Yes, No); (3) can you play a musical instrument (not at all, a little, pretty well); (4) how much did you spend on food last week? (5) height.

- (a) Classify each of these variables as categorical or quantitative and give reasons for your answers.
- (b) For each variable give the possible values.

## 1.12 What questions would you ask?

Refer to the previous exercise. Make up your own survey questions with at least six questions. Include at least two categorical variables and at least two quantitative variables. Tell which variables are categorical and which are quantitative. Give reasons for your answers. For each variable give the possible values.

## 1.13 How would you rate colleges?

Popular magazines rank colleges and universities on their “academic quality” in serving undergraduate students. Describe five variables that you would like to see measured for each college if you were choosing where to study. Give reasons for each of your choices.

### **1.14 Attending college in your state or in another state.**

The U.S. Census Bureau collects a large amount of information concerning higher education.<sup>1</sup> For example, the bureau provides a table that includes the following variables: state, number of students from the state who attend college, number of students who attend college in their home state.

- (a) What are the cases for this set of data?
- (b) Is there a label variable? If yes, what is it?
- (c) Identify each variable as categorical or quantitative.
- (d) Explain how you might use each of the quantitative variables to explain something about the states.
- (e) Consider a variable computed as the number of students in each state who attend college in the state divided by the total number of students from the state who attend college. Explain how you would use this variable explain something about the states.

### **1.15 Alcohol-impaired driving fatalities.**

A report on drunk-driving fatalities in the United States gives the number of alcohol-impaired driving fatalities for each state.<sup>2</sup> Discuss at least three different ways that these numbers could be converted to rates. Give the advantages and disadvantages of each.

## 1.2 Displaying Distributions with Graphs

When you complete this section, you will be able to

- Analyze the distribution of a categorical variable using a bar graph.
- Analyze the distribution of a categorical variable using a pie chart.
- Analyze the distribution of a quantitative variable using a stemplot.
- Analyze the distribution of a quantitative variable using a histogram.
- Examine the distribution of a quantitative variable with respect to the overall pattern of the data and deviations from that pattern.
- Identify the shape, center, and spread of the distribution of a quantitative variable.
- Identify and describe any outliers in the distribution of a quantitative variable.
- Use a time plot to describe the distribution of a quantitative variable that is measured over time.

Statistical tools and ideas help us examine data to describe their main features. This examination is called *exploratory data analysis*. Like an explorer crossing unknown lands, we want first to simply describe what we see. Here are two basic strategies that help us organize our exploration of a set of data:

### exploratory data analysis

- Begin by examining each variable by itself. Then move on to study the relationships among the variables.
- Begin with a graph or graphs. Then add numerical summaries of specific aspects of the data.

We will follow these principles in organizing our learning. This chapter presents methods for describing a single variable. We will study relationships among several variables in Chapter 2. Within each chapter, we will begin with graphical displays, then add numerical summaries for a more complete description.

### Categorical variables: bar graphs and pie charts

The values of a categorical variable are labels for the categories, such as “Yes” and “No.” The *distribution of a categorical variable* lists the categories and gives

either the **count** or the **percent** of cases that fall in each category.

distribution of a categorical variable

## EXAMPLE

### 1.7 How do you do online research?



A study of 552 first-year college students asked about their preferences for online resources. One question asked them to pick their favorite.<sup>3</sup> Here are the results:

Resource	Count ( <i>n</i> )
Google or Google Scholar	406
Library database or website	75
Wikipedia or online encyclopedia	52
Other	19
Total	552



Resource is the categorical variable in this example, and the values are the names of the online resources.

Note that the last value of the variable resource is “Other,” which includes all other online resources that were given as selection options. For data sets that have a large number of values for a categorical variable, we often create a category such as this that includes categories that have relatively small counts or percents.

Careful judgment is needed when doing this. You don't want to cover up some important piece of information contained in the data by combining data in this way.



## EXAMPLE

### 1.8 Favorites as percents.



When we look at the online resources data set, we see that Google is the clear winner. We see that 406 reported Google or Google Scholar as their favorite. To interpret this number, we need to know that the total number of students polled was 552. When we say that Google is the winner, we can describe this win by saying that 73.6% (406 divided by 552, expressed as a percent) of the students reported Google as their favorite. Here is a table of the preference percents:

Resource	Percent (%)
Google or Google Scholar	73.6
Library database or website	13.6
Wikipedia or online encyclopedia	9.4
Other	3.4
Total	100.0

The use of graphical methods will allow us to see this information and other characteristics of the data easily. We now examine two types of graph.

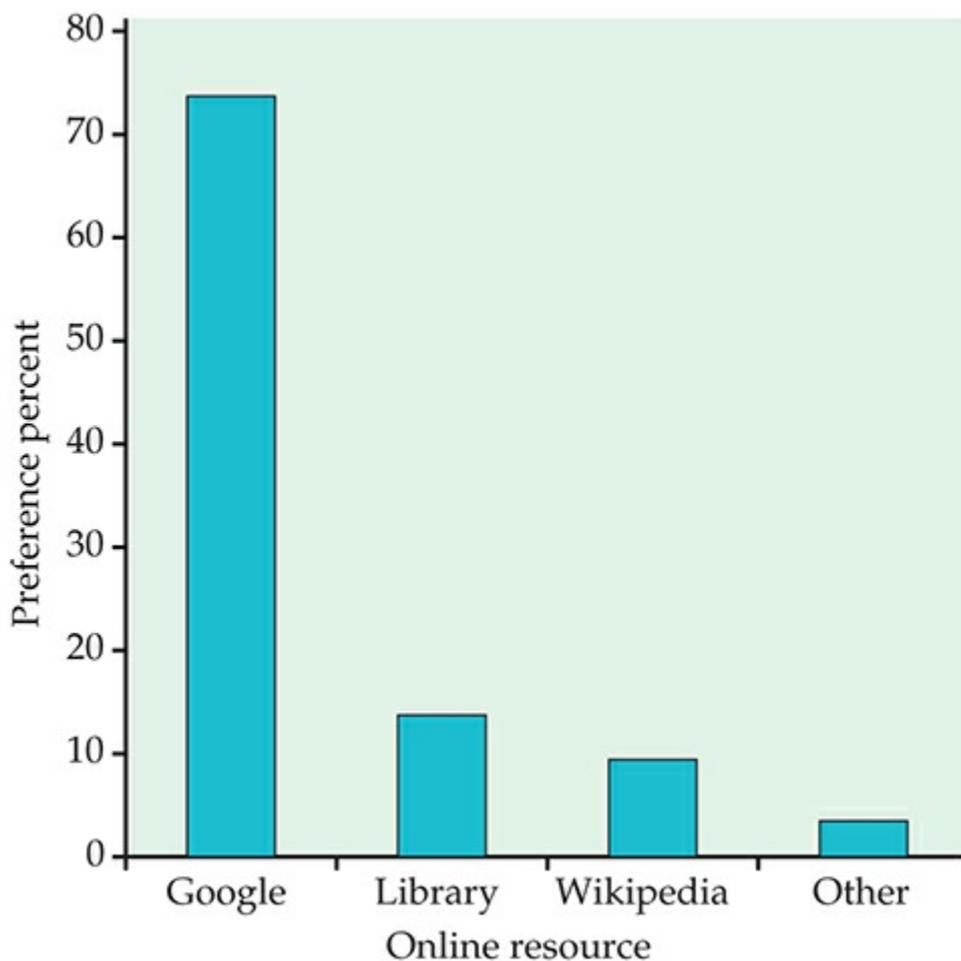
## EXAMPLE

## 1.9 Bar graph for the online resource preference data.



Figure 1.4 displays the online resource preference data using a *bar graph*. The heights of the four bars show the percents of the students who reported each of the resources as their favorite.

bar graph



**FIGURE 1.4**

Bar graph for the online resource preference data, for Example 1.9.

The categories in a bar graph can be put in any order. In Figure 1.4, we ordered the resources based on their preference percents. For other data sets, an alphabetical ordering or some other arrangement might produce a more useful graphical display.

You should always consider the best way to order the values of the categorical variable in a bar graph. Choose an ordering that will be useful to you. If you have

difficulty, ask a friend if your choice communicates what you expect.



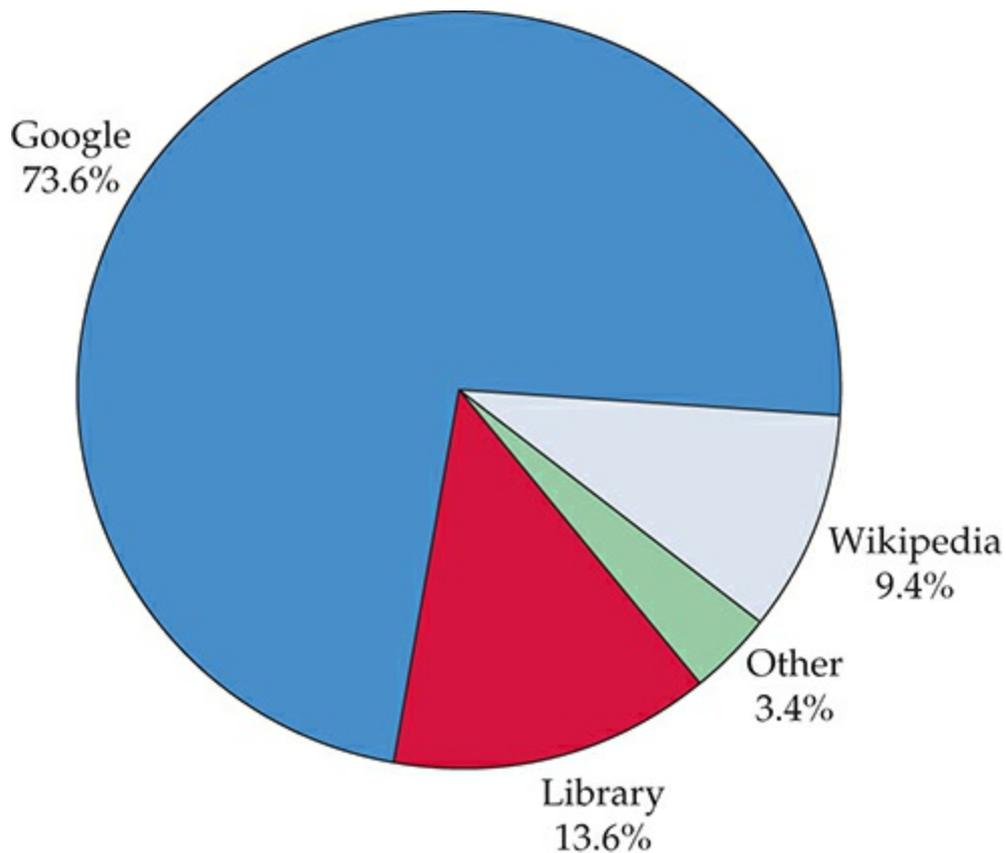
## EXAMPLE

### 1.10 Pie chart for the online resource preference data.



The *pie chart* in Figure 1.5 helps us see what part of the whole each group forms. Here it is very easy to see that Google is the favorite for about three-quarters of the students.

pie chart



**FIGURE 1.5**

Pie chart for the online resource preference data, for Example 1.10.

## USE YOUR KNOWLEDGE

### 1.16 Compare the bar graph with the pie chart.



ONLINE

Refer to the bar graph in Figure 1.4 and the pie chart in Figure 1.5 for the online resource preference data. Which graphical display does a better job of describing the data? Give reasons for your answer.

*To make a pie chart, you must include all the categories that make up a whole. A category such as “Other” in this example can be used, but the sum of the percents for all the categories should be 100%.*



This constraint makes bar graphs more flexible. For example, you can use a bar graph to compare the numbers of students at your college majoring in biology, business, and political science. A pie chart cannot make this comparison because not all students fall into one of these three majors.

## Quantitative variables: stemplots

A *stemplot* (also called a stem-and-leaf plot) gives a quick picture of the shape of a distribution while including the actual numerical values in the graph. Stemplots work best for small numbers of observations that are all greater than 0.

### STEMPLOT

To make a **stemplot**,

1. Separate each observation into a **stem** consisting of all but the final (rightmost) digit and a **leaf**, the final digit. Stems may have as many digits as needed, but each leaf contains only a single digit.
2. Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column.
3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.

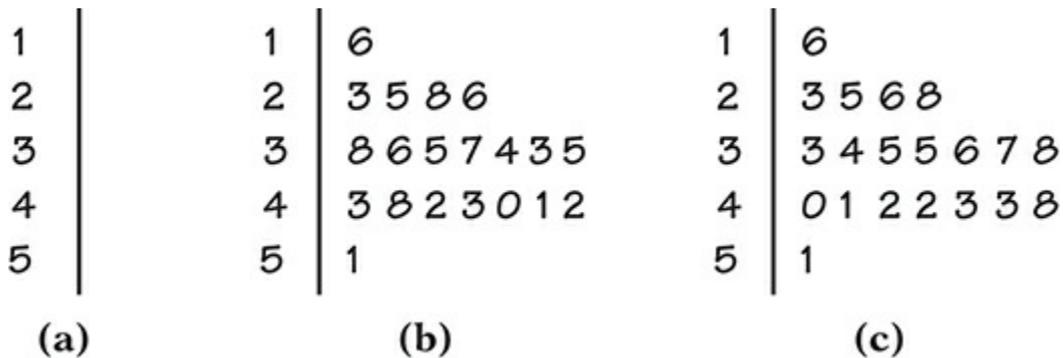
### EXAMPLE

#### 1.11 How much vitamin D do they have?



Your body needs vitamin D to use calcium when building bones. It is particularly important that young adolescents have adequate supplies of this vitamin because their bodies are growing rapidly. Vitamin D in the form 25-hydroxy vitamin D is measured in the blood and represents the stores available for the body to use. The units of measurement are nanograms per milliliter (ng/ml) of blood. Here are some values measured on a sample of 20 adolescent girls aged 11 to 14 years:<sup>4</sup>

16	43	38	48	42	23	36	35	37	34
25	28	26	43	51	33	40	35	41	42



**FIGURE 1.6**

Making a stemplot of the data in Example 1.11. (a) Write the stems. (b) Go through the data and write each leaf on the proper stem. For example, the values on the 2 stem are 23, 25, 28, and 26 in the order given in the display for the example. (c) Arrange the leaves on each stem in order out from the stem. The 2 stem now has leaves 3, 5, 6, and 8.

To make a stemplot of these data, use the first digits as stems and the second digits as leaves. Figure 1.6 shows the steps in making the plot. The girl with a measured value of 16 ng/ml for vitamin D appears on the first stem with a leaf of 6, while the girl with a measured value of 43 ng/ml appears on the stem labeled 4 with a leaf of 3.

The lowest value, 16 ng/ml, is somewhat far away from the next-highest value, 23. However, it is not particularly extreme.

## USE YOUR KNOWLEDGE

### 1.17 Make a stemplot.



Here are the scores on the first exam in an introductory statistics course for 30 students in one section of the course:

81	73	93	85	75	98	93	55	80	90	92	80	87	90	72
65	70	85	83	60	70	90	75	75	58	68	85	78	80	93

Use these data to make a stemplot. Then use the stemplot to describe the distribution of the first-exam scores for this course.

When you wish to compare two related distributions, a *back-to-back stemplot* with common stems is useful. The leaves on each side are ordered out from the common stem.

### back-to-back stemplot

## EXAMPLE

### 1.12 Vitamin D for boys.



Here are the 25-hydroxy vitamin D values for a sample of 20 adolescent boys aged 11 to 14 years:

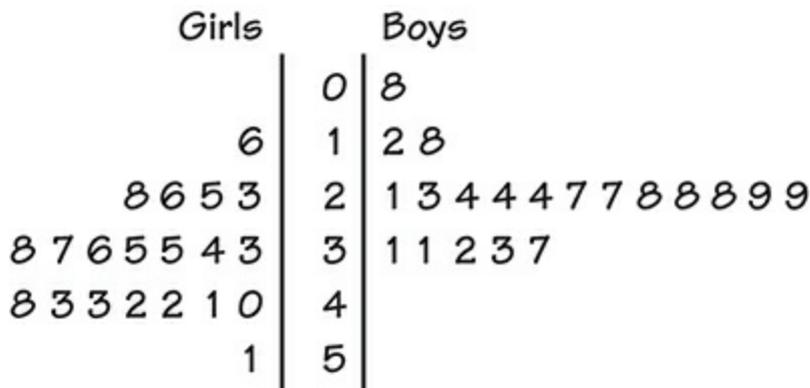
18	28	28	28	37	31	24	29	8	27
24	12	21	32	27	24	23	33	31	29

Figure 1.7 gives the back-to-back stemplot for the girls and the boys. The values on the left give the vitamin D measures for the girls, while the values on the right give the measures for the boys. The values for the boys tend to be lower than those for the girls.

There are two modifications of the basic stemplot that can be helpful in different situations. You can double the number of stems in a plot by *splitting each stem* into two: one with leaves 0 to 4 and the other with leaves 5 through 9. When the observed values have many digits, it is often best to *trim* the numbers by removing the last digit or digits before making a stemplot.

splitting stem

trimming



**FIGURE 1.7**

A back-to-back stemplot to compare the distributions of vitamin D for samples of adolescent girls and boys, for Example 1.12.

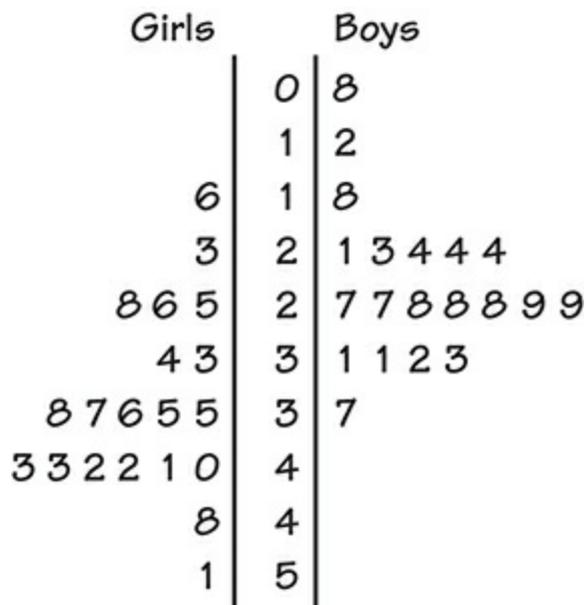
You must use your judgment in deciding whether to split stems and whether to trim, though statistical software will often make these choices for you. Remember that the purpose of a stemplot is to display the shape of a distribution. If there are many stems with no leaves or only one leaf, trimming will reduce the number of stems. Let's take a look at the effect of splitting the stems for our vitamin D data.

## EXAMPLE

### 1.13 Stemplot with split stems for vitamin D.



Figure 1.8 presents the data from Examples 1.11 and 1.12 in a stemplot with split stems. Notice that we needed only one stem for 0 because there are no values between 0 and 4.



**FIGURE 1.8**

A back-to-back stemplot with split stems to compare the distributions of vitamin D for samples of adolescent girls and boys, for Example 1.13.

## USE YOUR KNOWLEDGE

### 1.18 Which stemplot do you prefer?

Look carefully at the stemplots for the vitamin D data in Figures 1.7 and 1.8. Which do you prefer? Give reasons for your answer.

### 1.19 Why should you keep the space?

Suppose that you had a data set for girls similar to the one given in Example 1.11, but in which the observations of 33 ng/ml and 34 ng/ml were both changed to 35 ng/ml.

(a) Make a stemplot of these data for girls using split stems.

(b) Should you use one stem or two stems for the 30s? Give a reason for your answer. (*Hint:* How would your choice reveal or conceal a potentially important characteristic of the data?)

## Histograms

Stemplots display the actual values of the observations. This feature makes stemplots awkward for large data sets. Moreover, the picture presented by a

stemplot divides the observations into groups (stems) determined by the number system rather than by judgment.

Histograms do not have these limitations. A *histogram* breaks the range of values of a variable into classes and displays only the count or percent of the observations that fall into each class. You can choose any convenient number of classes, but you should always choose classes of equal width.

## histogram

145	139	126	122	125	130	96	110	118	118
101	142	134	124	112	109	134	113	81	113
123	94	100	136	109	131	117	110	127	124
106	124	115	133	116	102	127	117	109	137
117	90	103	114	139	101	122	105	97	89
102	108	110	128	114	112	114	102	82	101

Making a histogram by hand requires more work than a stemplot. Histograms do not display the actual values observed. For these reasons we prefer stemplots for small data sets.

The construction of a histogram is best shown by example. Most statistical software packages will make a histogram for you.

## EXAMPLE

### 1.14 Distribution of IQ scores.



You have probably heard that the distribution of scores on IQ tests is supposed to be roughly “bell-shaped.” Let’s look at some actual IQ scores. Table 1.1 displays the IQ scores of 60 fifth-grade students chosen at random from one school.

1. Divide the range of the data into classes of equal width. The scores in Table 1.1 range from 81 to 145, so we choose as our classes

			$75 \leq \text{IQ score} <$
85			$85 \leq \text{IQ score} <$
95			$\vdots$
			$145 \leq \text{IQ score} <$
155			

Be sure to specify the classes precisely so that each individual falls into exactly one class. A student with IQ 84 would fall into the first class, but IQ 85 falls into the second.

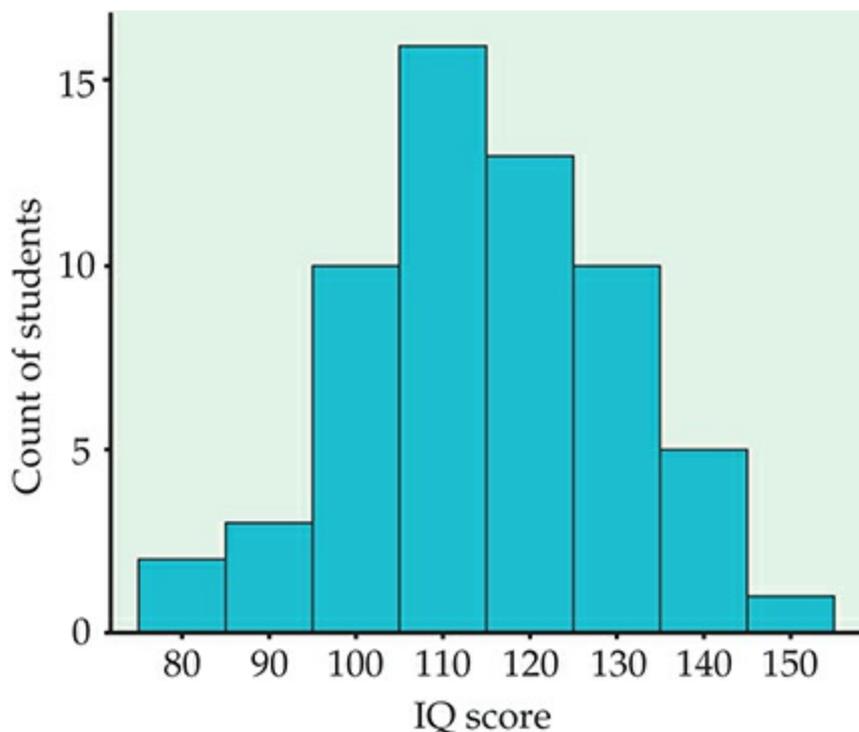
- Count the number of individuals in each class. These counts are called *frequencies*, frequency and a table of frequencies for all classes is a *frequency table*.

frequency

frequency table

Class	Count	Class	Count
$75 \leq \text{IQ score} < 85$	2	$115 \leq \text{IQ score} < 125$	13
$85 \leq \text{IQ score} < 95$	3	$125 \leq \text{IQ score} < 135$	10
$95 \leq \text{IQ score} < 105$	10	$135 \leq \text{IQ score} < 145$	5
$105 \leq \text{IQ score} < 115$	16	$145 \leq \text{IQ score} < 155$	1

- Draw the histogram. First, on the horizontal axis mark the scale for the variable whose distribution you are displaying. That's the IQ score. The scale runs from 75 to 155 because that is the span of the classes we chose. The vertical axis contains the scale of counts. Each bar represents a class. The base of the bar covers the class, and the bar height is the class count. There is no horizontal space between the bars unless a class is empty, so that its bar has height zero. Figure 1.9 is our histogram. It does look roughly "bell-shaped."



**FIGURE 1.9**

Histogram of the IQ scores of 60 fifth-grade students, for Example 1.14.

Large sets of data are often reported in the form of frequency tables when it is not practical to publish the individual observations. In addition to the frequency (count) for each class, we may be interested in the fraction or percent of the observations that fall in each class. A histogram of percents looks just like a frequency histogram such as Figure 1.9. Simply relabel the vertical scale to read in percents. *Use histograms of percents for comparing several distributions that have different numbers of observations.*

## USE YOUR KNOWLEDGE

### 1.20 Make a histogram.



Refer to the first-exam scores from Exercise 1.17 (page 14). Use these data to make a histogram with classes 50 to 59, 60 to 69, etc. Compare the histogram with the stemplot as a way of describing this distribution.

Which do you prefer for these data?

Our eyes respond to the *area* of the bars in a histogram. Because the classes are all the same width, area is determined by height, and all classes are fairly represented. There is no one right choice of the classes in a histogram. Too few classes will give a “skyscraper” graph, with all values in a few classes with tall bars. Too many will produce a “pancake” graph, with most classes having one or no observations. Neither choice will give a good picture of the shape of the distribution. You must use your judgment in choosing classes to display the shape. Statistical software will choose the classes for you. The software’s choice is often a good one, but you can change it if you want.



*You should be aware that the appearance of a histogram can change when you change the classes.* The histogram function in the *One-Variable Statistical Calculator* applet on the text website allows you to change the number of classes by dragging with the mouse, so that it is easy to see how the choice of classes affects the histogram.



## USE YOUR KNOWLEDGE

### 1.21 Change the classes in the histogram.

Refer to the first-exam scores from Exercise 1.17 (page 14) and the histogram that you produced in Exercise 1.20. Now make a histogram for these data using classes 40 to 59, 60 to 79, and 80 to 99. Compare this histogram with the one that you produced in Exercise 1.20. Which do you prefer? Give a reason for your answer.



## 1.22 Use smaller classes.

Repeat the previous exercise using classes 55 to 59, 60 to 64, 65 to 69, etc.

Although histograms resemble bar graphs, their details and uses are distinct. A histogram shows the distribution of counts or percents among the values of a single variable. A bar graph compares the counts of different items. The horizontal axis of a bar graph need not have any measurement scale but simply identifies the items being compared. Draw bar graphs with blank space between the bars to separate the items being compared. Draw histograms with no space, to indicate that all values of the variable are covered. *Some spreadsheet programs, which are not primarily intended for statistics, will draw histograms as if they were bar graphs, with space between the bars.* Often, you can tell the software to eliminate the space to produce a proper histogram.



## Data analysis in action: don't hang up on me

Many businesses operate call centers to serve customers who want to place an order or make an inquiry. Customers want their requests handled thoroughly. Businesses want to treat customers well, but they also want to avoid wasted time on the phone. They therefore monitor the length of calls and encourage their representatives to keep calls short.

### EXAMPLE

#### 1.15 How long are customer service center calls?

We have data on the lengths of all 31,492 calls made to the customer service center of a small bank in a month. Table 1.2 displays the lengths of the first 80 calls.<sup>5</sup>



## CALLS80

Take a look at the data in Table 1.2. In this data set the *cases* are calls made to the bank's call center. The *variable* recorded is the length of each call. The *units* are seconds. We see that the call lengths vary a great deal. The longest call lasted 2631 seconds, almost 44 minutes. More striking is that 8 of these 80 calls lasted less than 10 seconds. What's going on?

We started our study of the customer service center data by examining a few cases, the ones displayed in Table 1.2. It would be very difficult to examine all 31,492 cases in this way. How can we do this? Let's try a histogram.

## EXAMPLE

### 1.16 Histogram for customer service center call lengths.



Figure 1.10 is a histogram of the lengths of all 31,492 calls. We did not plot the few lengths greater than 1200 seconds (20 minutes). As expected, the graph shows that most calls last between about 1 and 5 minutes, with some lasting much longer when customers have complicated problems. More striking is the fact that 7.6% of all calls are no more than 10 seconds long. It turned out that the bank penalized representatives whose average call length was too long—so some representatives just hung up on customers to bring their average length down. Neither the customers nor the bank were happy about this. The bank changed its policy, and later data showed that calls under 10 seconds had almost disappeared.

TABLE 1.2

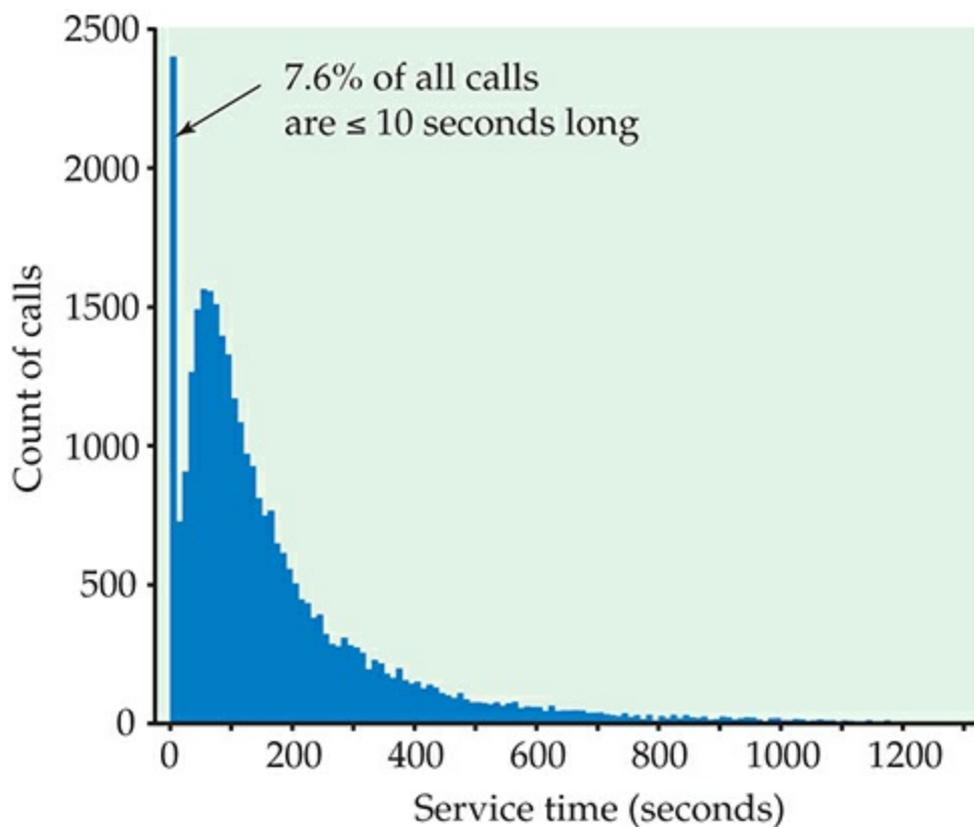
Service Times (Seconds) for Calls to a Customer Service Center

77	289	128	59	19	148	157	203
126	118	104	141	290	48	3	2
372	140	438	56	44	274	479	211
179	1	68	386	2631	90	30	57
89	116	225	700	40	73	75	51
148	9	115	19	76	138	178	76

67	102	35	80	143	951	106	55
4	54	137	367	277	201	52	9
700	182	73	199	325	75	103	64
121	11	9	88	1148	2	465	25

The extreme values of a distribution are in the *tails* of the distribution. The high values are in the upper, or right, tail, and the low values are in the lower, or left, tail. The overall pattern in Figure 1.10 is made up of the many moderate call lengths and the long right tail of more lengthy calls. The striking deviation from the overall pattern is the surprising number of very short calls in the left tail.

tails



**FIGURE 1.10**

The distribution of call lengths for 31,492 calls to a bank’s customer service center, for Example 1.16. The data show a surprising number of very short calls. These are mostly due to representatives deliberately hanging up in order to bring down their average call length.

Our examination of the call center data illustrates some important principles:

- After you understand the background of your data (cases, variables, units of measurement), the first thing to do is **plot** your data.
- When you look at a plot, look for an **overall pattern** and also for any **striking deviations** from the pattern.

## Examining distributions

Making a statistical graph is not an end in itself. The purpose of the graph is to help us understand the data. After you make a graph, always ask, “What do I see?” Once you have displayed a distribution, you can see its important features as follows.

### EXAMINING A DISTRIBUTION

In any graph of data, look for the **overall pattern** and for striking **deviations** from that pattern.

You can describe the overall pattern of a distribution by its **shape, center, and spread**.

An important kind of deviation is an **outlier**, an individual value that falls outside the overall pattern.

In Section 1.3, we will learn how to describe center and spread numerically. For now, we can describe the center of a distribution by its *midpoint*, the value with roughly half the observations taking smaller values and half taking larger values. We can describe the spread of a distribution by giving the *smallest and largest values*. Stemplots and histograms display the shape of a distribution in the same way. Just imagine a stemplot turned on its side so that the larger values lie to the right.

Some things to look for in describing shape are

- Does the distribution have one or several major peaks, called *modes*? A distribution with one major peak is called *unimodal*.

modes

unimodal

- Is it approximately symmetric or is it skewed in one direction? A distribution is *symmetric* if the values smaller and larger than its midpoint are mirror images of each other. It is *skewed to the right* if the right tail (larger values) is much longer than the left tail (smaller values).

symmetric

skewed

Some variables commonly have distributions with predictable shapes. Many biological measurements on specimens from the same species and sex—lengths of bird bills, heights of young women—have symmetric distributions. Money

amounts, on the other hand, usually have right-skewed distributions. There are many moderately priced houses, for example, but the few very expensive mansions give the distribution of house prices a strong right-skew.

## EXAMPLE

### 1.17 Examine the histogram of IQ scores.

What does the histogram of IQ scores (Figure 1.9, page 17) tell us?



**Shape:** The distribution is *roughly symmetric* with a *single peak* in the center. We don't expect real data to be perfectly symmetric, so in judging symmetry, we are satisfied if the two sides of the histogram are roughly similar in shape and extent.

**Center:** You can see from the histogram that the midpoint is not far from 110. Looking at the actual data shows that the midpoint is 114.

**Spread:** The histogram has a spread from 75 to 155. Looking at the actual data shows that the spread is from 81 to 145. There are no outliers or other strong deviations from the symmetric, unimodal pattern.

## EXAMPLE

### 1.18 Examine the histogram of call lengths.

The distribution of call lengths in Figure 1.10 (page 19), on the other hand, is strongly *skewed to the right*. The midpoint, the length of a typical call, is about 115 seconds, or just under 2 minutes. The spread is very large, from 1 second to 28,739 seconds.

The longest few calls are outliers. They stand apart from the long right tail of the distribution, though we can't see this from Figure 1.10, which omits the

largest observations. The longest call lasted almost 8 hours—that may well be due to equipment failure rather than an actual customer call.

## USE YOUR KNOWLEDGE

### 1.23 Describe the first-exam scores.



Refer to the first-exam scores from Exercise 1.17 (page 14). Use your favorite graphical display to describe the shape, the center, and the spread of these data. Are there any outliers?

## Dealing with outliers

In data sets smaller than the service call data, you can spot outliers by looking for observations that stand apart (either high or low) from the overall pattern of a histogram or stemplot. *Identifying outliers is a matter for judgment. Look for points that are clearly apart from the body of the data, not just the most extreme observations in a distribution. You should search for an explanation for any outlier.* Sometimes outliers point to errors made in recording the data. In other cases, the outlying observation may be caused by equipment failure or other unusual circumstances.



## EXAMPLE

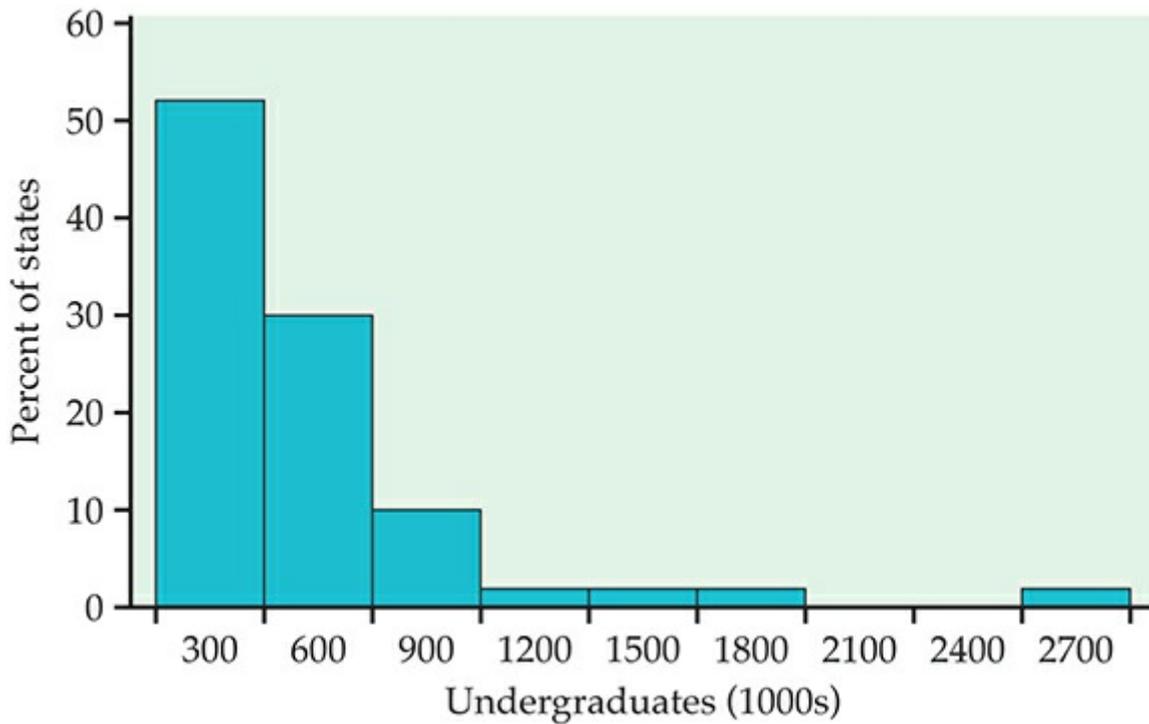
### 1.19 College students.



## COLLEGE

How does the number of undergraduate college students vary by state? Figure 1.11 is a histogram of the numbers of undergraduate students in each of the states.<sup>6</sup> Notice that over 50% of the states are included in the first bar of the histogram. These states have fewer than 300,000 undergraduates. The next bar includes another 30% of the states. These have between 300,000 and 600,000 students. The bar at the far right of the histogram corresponds to the state of California, which has 2,685,893 undergraduates. California certainly stands apart from the other states for this variable. It is an outlier.

The state of California is an outlier in the previous example because it has a very large number of undergraduate students. Since California has the largest population of all the states, we might expect it to have a large number of undergraduate students. Let's look at these data in a different way.



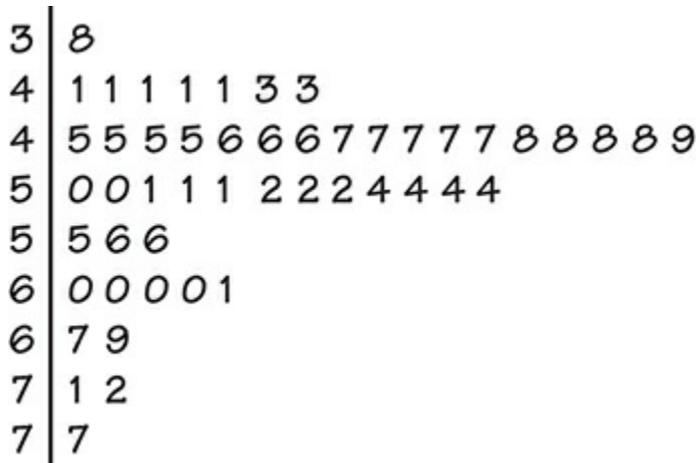
**FIGURE 1.11**

The distribution of the numbers of undergraduate college students for the 50 states, for Example 1.19.

## EXAMPLE

### 1.20 College students per 1000.

To account for the fact that there is large variation in the populations of the states, for each state we divide the number of undergraduate students by the population and then multiply by 1000. This gives the undergraduate college enrollment expressed as the number of students per 1000 people in each state. Figure 1.12 gives a stemplot of the distribution. California has 60 undergraduate students per 1000 people. This is one of the higher values in the distribution but it is clearly not an outlier.



**FIGURE 1.12**

Stemplot of the numbers of undergraduate college students per 1000 people in each of the 50 states, for Example 1.20.

## USE YOUR KNOWLEDGE

### 1.24 Four states with large populations.

There are four states with populations greater than 15 million.



COLLEGE

- (a) Examine the data file and report the names of these four states.

(b) Find these states in the distribution of number of undergraduate students per 1000 people. To what extent do these four states influence the distribution of number of undergraduate students per 1000 people?

In Example 1.19 we looked at the distribution of the number of undergraduate students, while in Example 1.20 we adjusted these data by expressing the counts as number per 1000 people in each state. Which way is correct? The answer depends upon why you are examining the data.

If you are interested in marketing a product to undergraduate students, the unadjusted numbers would be of interest. On the other hand, if you are interested in comparing states with respect to how well they provide opportunities for higher education to their residents, the population-adjusted values would be more suitable. *Always think about why you are doing a statistical analysis, and this will guide you in choosing an appropriate analytic strategy.*



Here is an example with a different kind of outlier.

## EXAMPLE

### 1.21 Healthy bones and PTH.

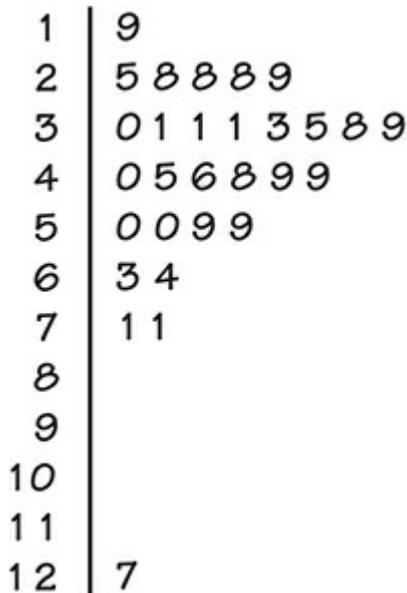


Bones are constantly being built up (bone formation) and torn down (bone resorption). Young people who are growing have more formation than resorption. When we age, resorption increases to the point where it exceeds formation. (The same phenomenon occurs when astronauts travel in space.) The result is osteoporosis, a disease associated with fragile bones that are more likely to break. The underlying mechanisms that control these processes are complex and involve a variety of substances. One of these is parathyroid hormone (PTH). Here are the values of PTH measured on a sample of 29 boys and girls aged 12 to 15 years:<sup>7</sup>

---

39 59 30 48 71 31 25 31 71 50 38 63 49 45 31  
 33 28 40 127 49 59 50 64 28 46 35 28 19 29

The data are measured in picograms per milliliter (pg/ml) of blood. The original data were recorded with one digit after the decimal point. They have been rounded to simplify our presentation here. Here is a stemplot of the data:



The observation 127 clearly stands out from the rest of the distribution. A PTH measurement on this individual taken on a different day was similar to the rest of the values in the data set. We conclude that this outlier was caused by a laboratory error or a recording error, and we are confident in discarding it for any additional analysis.

## Time plots

Whenever data are collected over time, it is a good idea to plot the observations in time order. *Displays of the distribution of a variable that ignore time order, such as stemplots and histograms, can be misleading when there is systematic change over time.*



### TIME PLOT

A **time plot** of a variable plots each observation against the time at which it was measured. Always put time on the horizontal scale of your plot and the

variable you are measuring on the vertical scale.

## EXAMPLE

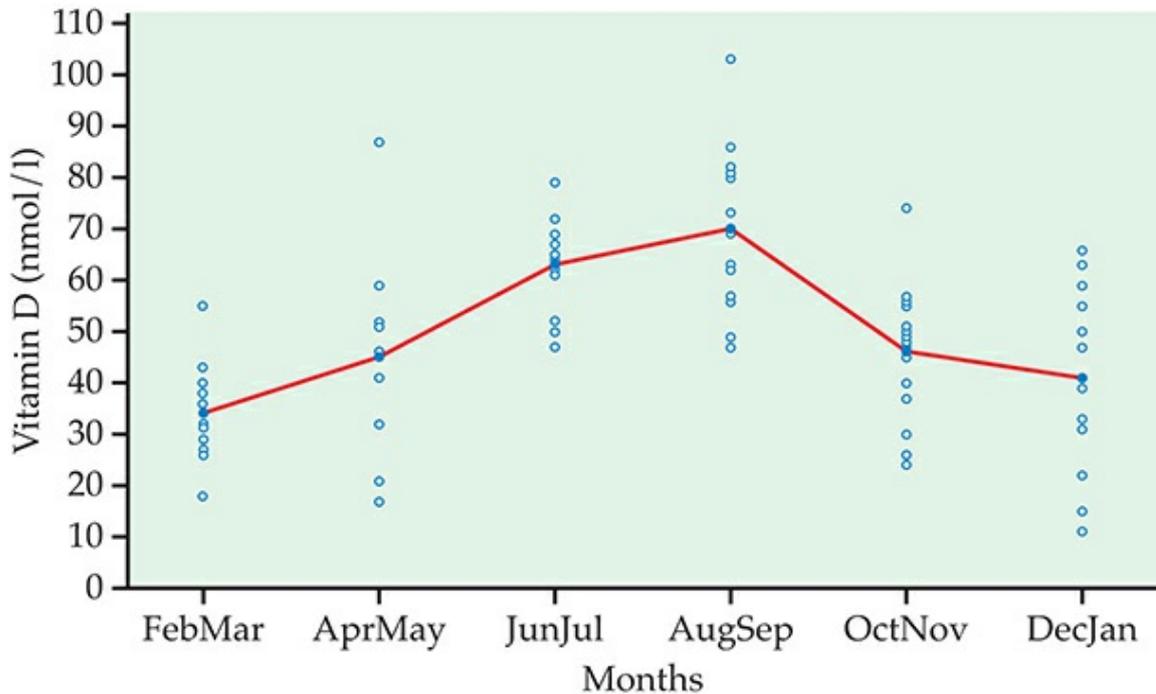
### 1.22 Seasonal variation in vitamin D.



Although we get some of our vitamin D from food, most of us get about 75% of what we need from the sun. Cells in the skin make vitamin D in response to sunlight. If people do not get enough exposure to the sun, they can become deficient in vitamin D, resulting in weakened bones and other health problems. The elderly, who need more vitamin D than younger people, and people who live in northern areas where there is relatively little sunlight in the winter, are particularly vulnerable to these problems.

Figure 1.13 is a plot of the serum levels of vitamin D versus time of year for samples of subjects from Switzerland.<sup>8</sup> The units for measuring vitamin D are nanomoles per liter (nmol/l) of blood. The observations are grouped into periods of two months for the plot. Means are marked by filled-in circles and are connected by a line in the plot. The effect of the lack of sunlight in the winter months on vitamin D levels is clearly evident in the plot.

The data described in the example above are based on a subset of the subjects in a study of 248 subjects. The researchers were particularly concerned about subjects whose levels were deficient, defined as a serum vitamin D level of less than 50 nmol/l. They found that there was a 3.8-fold higher deficiency rate in February–March than in August–September: 91.2% versus 24.3%. To ensure that individuals from this population have adequate levels of vitamin D, some form of supplementation is needed, particularly during certain times of the year.



**FIGURE 1.13**

Plot of vitamin D versus months of the year, for Example 1.22.

## SECTION 1.2 Summary

**Exploratory data analysis** uses graphs and numerical summaries to describe the variables in a data set and the relations among them.

The **distribution** of a variable tells us what values it takes and how often it takes these values.

**Bar graphs** and **pie charts** display the distributions of categorical variables. These graphs use the counts or percents of the categories.

**Stemplots** and **histograms** display the distributions of quantitative variables. Stemplots separate each observation into a **stem** and a one-digit **leaf**. Histograms plot the **frequencies** (counts) or the percents of equal-width classes of values.

When examining a distribution, look for **shape**, **center**, and **spread** and for clear **deviations** from the overall shape.

Some distributions have simple shapes, such as **symmetric** or **skewed**. The number of **modes** (major peaks) is another aspect of overall shape. Not all distributions have a simple overall shape, especially when there are few observations.

**Outliers** are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them.

When observations on a variable are taken over time, make a **time plot** that graphs time horizontally and the values of the variable vertically. A time plot can reveal changes over time.

## SECTION 1.2 Exercises

For Exercise 1.16, see page 12; for Exercise 1.17, see page 14; for Exercises 1.18 and 1.19, see page 15; for Exercise 1.20, see page 17; for Exercises 1.21 and 1.22, see page 18; for Exercise 1.23, see page 21; and for Exercise 1.24, see page 22.

### 1.25 The *Titanic* and class.

On April 15, 1912, on her maiden voyage, the *Titanic* collided with an iceberg and sank. The ship was luxurious but did not have enough lifeboats for the 2224 passengers and crew. As a result of the collision, 1502 people died.<sup>9</sup> The ship had three classes of passengers. The level of luxury and the price of the ticket varied with the class, with first class being the most luxurious. There were 323 passengers in first class, 277 in second class, and 709 in third class.<sup>10</sup>  TITANIC

- Make a bar graph of these data.
- Give a short summary of how the number of passengers varied with class.
- If you made a bar graph of the percents of passengers in each class, would the general features of the graph differ from the one you made in part (a)? Explain your answer.

### 1.26 Another look at the *Titanic* and class.

Refer to the previous exercise.  TITANIC

- Make a pie chart to display the data.
- Compare the pie chart with the bar graph. Which do you prefer? Give reasons for your answer.

### 1.27 Who survived?

Refer to the two previous exercises. The number of first-class passengers who survived was 200. For second and third class, the numbers were 119 and 181, respectively. Create a graphical summary that shows how the survival of passengers depended on class.  TITANIC

### 1.28 Do you use your Twitter account?

Although Twitter has more than 500,000,000 users, only about 170,000,000 are active. A study of Twitter account usage defined an active account as one with at least one message posted within a three-month period. Here are the percents of active accounts for 20 countries:<sup>11</sup>  TWITTC

Country	Percent	Country	Percent	Country	Percent
Argentina	25	India	19	South Korea	24
Brazil	25	Indonesia	28	Spain	29
Canada	28	Japan	30	Turkey	25
Chile	24	Mexico	26	United Kingdom	26
Colombia	26	Netherlands	33	United States	28
France	24	Philippines	22	Venezuela	28
Germany	23	Russia	26		

- (a) Make a stemplot of the distribution of percents of active accounts.
- (b) Describe the overall pattern of the data and any deviations from that pattern.
- (c) Identify the shape, center, and spread of the distribution.
- (d) Identify and describe any outliers.

### 1.29 Another look at Twitter account usage.

Refer to the previous exercise.  TWITTC

- (a) Use a histogram to summarize the distribution.
- (b) Use this histogram to answer parts (b), (c), and (d) of the previous exercise.
- (c) Which graphical display, stemplot or histogram, is more useful for describing this distribution? Give reasons for your answer.

### 1.30 Energy consumption.

The U.S. Energy Information Administration reports data summaries of various energy statistics. Let's look at the total amount of energy consumed, in quadrillions of British thermal units (Btu), for each month in 2011. Here are the data:<sup>12</sup>  ENERGY

Month	Energy (quadrillion Btu)	Month	Energy (quadrillion Btu)
January	9.33	July	8.41
February	8.13	August	8.43
March	8.38	September	7.58
April	7.54	October	7.61
May	7.61	November	7.81
June	7.92	December	8.60

- (a) Look at the table and describe how the energy consumption varies from month to month.
- (b) Make a time plot of the data and describe the patterns.
- (c) Suppose you wanted to communicate information about the month-to-month variation in energy consumption. Which would be more effective, the table of the data or the graph? Give reasons for your answer.

### 1.31 Energy consumption in a different year.

Refer to the previous exercise. Here are the data for 2010:  ENERGY

Month	Energy (quadrillion Btu)	Month	Energy (quadrillion Btu)
January	9.13	July	8.38
February	8.21	August	8.44
March	8.21	September	7.69
April	7.37	October	7.51
May	7.68	November	7.80

- (a) Analyze these data using the questions in the previous exercise as a guide.
- (b) Compare the patterns in 2010 with those in 2011. Describe any similarities and differences.

### 1.32 Favorite colors.

What is your favorite color? One survey produced the following summary of responses to that question: blue, 42%; green, 14%; purple, 14%; red, 8%; black, 7%; orange, 5%; yellow, 3%; brown, 3%; gray, 2%; and white, 2%.<sup>13</sup> Make a bar graph of the percents and write a short summary of the major features of your graph.  **FAVCOL**

### 1.33 Least-favorite colors.

Refer to the previous exercise. The same study also asked people about their least-favorite color. Here are the results: orange, 30%; brown, 23%; purple, 13%; yellow, 13%; gray, 12%; green, 4%; white, 4%; red, 1%; black, 0%; and blue, 0%. Make a bar graph of these percents and write a summary of the results.  **LFACOL**

### 1.34 Garbage.

The formal name for garbage is “municipal solid waste.” Here is a breakdown of the materials that make up American municipal solid waste.<sup>14</sup>  **GARBAGE**

Material	Weight (million tons)	Percent of total (%)
Food scraps	34.8	13.9
Glass	11.5	4.6
Metals	22.4	9.0
Paper, paperboard	71.3	28.5
Plastics	31.0	12.4
Rubber, leather, textiles	20.9	8.4
Wood	15.9	6.4
Yard trimmings	33.4	13.4
Other	8.6	3.2
Total	249.6	100.0

- (a) Add the weights and then the percents for the nine types of material given, including “Other.” Each entry, including the total, is separately rounded to the nearest tenth. So the sum and the total may slightly because of **roundoff error**.
- (b) Make a bar graph of the percents. The graph gives a clearer picture of the main contributors to garbage if you order the bars from tallest to shortest.
- (c) Make a pie chart of the percents. Compare the advantages and disadvantages of each graphical summary. Which do you prefer? Give reasons for your answer.

### 1.35 Recycled garbage.

Refer to the previous exercise. The following table gives the percent of the weight that was recycled for each of the categories.  **GARBAGE**

Material	Weight (million tons)	Percent recycled (%)
Food scraps	34.8	2.8
Glass	11.5	27.1
Metals	22.4	35.1
Paper, paperboard	71.3	62.5
Plastics	31.0	8.2
Rubber, leather, textiles	20.9	15.0
Wood	15.9	14.5
Yard trimmings	33.4	57.5
Other	8.6	16.3
Total	249.6	

- (a) Use a bar graph to display the percent recycled for these materials. Use the order of the materials given in the table above.
- (b) Make another bar graph where the materials are ordered by the percent recycled, largest percent to smallest percent.
- (c) Which bar graph, (a) or (b), do you prefer? Give a reason for your answer.
- (d) Explain why it is inappropriate to use a pie chart to display these data.

### 1.36 Market share for desktop browsers.

The following table gives the market share for the browsers used on desktop computers.<sup>15</sup>  **BROWSED**

Search engine	Market share (%)	Search engine	Market share (%)
Internet Explorer	54.76	Internet Explorer	5.33
Firefox	20.44	Opera	1.67
Chrome	17.24	Other	0.56

- (a) Use a bar graph to display the market shares.
- (b) Use a pie chart to display the market shares.
- (c) Summarize what these graphical summaries tell you about market shares for browsers on desktops.
- (d) Which graphical display do you prefer? Give reasons for your answer.

### 1.37 Market share for mobiles and tablet browsers.

The following table gives the market share for the browsers used on mobiles and tablets.  **BROWSEM**

Search engine	Market share (%)	Search engine	Market share (%)
Safari	61.50	Chrome	1.14
Android	26.09	Blackberry	1.09

- (a) Use a bar graph to display the market shares.
- (b) Use a pie chart to display the market shares.
- (c) Summarize what these graphical summaries tell you about market shares for browsers on mobiles and tablets.
- (d) Which graphical display do you prefer? Give reasons for your answer.

### 1.38 Compare the market shares for browsers.

Refer to the previous two exercises. Using the analyses that you have done for browsers for desktops and browsers for mobiles and tablets, write a short report comparing the market shares for these two types of devices.  **BROUSED, BROWSEM**

### 1.39 Vehicle colors.

Vehicle colors differ among regions of the world. Here are data on the most popular colors for vehicles in North America and Europe: <sup>16</sup>  **VCOLORS**

Color	North America (%)	Europe (%)
White	23	20
Black	18	25
Silver	16	15
Gray	13	18
Red	10	6
Blue	9	7
Brown/beige	5	5
Yellow/gold	3	1
Other	3	3

- (a) Make a bar graph for the North America percents.
- (b) Make a bar graph for the Europe percents.
- (c) Now, be creative: make one bar graph that compares the two regions as well as the colors. Arrange your graph so that it is easy to compare the two regions.

### 1.40 Facebook users by region.

The following table gives the numbers of Facebook users by region of the world as of November 2012: <sup>17</sup>  **FACER**

Region	Facebook users (in millions)	Region	Facebook users (in millions)
Africa	40	Middle East	20
Asia	195	North America	173
Caribbean	6	Oceania/Australia	14

Central America	41	South America	113
Europe	233		

- (a) Use a bar graph to describe these data.
- (b) Describe the major features of your graph in a short paragraph.

### 1.41 Facebook ratios.

One way to compare the numbers of Facebook users for different regions of the world is to take into account the populations of these regions. The market penetration for a product is the number of users divided by the number of potential users, expressed as a percent. For Facebook, we use the population as the number of potential users. Here are estimates of the populations in 2012 of the same geographic regions that we studied in the previous exercise.<sup>18</sup>  FACER

Region	Population (in millions)	Region	Population (in millions)
Africa	1026	Middle East	213
Asia	3900	North America	347
Caribbean	39	Oceania/Australia	36
Central America	155	South America	402
Europe	818		

- (a) Compute the market penetration for each region by dividing the number of users from the previous exercise by the population size given in this exercise. Multiply these ratios by 100 to make the ratios similar to percents, and make a table of the results. Use the values in this table to answer the remaining parts of this exercise.
- (b) Carefully examine your table, and summarize what it shows. Are there any extreme outliers? Which ones would you classify in this way?
- (c) Use a stemplot to describe these data. You can list any extreme outliers separately from the plot.
- (d) Describe the major features of these data using your plot and your list of outliers.
- (e) How effective is the stemplot for summarizing these data? Give reasons for your answer.
- (f) Explain why the values in the table that you constructed in part (a) are not the same as the percents of the population in each region who are users.

### 1.42 Sketch a skewed distribution.

Sketch a histogram for a distribution that is skewed to the left. Suppose that you and your friends emptied your pockets of coins and recorded the year marked on each coin. The distribution of dates would be skewed to the left. Explain why.

### 1.43 Grades and self-concept.

Table 1.3 presents data on 78 seventh-grade students in a rural midwestern school.<sup>19</sup> The researcher was interested in the relationship between the students' "self-concept" and their academic performance. The data we give here include each student's grade point average (GPA), score on a standard IQ test, and gender, taken from school records. Gender is coded as F for female and M for male. The students are identified only by an observation number (OBS). The missing OBS numbers show that some students

dropped out of the study. The final variable is each student's score on the Piers-Harris Children's Self-Concept Scale, a psychological test administered by the researcher. 

- (a) How many variables does this data set contain? Which are categorical variables and which are quantitative variables?
- (b) Make a stemplot of the distribution of GPA, after rounding to the nearest tenth of a point.
- (c) Describe the shape, center, and spread of the GPA distribution. Identify any suspected outliers from the overall pattern.
- (d) Make a back-to-back stemplot of the rounded GPAs for female and male students. Write a brief comparison of the two distributions.

### 1.44 Describe the IQ scores.

Make a graph of the distribution of IQ scores for the seventh-grade students in Table 1.3. Describe the shape, center, and spread of the distribution, as well as any outliers. IQ scores are usually said to be centered at 100. Is the midpoint for these students close to 100, clearly above, or clearly below?   
**SEVENGR**

### 1.45 Describe the self-concept scores.

Based on a suitable graph, briefly describe the distribution of self-concept scores for the students in Table 1.3. Be sure to identify any suspected outliers.   
**SEVENGR**

**TABLE 1.3** Educational Data for 78 Seventh-Grade Students

OBS	GPA	IQ	Gender	Selfconcept	OBS	GPA	IQ	Gender	Selfconcept
001	7.940	111	M	67	043	10.760	123	M	64
002	8.292	107	M	43	044	9.763	124	M	58
003	4.643	100	M	52	045	9.410	126	M	70
004	7.470	107	M	66	046	9.167	116	M	72
005	8.882	114	F	58	047	9.348	127	M	70
006	7.585	115	M	51	048	8.167	119	M	47
007	7.650	111	M	71	050	3.647	97	M	52
008	2.412	97	M	51	051	3.408	86	F	46
009	6.000	100	F	49	052	3.936	102	M	66
010	8.833	112	M	51	053	7.167	110	M	67
011	7.470	104	F	35	054	7.647	120	M	63
012	5.528	89	F	54	055	0.530	103	M	53
013	7.167	104	M	54	056	6.173	115	M	67
014	7.571	102	F	64	057	7.295	93	M	61
015	4.700	91	F	56	058	7.295	72	F	54
016	8.167	114	F	69	059	8.938	111	F	60
017	7.822	114	F	55	060	7.882	103	F	60
018	7.598	103	F	65	061	8.353	123	M	63
019	4.000	106	M	40	062	5.062	79	M	30

020	6.231	105	F	66	063	8.175	119	M	54
021	7.643	113	M	55	064	8.235	110	M	66
022	1.760	109	M	20	065	7.588	110	M	44
024	6.419	108	F	56	068	7.647	107	M	49
026	9.648	113	M	68	069	5.237	74	F	44
027	10.700	130	F	69	071	7.825	105	M	67
028	10.580	128	M	70	072	7.333	112	F	64
029	9.429	128	M	80	074	9.167	105	M	73
030	8.000	118	M	53	076	7.996	110	M	59
031	9.585	113	M	65	077	8.714	107	F	37
032	9.571	120	F	67	078	7.833	103	F	63
033	8.998	132	F	62	079	4.885	77	M	36
034	8.333	111	F	39	080	7.998	98	F	64
035	8.175	124	M	71	083	3.820	90	M	42
036	8.000	127	M	59	084	5.936	96	F	28
037	9.333	128	F	60	085	9.000	112	F	60
038	9.500	136	M	64	086	9.500	112	F	70
039	9.167	106	M	71	087	6.057	114	M	51
040	10.140	118	F	72	088	6.057	93	F	21
041	9.999	119	F	54	089	6.938	106	M	56

## 1.46 The Boston Marathon.

Women were allowed to enter the Boston Marathon in 1972. Here are the times (in minutes, rounded to the nearest minute) for the winning women from 1972 to 2012:

Year	Time	Year	Time	Year	Time	Year	Time
1972	190	1983	143	1994	142	2005	145
1973	186	1984	149	1995	145	2006	143
1974	167	1985	154	1996	147	2007	149
1975	162	1986	145	1997	146	2008	145
1976	167	1987	146	1998	143	2009	152
1977	168	1988	145	1999	143	2010	146
1978	165	1989	144	2000	146	2011	142
1979	155	1990	145	2001	144	2012	151
1980	154	1991	144	2002	141		
1981	147	1992	144	2003	145		
1982	150	1993	145	2004	144		

Make a graph that shows change over time. What overall pattern do you see? Have times stopped improving in recent years? If so, when did improvement end? 

## 1.3 Describing Distributions with Numbers

When you complete this section, you will be able to

- Describe the center of a distribution by using the mean.
- Describe the center of a distribution by using the median.
- Compare the mean and the median as measures of center for a particular set of data.
- Describe the spread of a distribution by using quartiles.
- Describe a distribution by using the five-number summary.
- Describe a distribution by using a boxplot and a modified boxplot.
- Compare one or more sets of data measured on the same variable by using side-by-side boxplots.
- Identify outliers by using the  $1.5 \times IQR$  rule.
- Describe the spread of a distribution by using the standard deviation.
- Choose measures of center and spread for a particular set of data.
- Compute the effects of a linear transformation on the mean, the median, the standard deviation, and the interquartile range.

We can begin our data exploration with graphs, but numerical summaries make our analysis more specific. A brief description of a distribution should include its *shape* and numbers describing its *center* and *spread*. We describe the shape of a distribution based on inspection of a histogram or a stemplot. Now we will learn specific ways to use numbers to measure the center and spread of a distribution. We can calculate these numerical measures for any quantitative variable. But to interpret measures of center and spread, and to choose among the several measures we will learn, you must think about the shape of the distribution and the meaning of the data. The numbers, like graphs, are aids to understanding, not “the answer” in themselves.

---

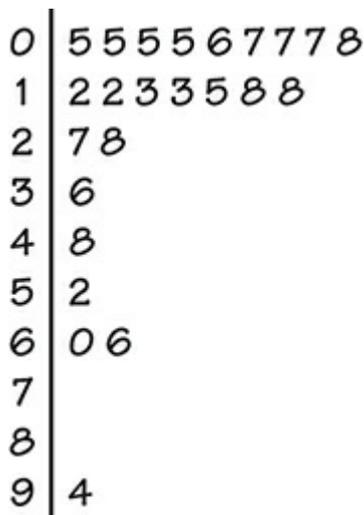
### EXAMPLE

### 1.23 The distribution of business start times.



An entrepreneur faces many bureaucratic and legal hurdles when starting a new business. The World Bank collects information about starting businesses throughout the world. They have determined the time, in days, to complete all the procedures required to start a business.<sup>20</sup> Data for 184 countries are included in the data file TIME. In this section we will examine data for a sample of 24 of these countries. Here are the data (start times, in days):

13	66	36	12	8	27	6	7	5	7	52	48
15	7	12	94	28	5	13	60	5	5	18	18



**FIGURE 1.14**

Stemplot for the sample of 24 business start times, for Example 1.23.

The stemplot in Figure 1.14 shows us the *shape*, *center*, and *spread* of the business start times. The stems are tens of days and the leaves are days. The distribution is highly skewed to the right. The largest value, 94, is separated from the rest of the distribution. We could consider this observation to be an outlier, but it appears to be part of a very long right tail. The values range from 5 to 94 days with a center somewhere around 10.

### Measuring center: the mean

Numerical description of a distribution begins with a measure of its center or average. The two common measures of center are the *mean* and the *median*. The mean is the “average value” and the median is the “middle value.” These are two

different ideas for “center,” and the two measures behave differently. We need precise recipes for the mean and the median.

## THE MEAN $\bar{x}$

To find the **mean**  $\bar{x}$  of a set of observations, add their values and divide by the number of observations. If the  $n$  observations are  $x_1, x_2, \dots, x_n$ , their mean is

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

or, in more compact notation,

$$\bar{x} = \frac{1}{n} \sum x_i$$

The  $\Sigma$  (capital Greek sigma) in the formula for the mean is short for “add them all up.” The bar over the  $x$  indicates the mean of all the  $x$ -values. Pronounce the mean  $\bar{x}$  as “x-bar.” This notation is so common that writers who are discussing data use  $\bar{x}$ ,  $\bar{y}$ , etc. without additional explanation. The subscripts on the observations  $x_i$  are a way of keeping the  $n$  observations separate.

## EXAMPLE

### 1.24 Mean time to start a business.



TIME24

The mean time to start a business is

$$\begin{aligned}\bar{x} &= \frac{x_1 + x_2 + \dots + x_n}{n} \\ &= \frac{13 + 66 + \dots + 1824}{78} \\ &= 23.625\end{aligned}$$

The mean time to start a business for the 24 countries in our data set is 23.6 days. Note that we have rounded the answer. Our goal is to use the mean to

describe the center of a distribution; it is not to demonstrate that we can compute with great accuracy. The extra digits do not provide any additional useful information. In fact, they distract our attention from the important digits that are meaningful. Do you think it would be better to report the mean as 24 days?

The value of the mean will not necessarily be equal to the value of one of the observations in the data set. Our example of time to start a business illustrates this fact.

## USE YOUR KNOWLEDGE

### 1.47 Include the outlier.

The complete business start time data set with 184 countries has a few with very large start times. In constructing the data set for Example 1.23, a random sample of 25 countries was selected. This sample included the South American country of Suriname, where the start time is 694 days. This country was deleted for Example 1.23. Reconstruct the original random sample by including Suriname. Show that the mean has increased to 50 days. (This is a rounded number. You should report the mean with one digit after the decimal.) The effect of the outlier is to more than double the mean.



### 1.48 Find the mean.



Here are the scores on the first exam in an introductory statistics course for 10 students:

81 73 93 85 75 98 93 55 80 90

Find the mean first-exam score for these students.

Exercise 1.47 illustrates an important weakness of the mean as a measure of center: *the mean is sensitive to the influence of a few extreme observations*. These may be outliers, but a skewed distribution that has no outliers will also pull the mean toward its long tail. Because the mean cannot resist the influence of extreme observations, we say that it is not a **resistant measure** of center.

resistant measure



A measure that is resistant does more than limit the influence of outliers. Its value does not respond strongly to changes in a few observations, no matter how large those changes may be. The mean fails this requirement because we can make the mean as large as we wish by making a large enough increase in just one observation. A resistant measure is sometimes called a **robust measure**.

robust measure

## Measuring center: the median

We used the midpoint of a distribution as an informal measure of center in Section 1.2. The *median* is the formal version of the midpoint, with a specific rule for calculation.

### THE MEDIAN $M$

The **median  $M$**  is the midpoint of a distribution. Half the observations are smaller than the median, and the other half are larger than the median. Here is a rule for finding the median:

1. Arrange all observations in order of size, from smallest to largest.
2. If the number of observations  $n$  is odd, the median  $M$  is the center observation in the ordered list. Find the location of the median by counting  $(n + 1)/2$  observations up from the bottom of the list.
3. If the number of observations  $n$  is even, the median  $M$  is the mean of the two center observations in the ordered list. The location of the median is again  $(n + 1)/2$  from the bottom of the list.

*Note that the formula  $(n + 1)/2$  does not give the median, just the location of the*

*median in the ordered list.* Medians require little arithmetic, so they are easy to find by hand for small sets of data. Arranging even a moderate number of observations in order is tedious, however, so that finding the median by hand for larger sets of data is unpleasant. Even simple calculators have an  $\bar{x}$  button, but you will need computer software or a graphing calculator to automate finding the median.



## EXAMPLE

### 1.25 Median time to start a business.

To find the median time to start a business for our 24 countries, we first arrange the data in order from smallest to largest.



5	5	5	5	6	7	7	7	8	12	12	13
13	15	18	18	27	28	36	48	52	60	66	94

The count of observations  $n = 24$  is even. The median, then, is the average of the two center observations in the ordered list. To find the location of the center observations, we first compute

$$\text{location of } M = n + 12 = 252 = 12.5$$

Therefore, the center observations are the 12th and 13th observations in the ordered list. The median is

$$M = 13 + 132 = 13$$

Note that you can use the stemplot in Figure 1.14 directly to compute the median. In the stemplot the cases are already ordered and you simply need to count from the top or the bottom to the desired location.

## USE YOUR KNOWLEDGE

### 1.49 Include the outlier.

Include Suriname, where the start time is 694 days, in the data set, and show that the median is 13 days. Note that with this case included, the sample size is now 25 and the median is the 13th observation in the ordered list. Write out the ordered list and circle the outlier. Describe the effect of the outlier on the median for this set of data.



### 1.50 Calls to a customer service center.

The service times for 80 calls to a customer service center are given in Table 1.2 (page 19). Use these data to compute the median service time.



### 1.51 Find the median.



Here are the scores on the first exam in an introductory statistics course for 10 students:

81 73 93 85 75 98 93 55 80 90

Find the median first-exam score for these students.

## Mean versus median

Exercises 1.47 and 1.49 illustrate an important difference between the mean and the median. Suriname is an outlier. It pulls the mean time to start a business up

from 24 days to 54 days. The median remained at 13 days.

The median is more *resistant* than the mean. If the largest start time in the data set was 1200 days, the median for all 25 countries would still be 13 days. The largest observation just counts as one observation above the center, no matter how far above the center it lies. The mean uses the actual value of each observation and so will chase a single large observation upward.

The best way to compare the response of the mean and median to extreme observations is to use an interactive applet that allows you to place points on a line and then drag them with your computer's mouse. Exercises 1.85 to 1.87 use the *Mean and Median* applet on the website for this book, [whfreeman.com/ips8e](http://whfreeman.com/ips8e), to compare the mean and the median.



The median and mean are the most common measures of the center of a distribution. The mean and median of a symmetric distribution are close together. If the distribution is exactly symmetric, the mean and median are exactly the same. In a skewed distribution, the mean is farther out in the long tail than is the median.

The endowment for a college or university is money set aside and invested. The income from the endowment is usually used to support various programs. The distribution of the sizes of the endowments of colleges and universities is strongly skewed to the right. Most institutions have modest endowments, but a few are very wealthy. The median endowment of colleges and universities in a recent year was \$93 million—but the mean endowment was \$498 million.<sup>21</sup> The few wealthy institutions pull the mean up but do not affect the median. *Don't confuse the "average" value of a variable (the mean) with its "typical" value, which we might describe by the median.*



We can now give a better answer to the question of how to deal with outliers in data. First, look at the data to identify outliers and investigate their causes. You can then correct outliers if they are wrongly recorded, delete them for good reason, or otherwise give them individual attention. The outlier in Example 1.21 (page 23) can be dropped from the data once we discover that it is an error. If you have no clear reason to drop outliers, you may want to use resistant methods in your analysis, so that outliers have little influence over your conclusions. The choice is often a matter for judgment.

## Measuring spread: the quartiles

A measure of center alone can be misleading. Two countries with the same median family income are very different if one has extremes of wealth and poverty and the other has little variation among families. A drug manufactured with the correct mean concentration of active ingredient is dangerous if some batches are much too high and others much too low.

We are interested in the *spread* or *variability* of incomes and drug potencies as well as their centers. **The simplest useful numerical description of a distribution consists of both a measure of center and a measure of spread.**

We can describe the spread or variability of a distribution by giving several percentiles. The median divides the data in two; half of the observations are above the median and half are below the median. We could call the median the 50th percentile. The upper *quartile* is the median of the upper half of the data. Similarly, the lower quartile is the median of the lower half of the data. With the median, the quartiles divide the data into four equal parts; 25% of the data are in each part.

### quartile

We can do a similar calculation for any percent. The *p*th percentile of a distribution is the value that has *p* percent of the observations fall at or below it. To calculate a percentile, arrange the observations in increasing order and count up the required percent from the bottom of the list.

### percentile

Our definition of percentiles is a bit inexact because there is not always a value with exactly *p* percent of the data at or below it. We will be content to take the nearest observation for most percentiles, but the quartiles are important enough to require an exact rule.

## THE QUARTILES $Q_1$ AND $Q_3$

To calculate the quartiles:

1. Arrange the observations in increasing order and locate the median  $M$  in the ordered list of observations.
2. The **first quartile**  $Q_1$  is the median of the observations whose positions in the ordered list are to the left of the location of the overall median.
3. The **third quartile**  $Q_3$  is the median of the observations whose positions in the ordered list are to the right of the location of the overall median.

Here is an example.

## EXAMPLE

### 1.26 Finding the quartiles.



Here is the ordered list of the times to start a business in our sample of 24 countries:

5	5	5	5	6	7	7	7	8	12	12	13
13	15	18	18	27	28	36	48	52	60	66	94

The count of observations  $n = 24$  is even, so the median is at position  $(24 + 1)/2 = 12.5$ , that is, between the 12th and the 13th observation in the ordered list. There are 12 cases above this position and 12 below it. The first quartile is the median of the first 12 observations, and the third quartile is the median of the last 12 observations. Check that  $Q_1 = 7$  and  $Q_3 = 32$ .

Notice that the quartiles are resistant. For example,  $Q_3$  would have the same value if the highest start time was 940 days rather than 94 days.



*Be careful when several observations take the same numerical value. Write down all the observations and apply the rules just as if they all had distinct values.*

## USE YOUR KNOWLEDGE

### 1.52 Find the quartiles.



Here are the scores on the first exam in an introductory statistics course for 10 students:

81 73 93 85 75 98 93 55 80 90

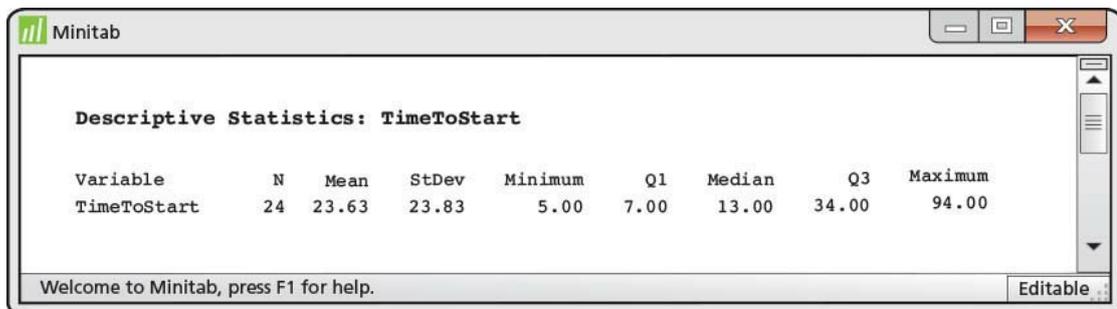
Find the quartiles for these first-exam scores.

## EXAMPLE

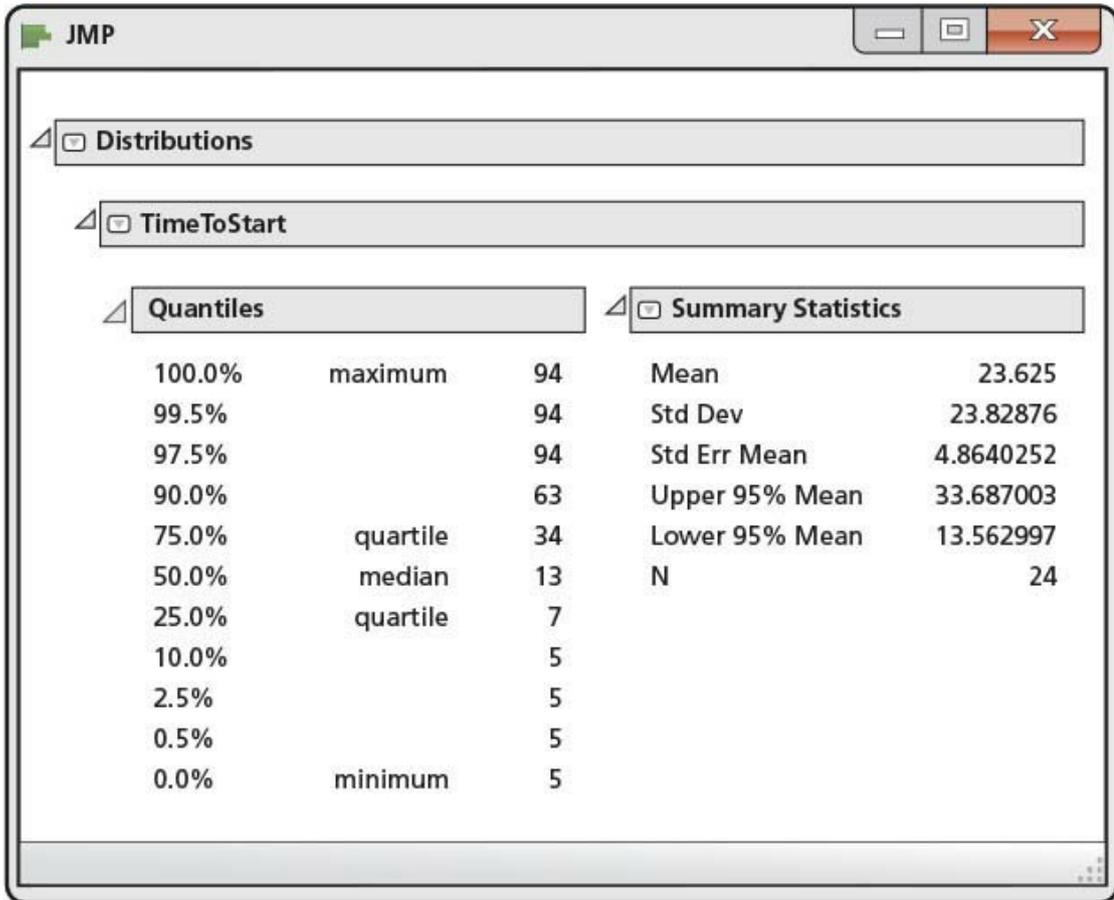
### 1.27 Results from software.



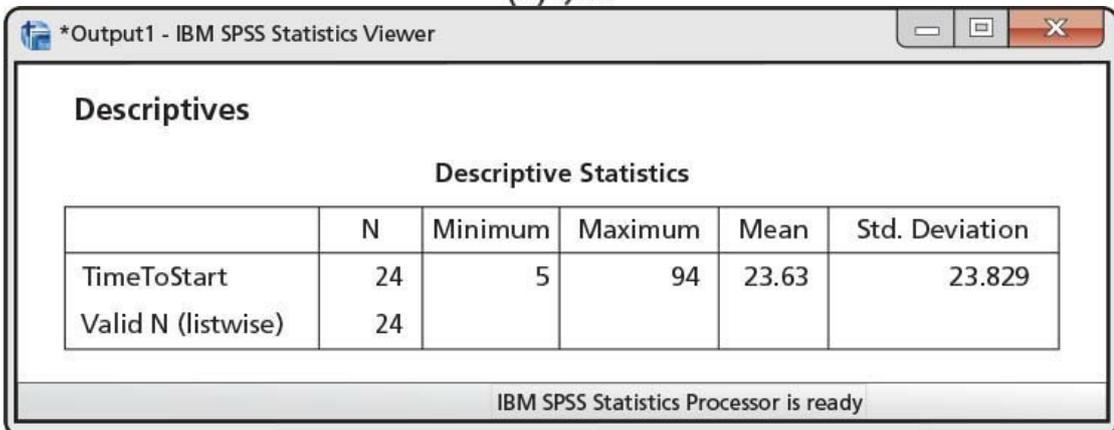
Statistical software often provides several numerical measures in response to a single command. Figure 1.15 displays such output from Minitab, JMP, and SPSS software for the data on the time to start a business. Examine the outputs carefully. Notice that they give different numbers of significant digits for some of these numerical summaries. Which output do you prefer?



(a) Minitab



(b) JMP



(c) SPSS

**FIGURE 1.15**

Descriptive statistics from (a) Minitab (b) JMP, and (c) SPSS for the time to start a business, for Example 1.27.

*There are several rules for calculating quartiles, which often give slightly different values. The differences are generally small. For describing data, just report the values that your software gives.*



## The five-number summary and boxplots

In Section 1.2, we used the smallest and largest observations to indicate the spread of a distribution. These single observations tell us little about the distribution as a whole, but they give information about the tails of the distribution that is missing if we know only  $Q_1$ ,  $M$ , and  $Q_3$ . To get a quick summary of both center and spread, use all five numbers.

### THE FIVE-NUMBER SUMMARY

The **five-number summary** of a set of observations consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is

Minimum  $Q_1$   $M$   $Q_3$  Maximum

### EXAMPLE

#### 1.28 Service center call lengths.



CALLS80

Table 1.2 (page 19) gives the service center call lengths for the sample of 80 calls that we discussed in Example 1.15. The five-number summary for these data is 1.0, 54.5, 103.5, 200, and 2631. The distribution is highly skewed. The mean is 197 seconds, a value that is very close to the third quartile.

## USE YOUR KNOWLEDGE

### 1.53 Verify the calculations.

Refer to the five-number summary and the mean for service center call lengths given in Example 1.28. Verify these results. Do not use software for this exercise and be sure to show all your work.



CALLS80

### 1.54 Find the five-number summary.

Here are the scores on the first exam in an introductory statistics course for 10 students:

81 73 93 85 75 98 93 55 80 90



STAT

Find the five-number summary for these first-exam scores.

The five-number summary leads to another visual representation of a distribution, the *boxplot*.

## BOXPLOT

A **boxplot** is a graph of the five-number summary.

- A central box spans the quartiles  $Q_1$  and  $Q_3$
- A line in the box marks the median  $M$
- Lines extend from the box out to the smallest and largest observations.

The lines extending to the smallest and largest observations are sometimes called *whiskers*, and boxplots are sometimes called *box-and-whisker plots*. Software provides many varieties of boxplots, some of which use different choices

for the placement of the whiskers.

whiskers

box-and-whisker plots

When you look at a boxplot, first locate the median, which marks the center of the distribution. Then look at the spread. The quartiles show the spread of the middle half of the data, and the extremes (the smallest and largest observations) show the spread of the entire data set.

## EXAMPLE

### 1.29 IQ scores.



In Example 1.14 (page 16), we used a histogram to examine the distribution of a sample of 60 IQ scores. A boxplot for these data is given in Figure 1.16. Note that the mean is marked with a “+” and appears very close to the median. The two quartiles are each approximately the same distance from the median, and the two whiskers are approximately the same distance from the corresponding quartiles. All these characteristics are consistent with a symmetric distribution, as illustrated by the histogram in Figure 1.9.

## USE YOUR KNOWLEDGE

### 1.55 Make a boxplot.

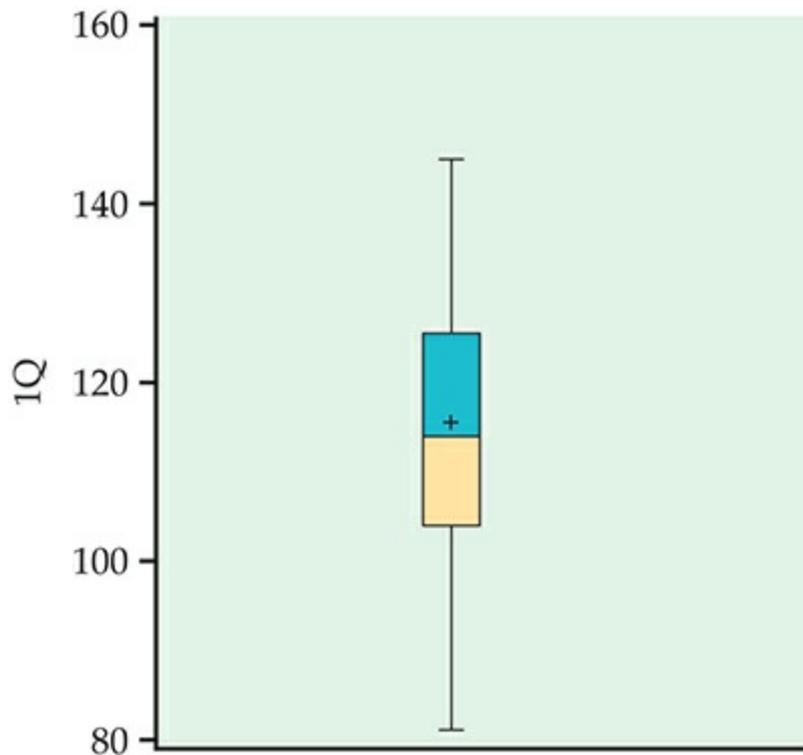
Here are the scores on the first exam in an introductory statistics course for 10 students:

81 73 93 85 75 98 93 55 80 90



Make a boxplot for these first-exam scores.

### The $1.5 \times IQR$ rule for suspected outliers



**FIGURE 1.16**  
Boxplot for sample of 60 IQ scores, for Example 1.29.

If we look at the data in Table 1.2 (page 19), we can spot a clear outlier, a call lasting 2631 seconds, more than twice the length of any other call. How can we describe the spread of this distribution? The smallest and largest observations are extremes that do not describe the spread of the majority of the data. The distance between the quartiles (the range of the center half of the data) is a more resistant measure of spread than the range. This distance is called the *interquartile range*.

### THE INTERQUARTILE RANGE *IQR*

The **interquartile range *IQR*** is the distance between the first and third quartiles:

$$IQR = Q_3 - Q_1$$

## EXAMPLE

### 1.30 *IQR* for service center call length data.

In Exercise 1.53 (page 38) you verified that the five-number summary for our data on service center call lengths was 1.0, 54.5, 103.5, 200, and 2631. Therefore, we calculate

$$\begin{aligned}IQR &= Q_3 - Q_1 \\IQR &= 200 - 54.5 \\&= 145.5\end{aligned}$$

The quartiles and the *IQR* are not affected by changes in either tail of the distribution. They are therefore resistant, because changes in a few data points have no further effect once these points move outside the quartiles.

However, *no single numerical measure of spread, such as IQR, is very useful for describing skewed distributions.* The two sides of a skewed distribution have different spreads, so one number can't summarize them. We can often detect skewness from the five-number summary by comparing how far the first quartile and the minimum are from the median (left tail) with how far the third quartile and the maximum are from the median (right tail). The interquartile range is mainly used as the basis for a rule of thumb for identifying suspected outliers.



## THE $1.5 \times IQR$ RULE FOR OUTLIERS

Call an observation a suspected outlier if it falls more than  $1.5 \times IQR$  above the third quartile or below the first quartile.

## EXAMPLE

### 1.31 Outliers for call length data.

For the call length data in Table 1.2 (page 19),

$$1.5 \times IQR = 1.5 \times 145.5 = 218.25$$



CALLS80

Any values below  $54.5 - 218.25 = -163.75$  or above  $200 + 218.25 = 418.25$  are flagged as possible outliers. There are no low outliers, but the 8 longest calls are flagged as possible high outliers. Their lengths are

438 465 479 700 700 951 1148 2631

## USE YOUR KNOWLEDGE

### 1.56 Find the *IQR*.



STAT

Here are the scores on the first exam in an introductory statistics course for 10 students:

81 73 93 85 75 98 93 55 80 90

Find the interquartile range and use the  $1.5 \times IQR$  rule to check for outliers. How low would the lowest score need to be for it to be an outlier according to this rule?

Two variations on the basic boxplot can be very useful. The first, called a *modified boxplot*, uses the  $1.5 \times IQR$  rule. The lines that extend out from the quartiles are terminated in whiskers that are  $1.5 \times IQR$  in length. Points beyond the

whiskers are plotted individually and are classified as outliers according to the  $1.5 \times IQR$  rule.

### modified boxplot

The other variation is to use two or more boxplots in the same graph to compare groups measured on the same variable. These are called *side-by-side boxplots*. The following example illustrates these two variations.

### side-by-side boxplots

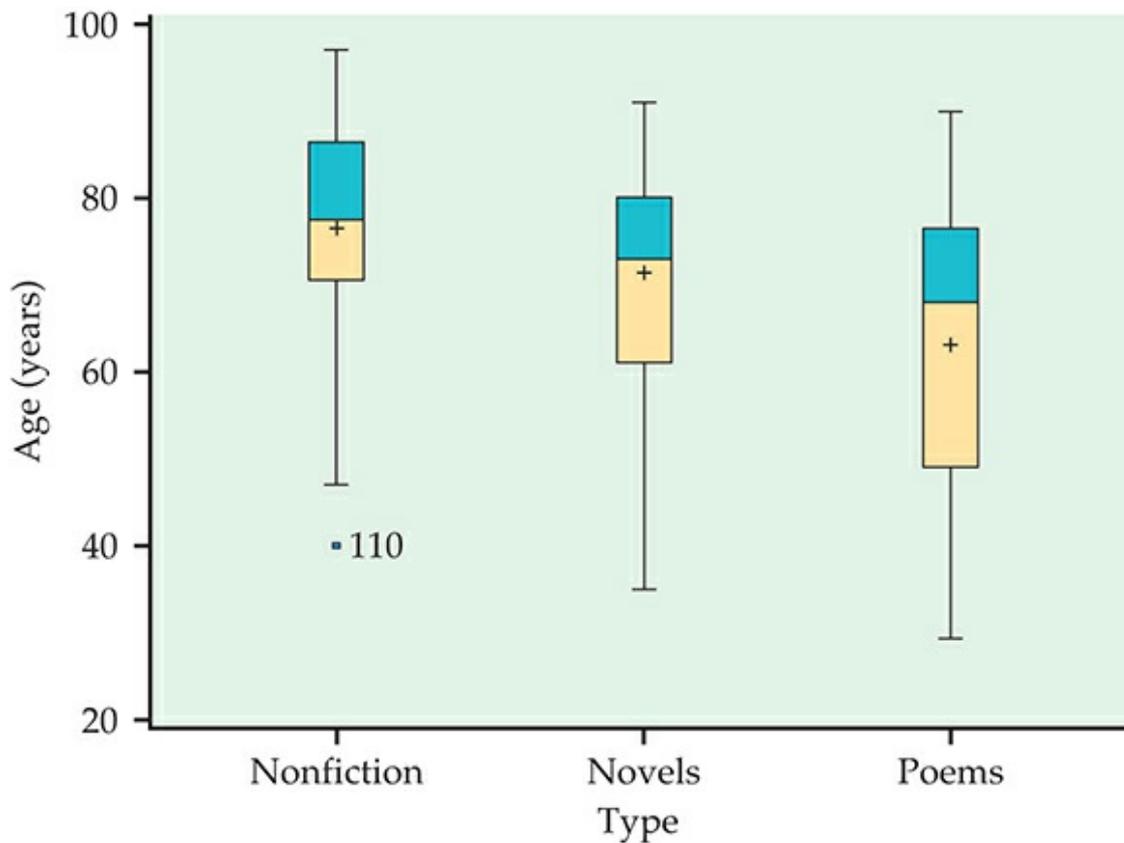
## EXAMPLE

### 1.32 Do poets die young?

According to William Butler Yeats, “She is the Gaelic muse, for she gives inspiration to those she persecutes. The Gaelic poets die young, for she is restless, and will not let them remain long on earth.” One study designed to investigate this issue examined the age at death for writers from different cultures and genders.<sup>22</sup>



Three categories of writers examined were novelists, poets, and nonfiction writers. We examine the ages at death for female writers in these categories from North America. Figure 1.17 shows modified side-by-side boxplots for the three categories of writers.



**FIGURE 1.17**

Modified side-by-side boxplots for the data on writers' age at death, for Example 1.32.

Displaying the boxplots for the three categories of writers lets us compare the three distributions. We see that nonfiction writers tend to live the longest, followed by novelists. The poets do appear to die young! There is one outlier among the nonfiction writers, which is plotted individually along with the value of its label. This writer died at the age of 40, young for a nonfiction writer, but not for a novelist or a poet!

## Measuring spread: the standard deviation

The five-number summary is not the most common numerical description of a distribution. That distinction belongs to the combination of the mean to measure center and the *standard deviation* to measure spread, or variability. The standard deviation measures spread by looking at how far the observations are from their mean.

### THE STANDARD DEVIATION $s$

The **variance**  $s^2$  of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of  $n$

observations  $x_1, x_2, \dots, x_n$  is

$$s^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 \cdot n - 1$$

or, in more compact notation,

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

The **standard deviation**  $s$  is the square root of the variance  $s^2$ :

$$s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

The idea behind the variance and the standard deviation as measures of spread is as follows: The deviations  $x_i - \bar{x}$  display the spread of the values  $x_i$  about their mean  $\bar{x}$ . Some of these deviations will be positive and some negative because some of the observations fall on each side of the mean. In fact, *the sum of the deviations of the observations from their mean will always be zero*. Squaring the deviations makes the negative deviations positive, so that observations far from the mean in either direction have large positive squared deviations. The variance is the average squared deviation. Therefore,  $s^2$  and  $s$  will be large if the observations are widely spread about their mean, and small if the observations are all close to the mean.

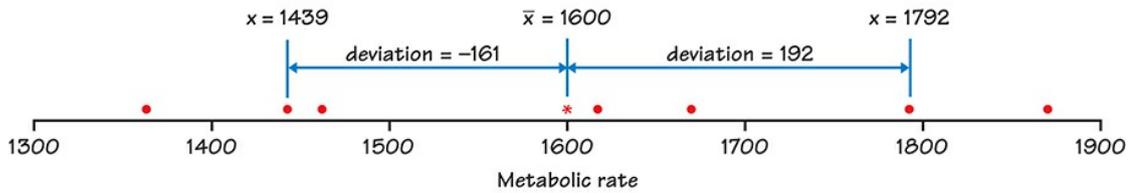
## EXAMPLE

### 1.33 Metabolic rate.



#### METABOL

A person's metabolic rate is the rate at which the body consumes energy. Metabolic rate is important in studies of weight gain, dieting, and exercise. Here are the metabolic rates of 7 men who took part in a study of dieting. (The units are calories per 24 hours. These are the same calories used to describe the energy content of foods.)



**FIGURE 1.18**

Metabolic rates for seven men, with the mean (\*) and the deviations of two observations from the mean, for Example 1.33.

1792 1666 1362 1614 1460 1867 1439

Enter these data into your calculator or software and verify that

$$\bar{x} = 1600 \text{ calories} \quad s = 189.24 \text{ calories}$$

Figure 1.18 plots these data as dots on the calorie scale, with their mean marked by an asterisk (\*). The arrows mark two of the deviations from the mean. If you were calculating  $s$  by hand, you would find the first deviation as

$$x_i - \bar{x} = 1792 - 1600 = 192$$

Exercise 1.82 asks you to calculate the seven deviations from Example 1.33, square them, and find  $s^2$  and  $s$  directly from the deviations. Working one or two short examples by hand helps you understand how the standard deviation is obtained. In practice, you will use either software or a calculator that will find  $s$ . The software outputs in Figure 1.15 (page 37) give the standard deviation for the data on the time to start a business.

## USE YOUR KNOWLEDGE

### 1.57 Find the variance and the standard deviation.



Here are the scores on the first exam in an introductory statistics course for 10 students:

81 73 93 85 75 98 93 55 80 90

Find the variance and the standard deviation for these first-exam scores.

The idea of the variance is straightforward: it is the average of the squares of the deviations of the observations from their mean. The details we have just presented, however, raise some questions.

*Why do we square the deviations?*

- First, the sum of the squared deviations of any set of observations from their mean is the smallest that the sum of squared deviations from any number can possibly be. This is not true of the unsquared distances. So squared deviations point to the mean as center in a way that distances do not.
- Second, the standard deviation turns out to be the natural measure of spread for a particularly important class of symmetric unimodal distributions, the *Normal distributions*. We will meet the Normal distributions in the next section.

*Why do we emphasize the standard deviation rather than the variance?*

- One reason is that  $s$ , not  $s^2$ , is the natural measure of spread for Normal distributions, which are introduced in the next section.
- There is also a more general reason to prefer  $s$  to  $s^2$ . Because the variance involves squaring the deviations, it does not have the same unit of measurement as the original observations. The variance of the metabolic rates, for example, is measured in squared calories. Taking the square root gives us a description of the spread of the distribution in the original measurement units.

*Why do we average by dividing by  $n - 1$  rather than  $n$  in calculating the variance?*

- Because the sum of the deviations is always zero, the last deviation can be found once we know the other  $n - 1$ . So we are not averaging  $n$  unrelated numbers. Only  $n - 1$  of the squared deviations can vary freely, and we average by dividing the total by  $n - 1$ .
- The number  $n - 1$  is called the *degrees of freedom* of the variance or standard deviation. Many calculators offer a choice between dividing by  $n$  and dividing by  $n - 1$ , so be sure to use  $n - 1$ .

degrees of freedom

## Properties of the standard deviation

Here are the basic properties of the standard deviation  $s$  as a measure of spread.

### PROPERTIES OF THE STANDARD DEVIATION

- $s$  measures spread about the mean and should be used only when the mean is chosen as the measure of center.
- $s = 0$  only when there is *no spread*. This happens only when all observations

have the same value. Otherwise,  $s > 0$ . As the observations become more spread out about their mean,  $s$  gets larger.

- $s$ , like the mean  $\bar{x}$ , is not resistant. A few outliers can make  $s$  very large.

## USE YOUR KNOWLEDGE

### 1.58 A standard deviation of zero.

Construct a data set with 5 cases that has a variable with  $s = 0$ .

*The use of squared deviations renders  $s$  even more sensitive than  $\bar{x}$  to a few extreme observations.* For example, when we add Suriname to our sample of 24 countries for the analysis of the time to start a business (Example 1.24 and Exercise 1.47), we increase the standard deviation from 23.8 to 137.9! Distributions with outliers and strongly skewed distributions have standard deviations that do not give much helpful information about such distributions.



## USE YOUR KNOWLEDGE

### 1.59 Effect of an outlier on the *IQR*.

Find the *IQR* for the time to start a business with and without Suriname. What do you conclude about the sensitivity of this measure of spread to the inclusion of an outlier?



TIME24, TIME25

## Choosing measures of center and spread

How do we choose between the five-number summary and  $\bar{x}$  and  $s$  to describe the center and spread of a distribution? Because the two sides of a strongly skewed distribution have different spreads, no single number such as  $s$  describes the spread well. The five-number summary, with its two quartiles and two extremes, does a better job.

## CHOOSING A SUMMARY

The five-number summary is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with strong outliers. Use  $\bar{x}$  and  $s$  for reasonably symmetric distributions that are free of outliers.

*Remember that a graph gives the best overall picture of a distribution. Numerical measures of center and spread report specific facts about a distribution, but they do not describe its shape. Numerical summaries do not disclose the presence of multiple modes or gaps, for example. **Always plot your data.***



## Changing the unit of measurement

The same variable can be recorded in different units of measurement. Americans commonly record distances in miles and temperatures in degrees Fahrenheit, while the rest of the world measures distances in kilometers and temperatures in degrees Celsius. Fortunately, it is easy to convert numerical descriptions of a distribution from one unit of measurement to another. This is true because a change in the measurement unit is a *linear transformation* of the measurements.

## LINEAR TRANSFORMATIONS

A **linear transformation** changes the original variable  $x$  into the new variable  $x_{\text{new}}$  given by an equation of the form

$$x_{\text{new}} = a + bx$$

Adding the constant  $a$  shifts all values of  $x$  upward or downward by the same amount. In particular, such a shift changes the origin (zero point) of the variable. Multiplying by the positive constant  $b$  changes the size of the unit of

**EXAMPLE****1.34 Change the units.**

(a) If a distance  $x$  is measured in kilometers, the same distance in miles is

$$x_{\text{new}} = 0.62x$$

For example, a 10-kilometer race covers 6.2 miles. This transformation changes the units without changing the origin—a distance of 0 kilometers is the same as a distance of 0 miles.

(b) A temperature  $x$  measured in degrees Fahrenheit must be reexpressed in degrees Celsius to be easily understood by the rest of the world. The transformation is

$$x_{\text{new}} = 59(x - 32) = -1609 + 59x$$

Thus, the high of 95°F on a hot American summer day translates into 35°C. In this case

$$a = -1609 \text{ and } b = 59$$

This linear transformation changes both the unit size and the origin of the measurements. The origin in the Celsius scale (0°C, the temperature at which water freezes) is 32° in the Fahrenheit scale.

*Linear transformations do not change the shape of a distribution.* If measurements on a variable  $x$  have a right-skewed distribution, any new variable  $x_{\text{new}}$  obtained by a linear transformation  $x_{\text{new}} = a + bx$  (for  $b > 0$ ) will also have a right-skewed distribution. If the distribution of  $x$  is symmetric and unimodal, the distribution of  $x_{\text{new}}$  remains symmetric and unimodal.

Although a linear transformation preserves the basic shape of a distribution, the center and spread will change. Because linear changes of measurement scale are common, we must be aware of their effect on numerical descriptive measures of center and spread. Fortunately, the changes follow a simple pattern.

## EXAMPLE

### 1.35 Use scores to find the points.

In an introductory statistics course, homework counts for 300 points out of a total of 1000 possible points for all course requirements. During the semester there were 12 homework assignments, and each was given a grade on a scale of 0 to 100. The maximum total score for the 12 homework assignments is therefore 1200. To convert the homework scores to final grade points, we need to convert the scale of 0 to 1200 to a scale of 0 to 300. We do this by multiplying the homework scores by  $300/1200$ . In other words, we divide the homework scores by 4. Here are the homework scores and the corresponding final grade points for 5 students:

Student	1	2	3	4	5
Score	1056	1080	900	1164	1020
Points	264	270	225	291	255

These two sets of numbers measure the same performance on homework for the course. Since we obtained the points by dividing the scores by 4, the mean of the points will be the mean of the scores divided by 4. Similarly, the standard deviation of points will be the standard deviation of the scores divided by 4.

## USE YOUR KNOWLEDGE

### 1.60 Calculate the points for a student.

Use the setting of Example 1.35 to find the points for a student whose score is 850.

Here is a summary of the rules for linear transformations.

## EFFECT OF A LINEAR TRANSFORMATION

To see the effect of a linear transformation on measures of center and spread,

apply these rules:

- Multiplying each observation by a positive number  $b$  multiplies both measures of center (mean and median) and measures of spread (interquartile range and standard deviation) by  $b$ .
- Adding the same number  $a$  (either positive or negative) to each observation adds  $a$  to measures of center and to quartiles and other percentiles but does not change measures of spread.

In Example 1.35, when we converted from score to points, we described the transformation as dividing by 4. The multiplication part of the summary of the effect of a linear transformation applies to this case because division by 4 is the same as multiplication by 0.25. Similarly, the second part of the summary applies to subtraction as well as addition because subtraction is simply the addition of a negative number.

The measures of spread  $IQR$  and  $s$  do not change when we add the same number  $a$  to all the observations because adding a constant changes the location of the distribution but leaves the spread unaltered. You can find the effect of a linear transformation  $x_{\text{new}} = a + bx$  by combining these rules. For example, if  $x$  has mean  $\bar{x}$ , the transformed variable  $x_{\text{new}}$  has mean  $a + b\bar{x}$ .

### SECTION 1.3 Summary

A numerical summary of a distribution should report its **center** and its **spread**, or **variability**.

The **mean**  $\bar{x}$  and the **median**  $M$  describe the center of a distribution in different ways. The mean is the arithmetic average of the observations, and the median is their midpoint.

When you use the median to describe the center of a distribution, describe its spread by giving the **quartiles**. The **first quartile**  $Q_1$  has one-fourth of the observations below it, and the **third quartile**  $Q_3$  has three-fourths of the observations below it.

The **interquartile range** is the difference between the quartiles. It is the spread of the center half of the data. The **1.5 × IQR rule** flags observations more than  $1.5 \times IQR$  beyond the quartiles as possible outliers.

The **five-number summary** consisting of the median, the quartiles, and the smallest and largest individual observations provides a quick overall description of a distribution. The median describes the center, and the quartiles and extremes show the spread.

**Boxplots** based on the five-number summary are useful for comparing several distributions. The box spans the quartiles and shows the spread of the central half of the distribution. The median is marked within the box. Lines extend from the box to the extremes and show the full spread of the data. In a **modified boxplot**,

points identified by the  $1.5 \times IQR$  rule are plotted individually. **Side-by-side boxplots** can be used to display boxplots for more than one group on the same graph.

The **variance**  $s^2$  and especially its square root, the **standard deviation**  $s$ , are common measures of spread about the mean as center. The standard deviation  $s$  is zero when there is no spread and gets larger as the spread increases.

A **resistant measure** of any aspect of a distribution is relatively unaffected by changes in the numerical value of a small proportion of the total number of observations, no matter how large these changes are. The median and quartiles are resistant, but the mean and the standard deviation are not.

The mean and standard deviation are good descriptions for symmetric distributions without outliers. They are most useful for the Normal distributions introduced in the next section. The five-number summary is a better exploratory description for skewed distributions.

**Linear transformations** have the form  $x_{\text{new}} = a + bx$ . A linear transformation changes the origin if  $a \neq 0$  and changes the size of the unit of measurement if  $b > 0$ . Linear transformations do not change the overall shape of a distribution. A linear transformation multiplies a measure of spread by  $b$  and changes a percentile or measure of center  $m$  into  $a + bm$ .

Numerical measures of particular aspects of a distribution, such as center and spread, do not report the entire shape of most distributions. In some cases, particularly distributions with multiple peaks and gaps, these measures may not be very informative.

## SECTION 1.3 Exercises

*For Exercises 1.47 and 1.48, see page 32; for Exercises 1.49 to 1.51, see page 34; for Exercise 1.52, see page 36; for Exercises 1.53 and 1.54, see page 38; for Exercise 1.55, see page 39; for Exercise 1.56, see page 41; for Exercise 1.57, see page 43; for Exercise 1.58, see page 44; for Exercise 1.59, see page 45; and for Exercise 1.60, see page 47.*

### 1.61 Gosset's data on double stout sales.

William Sealy Gosset worked at the Guinness Brewery in Dublin and made substantial contributions to the practice of statistics.<sup>23</sup> In his work at the brewery he collected and analyzed a great deal of data. Archives with Gosset's handwritten tables, graphs, and notes have been preserved at the Guinness Storehouse in Dublin.<sup>24</sup> In one study, Gosset examined the change in the double stout market before and after World War I (1914–1918). For various regions in England and Scotland, he calculated the ratio of sales in 1925, after the war, as a percent of sales in 1913, before the war. Here are the data: 

Bristol	94	Glasgow	66
Cardiff	112	Liverpool	140
English Agents	78	London	428
English O	68	Manchester	190
English P	46	Newcastle-on-Tyne	118
English R	111	Scottish	24

- (a) Compute the mean for these data.
- (b) Compute the median for these data.
- (c) Which measure do you prefer for describing the center of this distribution? Explain your answer. (You may include a graphical summary as part of your explanation.)

### 1.62 Measures of spread for the double stout data.

Refer to the previous exercise.  **STOUT**

- (a) Compute the standard deviation for these data.
- (b) Compute the quartiles for these data.
- (c) Which measure do you prefer for describing the spread of this distribution? Explain your answer. (You may include a graphical summary as part of your explanation.)

### 1.63 Are there outliers in the double stout data?

Refer to Exercise 1.61.  **STOUT**

- (a) Find the *IQR* for these data.
- (b) Use the  $1.5 \times IQR$  rule to identify and name any outliers.
- (c) Make a boxplot for these data and describe the distribution using only the information in the boxplot.
- (d) Make a modified boxplot for these data and describe the distribution using only the information in the boxplot.
- (e) Make a stemplot for these data.
- (f) Compare the boxplot, the modified boxplot, and the stemplot. Evaluate the advantages and disadvantages of each graphical summary for describing the distribution of the double stout data.

### 1.64 Smolts.

Smolts are young salmon at a stage when their skin becomes covered with silvery scales and they start to migrate from freshwater to the sea. The reflectance of a light shined on a smolt's skin is a measure of the smolt's readiness for the migration. Here are the reflectances, in percents, for a sample of 50 smolts:<sup>25</sup>  **SMOLTS**

57.6	54.8	63.4	57.0	54.7	42.3	63.6	55.5	33.5	63.3
58.3	42.1	56.1	47.8	56.1	55.9	38.8	49.7	42.3	45.6
69.0	50.4	53.0	38.3	60.4	49.3	42.8	44.5	46.4	44.3
58.9	42.1	47.6	47.9	69.2	46.6	68.1	42.8	45.6	47.3
59.6	37.8	53.9	43.2	51.4	64.5	43.8	42.7	50.9	43.8

- (a) Find the mean reflectance for these smolts.
- (b) Find the median reflectance for these smolts.

(c) Do you prefer the mean or the median as a measure of center for these data? Give reasons for your preference.

### 1.65 Measures of spread for smolts.

Refer to the previous exercise.  **SMOLTS**

- (a) Find the standard deviation of the reflectance for these smolts.
- (b) Find the quartiles of the reflectance for these smolts.
- (c) Do you prefer the standard deviation or the quartiles as a measure of spread for these data? Give reasons for your preference.

### 1.66 Are there outliers in the smolt data?

Refer to Exercise 1.64.  **SMOLTS**

- (a) Find the *IQR* for the smolt data.
- (b) Use the  $1.5 \times IQR$  rule to identify any outliers.
- (c) Make a boxplot for the smolt data and describe the distribution using only the information in the boxplot.
- (d) Make a modified boxplot for these data and describe the distribution using only the information in the boxplot.
- (e) Make a stemplot for these data.
- (f) Compare the boxplot, the modified boxplot, and the stemplot. Evaluate the advantages and disadvantages of each graphical summary for describing the distribution of the smolt reflectance data.

### 1.67 The value of brands.

A brand is a symbol or images that are associated with a company. An effective brand identifies the company and its products. Using a variety of measures, dollar values for brands can be calculated.<sup>26</sup> The most valuable brand is Apple, with a value of \$76.568 million. Apple is followed by Google at \$69.726 million, Coca-Cola at \$67.839 million, Microsoft at \$57.853 million, and IBM at \$57.532 million. For this exercise you will use the brand values (in millions of dollars) for the top 100 brands in the data file

**BRANDS.**  **BRANDS**

- (a) Graphically display the distribution of the values of these brands.
- (b) Use numerical measures to summarize the distribution.
- (c) Write a short paragraph discussing the dollar values of the top 100 brands. Include the results of your analysis.

### 1.68 Alcohol content of beer.

Brewing beer involves a variety of steps that can affect the alcohol content. The data file BEER gives the

percent alcohol for 153 domestic brands of beer.<sup>27</sup>  **BEER**

- (a) Use graphical and numerical summaries of your choice to describe these data. Give reasons for your choices.
- (b) Give the alcohol content and the brand of any outliers. Explain how you determined that they were outliers.

### 1.69 Remove the outliers for alcohol content of beer.

Refer to the previous exercise.  **BEER**

- (a) Calculate the mean with and without the outliers. Do the same for the median. Explain how these statistics change when the outliers are excluded.
- (b) Calculate the standard deviation with and without the outliers. Do the same for the quartiles. Explain how these statistics change when the outliers are excluded.
- (c) Write a short paragraph summarizing what you have learned in this exercise.

### 1.70 Calories in beer.

Refer to the previous two exercises. The data file also gives the calories per 12 ounces of beverage.  **BEER**

- (a) Analyze the data and summarize the distribution of calories for these 153 brands of beer.
- (b) In Exercise 1.68 you identified outliers. To what extent are these brands outliers in the distribution of calories? Explain your answer.

### 1.71 Potatoes.

A quality product is one that is consistent and has very little variability in its characteristics. Controlling variability can be more difficult with agricultural products than with those that are manufactured. The following table gives the weights, in ounces, of the 25 potatoes sold in a 10-pound bag.  **POTATO**

7.6	7.9	8.0	6.9	6.7	7.9	7.9	7.9	7.6	7.8	7.0	4.7	7.6
6.3	4.7	4.7	4.7	6.3	6.0	5.3	4.3	7.9	5.2	6.0	3.7	

- (a) Summarize the data graphically and numerically. Give reasons for the methods you chose to use in your summaries.
- (b) Do you think that your numerical summaries do an effective job of describing these data? Why or why not?
- (c) There appear to be two distinct clusters of weights for these potatoes. Divide the sample into two subsamples based on the clustering. Give the mean and standard deviation for each subsample. Do you think that this way of summarizing these data is better than a numerical summary that uses all the data as a single sample? Give a reason for your answer.

### 1.72 Longleaf pine trees.

The Wade Tract in Thomas County, Georgia, is an old-growth forest of longleaf pine trees (*Pinus palustris*) that has survived in a relatively undisturbed state since before the settlement of the area by Europeans. A study collected data on 584 of these trees.<sup>28</sup> One of the variables measured was the diameter at breast height (DBH). This is the diameter of the tree at 4.5 feet and the units are centimeters (cm). Only trees with DBH greater than 1.5 cm were sampled. Here are the diameters of a random sample of 40 of these trees:  PINES

10.5	13.3	26.0	18.3	52.2	9.2	26.1	17.6	40.5	31.8
47.2	11.4	2.7	69.3	44.4	16.9	35.7	5.4	44.2	2.2
4.3	7.8	38.1	2.2	11.4	51.5	4.9	39.7	32.6	51.8
43.6	2.3	44.6	31.5	40.3	22.3	43.3	37.5	29.1	27.9

- Find the five-number summary for these data.
- Make a boxplot.
- Make a histogram.
- Write a short summary of the major features of this distribution. Do you prefer the boxplot or the histogram for these data?

### 1.73 Blood proteins in children from Papua New Guinea.

C-reactive protein (CRP) is a substance that can be measured in the blood. Values increase substantially within 6 hours of an infection and reach a peak within 24 to 48 hours. In adults, chronically high values have been linked to an increased risk of cardiovascular disease. In a study of apparently healthy children aged 6 to 60 months in Papua New Guinea, CRP was measured in 90 children.<sup>29</sup> The units are milligrams per liter (mg/l). Here are the data from a random sample of 40 of these children:  CRP

0.00	3.90	5.64	8.22	0.00	5.62	3.92	6.81	30.61	0.00
73.20	0.00	46.70	0.00	0.00	26.41	22.82	0.00	0.00	3.49
0.00	0.00	4.81	9.57	5.36	0.00	5.66	0.00	59.76	12.38
15.74	0.00	0.00	0.00	0.00	9.37	20.78	7.10	7.89	5.53

- Find the five-number summary for these data.
- Make a boxplot.
- Make a histogram.
- Write a short summary of the major features of this distribution. Do you prefer the boxplot or the histogram for these data?

### 1.74 Does a log transform reduce the skewness?

Refer to the previous exercise. With strongly skewed distributions such as this, we frequently reduce the skewness by taking a log transformation. We have a bit of a problem here, however, because some of the data are recorded as 0.00, and the logarithm of zero is not defined. For this variable, the value 0.00 is recorded whenever the amount of CRP in the blood is below the level that the measuring instrument is capable of detecting. The usual procedure in this circumstance is to add a small number to each observation before taking the logs. Transform these data by adding 1 to each observation and then taking the logarithm. Use the questions in the previous exercise as a guide to your analysis, and prepare a summary contrasting

this analysis with the one that you performed in the previous exercise.  CRP

### 1.75 Vitamin A deficiency in children from Papua New Guinea.

In the Papua New Guinea study that provided the data for the previous two exercises, the researchers also measured serum retinol. A low value of this variable can be an indicator of vitamin A deficiency. Here are the data on the same sample of 40 children from this study. The units are micromoles per liter  $\mu\text{mol/l}$ .

1.15	1.36	0.38	0.34	0.35	0.37	1.17	0.97	0.97	0.67
0.31	0.99	0.52	0.70	0.88	0.36	0.24	1.00	1.13	0.31
1.44	0.35	0.34	1.90	1.19	0.94	0.34	0.35	0.33	0.69
0.69	1.04	0.83	1.11	1.02	0.56	0.82	1.20	0.87	0.41

Analyze these data. Use the questions in the previous two exercises as a guide.  VITA

### 1.76 Luck and puzzle solving.

Children in a psychology study were asked to solve some puzzles and were then given feedback on their performance. They then were asked to rate how luck played a role in determining their scores.<sup>30</sup> This variable was recorded on a 1 to 10 scale with 1 corresponding to very lucky and 10 corresponding to very unlucky. Here are the scores for 60 children:

1	10	1	10	1	1	10	5	1	1	8	1	10	2	1
9	5	2	1	8	10	5	9	10	10	9	6	10	1	5
1	9	2	1	7	10	9	5	10	10	10	1	8	1	6
10	1	6	10	10	8	10	3	10	8	1	8	10	4	2

Use numerical and graphical methods to describe these data. Write a short report summarizing your work.

 LUCK

### 1.77 Median versus mean for net worth.

A report on the assets of American households says that the median net worth of U.S. families is \$77,300. The mean net worth of these families is \$498,800.<sup>31</sup> What explains the difference between these two measures of center?

### 1.78 Create a data set.

Create a data set with 9 observations for which the median would change by a large amount if the smallest observation were deleted.

### 1.79 Mean versus median.

A small accounting firm pays each of its six clerks \$45,000, two junior accountants \$70,000 each, and the firm's owner \$420,000. What is the mean salary paid at this firm? How many of the employees earn less than the mean? What is the median salary?

## 1.80 Be careful about how you treat the zeros.

In computing the median income of any group, some federal agencies omit all members of the group who had no income. Give an example to show that the reported median income of a group can go down even though the group becomes economically better off. Is this also true of the mean income?

## 1.81 How does the median change?

The firm in Exercise 1.79 gives no raises to the clerks and junior accountants, while the owner's take increases to \$500,000. How does this change affect the mean? How does it affect the median?

## 1.82 Metabolic rates.

Calculate the mean and standard deviation of the metabolic rates in Example 1.33 (page 42), showing each step in detail. First find the mean  $\bar{x}$  by summing the 7 observations and dividing by 7. Then find each of the deviations  $x_i - \bar{x}$  and their squares. Check that the deviations have sum 0. Calculate the variance as an average of the squared deviations (remember to divide by  $n - 1$ ). Finally, obtain  $s$  as the square root of the variance.  **METABOL**

## 1.83 Earthquakes.

Each year there are about 900,000 earthquakes of magnitude 2.5 or less that are usually not felt. In contrast, there are about 10 of magnitude 7.0 that cause serious damage.<sup>32</sup> Explain why the average magnitude of earthquakes is not a good measure of their impact.

## 1.84 IQ scores.

Many standard statistical methods that you will study in Part II of this book are intended for use with distributions that are symmetric and have no outliers. These methods start with the mean and standard deviation,  $\bar{x}$  and  $s$ . For example, standard methods would typically be used for the IQ and GPA data in Table 1.3 (page 29).  **IQGPA**

- Find  $\bar{x}$  and  $s$  for the IQ data. In large populations, IQ scores are standardized to have mean 100 and standard deviation 15. In what way does the distribution of IQ among these students differ from the overall population?
- Find the median IQ score. It is, as we expect, close to the mean.
- Find the mean and median for the GPA data. The two measures of center differ a bit. What feature of the data (see your stemplot in Exercise 1.43 or make a new stemplot) explains the difference?

## 1.85 Mean and median for two observations.

The *Mean and Median* applet allows you to place observations on a line and see their mean and median visually. Place two observations on the line by clicking below it. Why does only one arrow appear?

## 1.86 Mean and median for three observations.

In the *Mean and Median* applet, place three observations on the line by clicking below it, two close together near the center of the line and one somewhat to the right of these two.

(a) Pull the single rightmost observation out to the right. (Place the cursor on the point, hold down a mouse button, and drag the point.) How does the mean behave? How does the median behave? Explain briefly why each measure acts as it does.

(b) Now drag the rightmost point to the left as far as you can. What happens to the mean? What happens to the median as you drag this point past the other two (watch carefully)?



### 1.87 Mean and median for five observations.

Place five observations on the line in the *Mean and Median* applet by clicking below it.

(a) Add one additional observation *without changing the median*. Where is your new point?

(b) Use the applet to convince yourself that when you add yet another observation (there are now seven in all), the median does not change no matter where you put the seventh point. Explain why this must be true.

### 1.88 Hummingbirds and flowers.

Different varieties of the tropical flower *Heliconia* are fertilized by different species of hummingbirds. Over time, the lengths of the flowers and the form of the hummingbirds' beaks have evolved to match each other. Here are data on the lengths in millimeters of three varieties of these flowers on the island of Dominica.<sup>33</sup>

<i>H. bihai</i>							
47.12	46.75	46.81	47.12	46.67	47.43	46.44	46.64
48.07	48.34	48.15	50.26	50.12	46.34	46.94	48.36
<i>H. caribaea red</i>							
41.90	42.01	41.93	43.09	41.47	41.69	39.78	40.57
39.63	42.18	40.66	37.87	39.16	37.40	38.20	38.07
38.10	37.97	38.79	38.23	38.87	37.78	38.01	
<i>H. caribaea yellow</i>							
36.78	37.02	36.52	36.11	36.03	35.45	38.13	37.1
35.17	36.82	36.66	35.68	36.03	34.57	34.63	

Make boxplots to compare the three distributions. Report the five-number summaries along with your graph. What are the most important differences among the three varieties of flowers?  HELICON

### 1.89 Compare the three varieties of flowers.

The biologists who collected the flower length data in the previous exercise compared the three *Heliconia* varieties using statistical methods based on  $\bar{x}$  and  $s$ .  HELICON

(a) Find  $\bar{x}$  and  $s$  for each variety.

(b) Make a stemplot of each set of flower lengths. Do the distributions appear suitable for use of  $\bar{x}$  and  $s$  as summaries?

## 1.90 Imputation.

Various problems with data collection can cause some observations to be missing. Suppose a data set has 20 cases. Here are the values of the variable  $x$  for 10 of these cases:  **IMPUTE**

17 6 12 14 20 23 9 12 16 21

The values for the other 10 cases are missing. One way to deal with missing data is called **imputation**. The basic idea is that missing values are replaced, or imputed, with values that are based on an analysis of the data that are not missing. For a data set with a single variable, the usual choice of a value for imputation is the mean of the values that are not missing. The mean for this data set is 15.

- Verify that the mean is 15 and find the standard deviation for the 10 cases for which  $x$  is not missing.
- Create a new data set with 20 cases by setting the values for the 10 missing cases to 15. Compute the mean and standard deviation for this data set.
- Summarize what you have learned about the possible effects of this type of imputation on the mean and the standard deviation.



### 1.91 Create a data set.

Give an example of a small set of data for which the mean is smaller than the third quartile.

### 1.92 Create another data set.

Create a set of 5 positive numbers (repeats allowed) that have median 11 and mean 8. What thought process did you use to create your numbers?



### 1.93 A standard deviation contest.

This is a standard deviation contest. You must choose four numbers from the whole numbers 0 to 20, with repeats allowed.

- Choose four numbers that have the smallest possible standard deviation.
- Choose four numbers that have the largest possible standard deviation.
- Is more than one choice possible in either (a) or (b)? Explain.



### 1.94 Deviations from the mean sum to zero.

Use the definition of the mean  $\bar{x}$  to show that the sum of the deviations  $x_i - \bar{x}$  of the observations from their mean is always zero. This is one reason why the variance and standard deviation use squared deviations.

### 1.95 Does your software give incorrect answers?

This exercise requires a calculator with a standard deviation button or statistical software on a computer. The observations

30,001 30,002 30,003

have mean  $\bar{x}=30,002$  and standard deviation  $s = 1$ . Adding a 0 in the center of each number, the next set becomes

300,001 300,003 300,003

The standard deviation remains  $s = 1$  as more 0s are added. Use your calculator or computer to calculate the standard deviation of these numbers, adding extra 0s until you get an incorrect answer. How soon did you go wrong? This demonstrates that calculators and computers cannot handle an arbitrary number of digits correctly.

## 1.96 Compare three varieties of flowers.

Exercise 1.88 reports data on the lengths in millimeters of flowers of three varieties of *Heliconia*. In Exercise 1.89 you found the mean and standard deviation for each variety. Starting from the  $\bar{x}$ - and  $s$ -values in millimeters, find the means and standard deviations in inches. (A millimeter is 1/1000 of a meter. A meter is 39.37 inches.)

## 1.97 Weight gain.

A study of diet and weight gain deliberately overfed 12 volunteers for eight weeks. The mean increase in fat was  $\bar{x}=2.32$  kilograms, and the standard deviation was  $s = 1.21$  kilograms. What are  $\bar{x}$  and  $s$  in pounds? (A kilogram is 2.2 pounds.)



## 1.98 Changing units from inches to centimeters.

Changing the unit of length from inches to centimeters multiplies each length by 2.54 because there are 2.54 centimeters in an inch. This change of units multiplies our usual measures of spread by 2.54. This is true of *IQR* and the standard deviation. What happens to the variance when we change units in this way?

## 1.99 A different type of mean.

The **trimmed mean** is a measure of center that is more resistant than the mean but uses more of the available information than the median. To compute the 10% trimmed mean, discard the highest 10% and the lowest 10% of the observations and compute the mean of the remaining 80%. Trimming eliminates the effect of a small number of outliers. Compute the 10% trimmed mean of the service time data in Table 1.2 (page 19). Then compute the 20% trimmed mean. Compare the values of these measures with the median and the ordinary untrimmed mean.



## 1.100 Changing units from centimeters to inches.

Refer to Exercise 1.72 (page 50). Change the measurements from centimeters to inches by multiplying each value by 0.39. Answer the questions from that exercise and explain the effect of the transformation on these data.

## 1.4 Density Curves and Normal Distributions

When you complete this section, you will be able to

- Compare the mean and the median for symmetric and skewed distributions.
- Sketch a Normal distribution for any given mean and standard deviation.
- Apply the 68–95–99.7 rule to find proportions of observations within 1, 2, and 3 standard deviations of the mean for any Normal distribution.
- Transform values of a variable from a general Normal distribution to the standard Normal distribution.
- Compute areas under a Normal curve using software or Table A.
- Perform inverse Normal calculations to find values of a Normal variable corresponding to various areas.
- Assess the extent to which the distribution of a set of data can be approximated by a Normal distribution.

We now have a kit of graphical and numerical tools for describing distributions. What is more, we have a clear strategy for exploring data on a single quantitative variable:

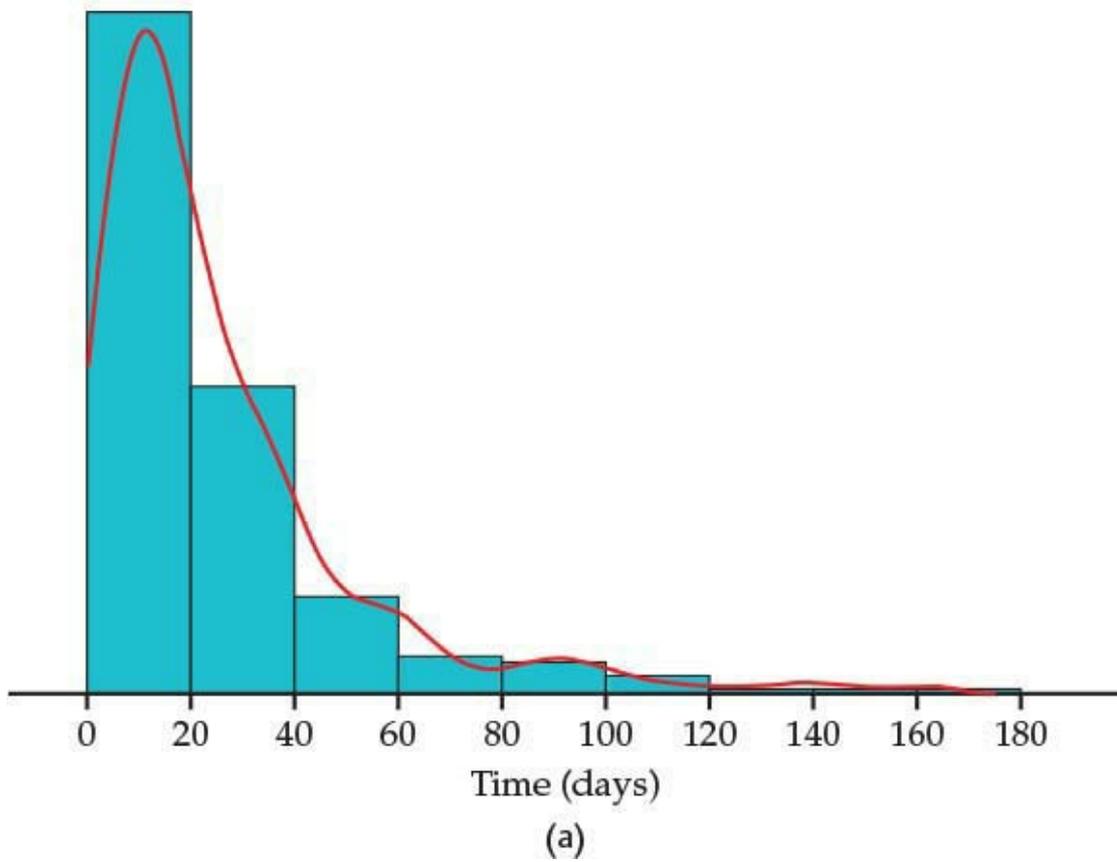
1. Always plot your data: make a graph, usually a stemplot or a histogram.
2. Look for the overall pattern and for striking deviations such as outliers.
3. Calculate an appropriate numerical summary to briefly describe center and spread.

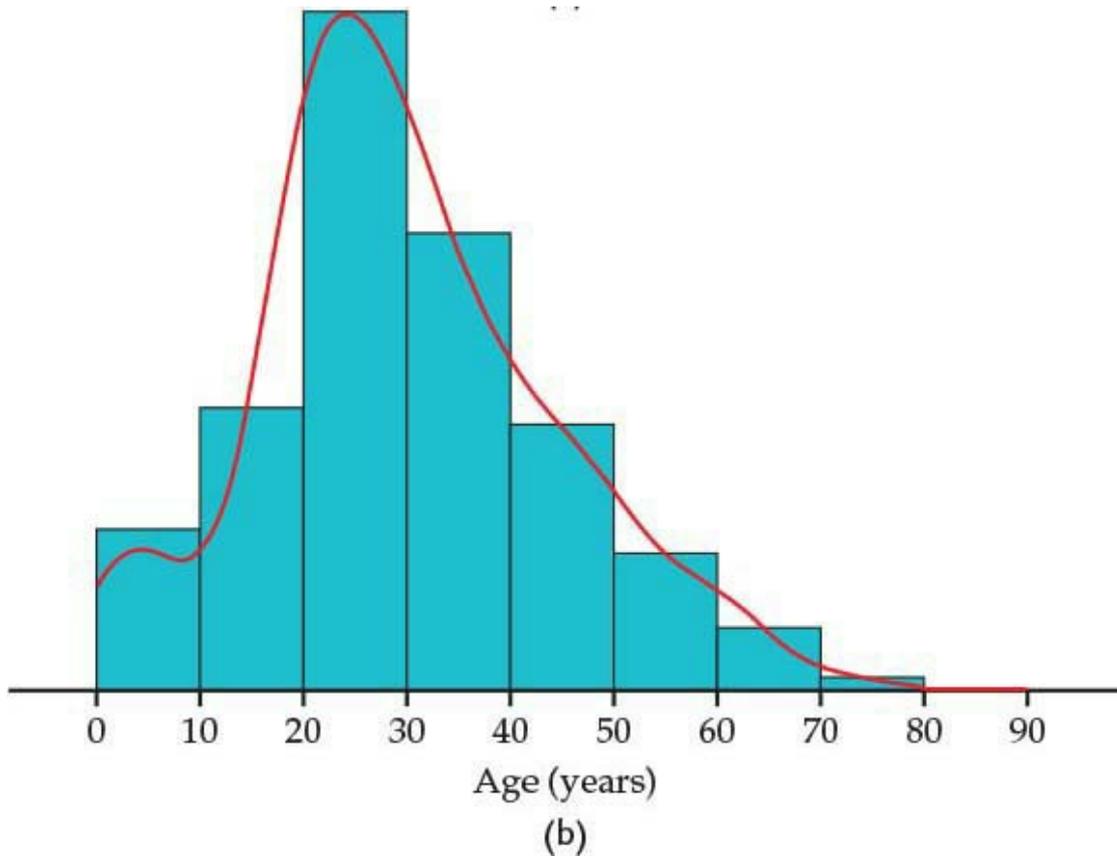
Technology has expanded the set of graphs that we can choose for Step 1. It is possible, though painful, to make histograms by hand. Using software, clever algorithms can describe a distribution in a way that is not feasible by hand, by fitting a smooth curve to the data in addition to or instead of a histogram. The curves used are called *density curves*. Before we examine density curves in detail, here is an example of what software can do.

density curve

## EXAMPLE

1.36 Density curves for times to start a business and *Titanic* passenger ages.





**FIGURE 1.19**

(a) The distribution of the time to start a business, for Example 1.36. The distribution is pictured with both a histogram and a density curve. (b) The distribution of the ages of the *Titanic* passengers, for Example 1.36. These distributions have a single mode with tails of two different lengths.

Figure 1.19 illustrates the use of density curves along with histograms to describe distributions. Figure 1.19(a) shows the distribution of the times to start a business for 194 countries (see Example 1.23, page 31). The outlier, Surinam, described in Exercise 1.47 (page 32), has been deleted from the data set. The distribution is highly skewed to the right. Most of the data are in the first two classes, with 40 or fewer days to start a business.

Exercise 1.25 (page 33) describes data on the class of the ticket of the *Titanic* passengers, and Figure 1.19(b) shows the distribution of the ages of these passengers. It has a single mode, a long right tail, and a relatively short left tail.



TITANIC

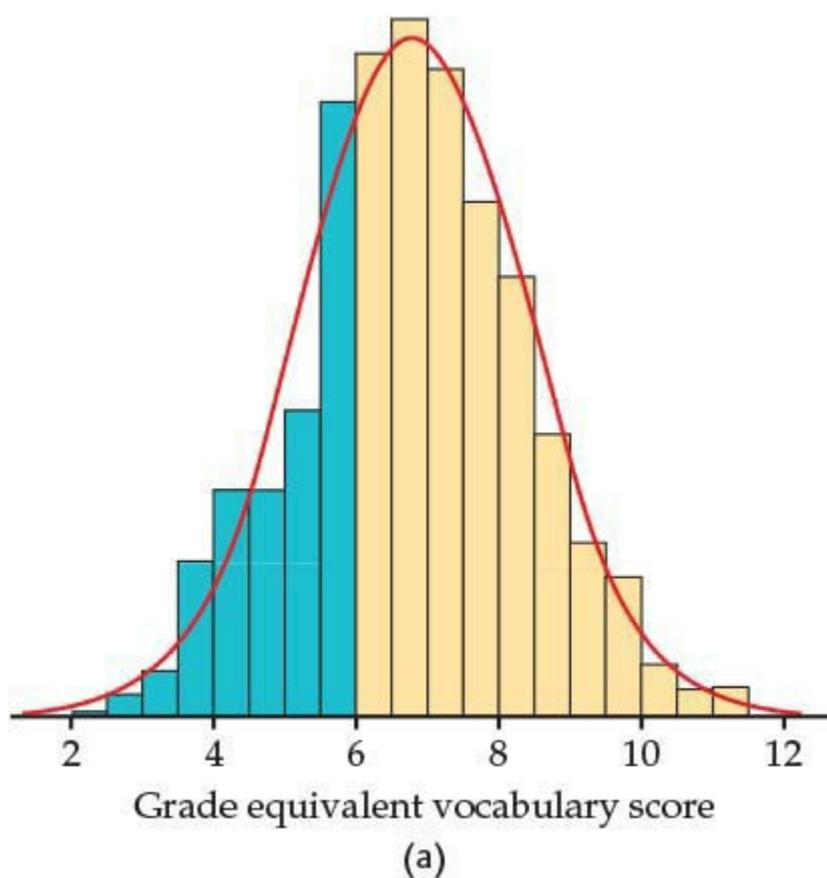
A smooth density curve is an idealization that gives the overall pattern of the data but ignores minor irregularities. We turn now to a special class of density curves, the bell-shaped Normal curves.

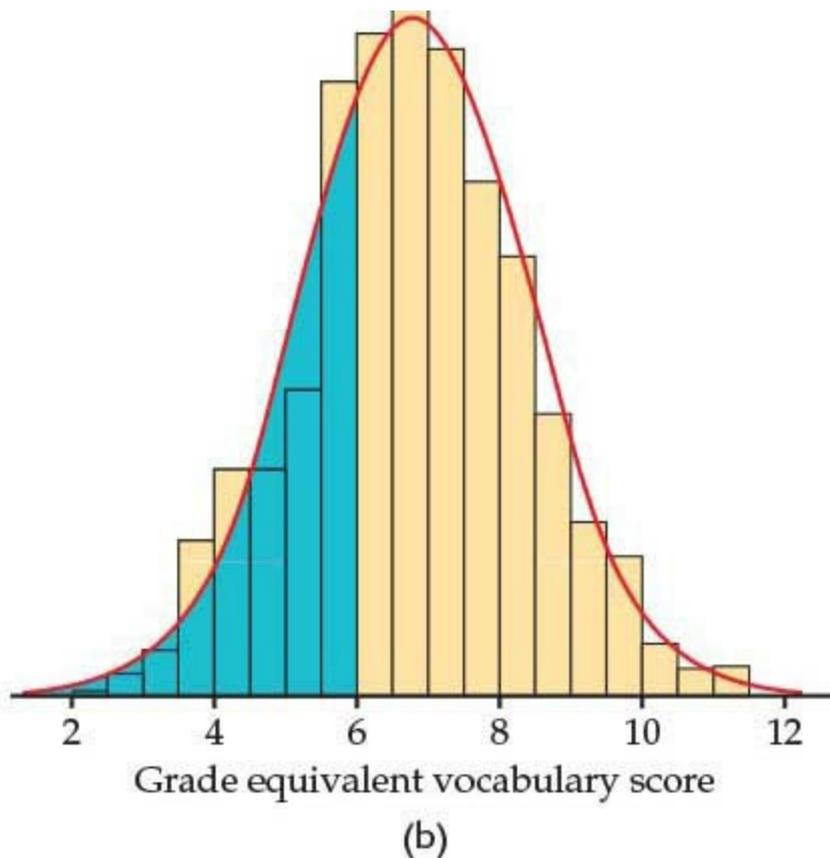
## Density curves

One way to think of a density curve is as a smooth approximation to the irregular bars of a histogram. Figure 1.20 shows a histogram of the scores of all 947 seventh-grade students in Gary, Indiana, on the vocabulary part of the Iowa Test of Basic Skills. Scores of many students on this national test have a very regular distribution. The histogram is symmetric, and both tails fall off quite smoothly from a single center peak. There are no large gaps or obvious outliers. The curve drawn through the tops of the histogram bars in Figure 1.20 is a good description of the overall pattern of the data.

### EXAMPLE

#### 1.37 Vocabulary scores.





**FIGURE 1.20**

- (a) The distribution of Iowa Test vocabulary scores for Gary, Indiana, seventh-graders, for Example 1.37. The shaded bars in the histogram represent scores less than or equal to 6.0.  
 (b) The shaded area under the Normal density curve also represents scores less than or equal to 6.0. This area is 0.293, close to the true 0.303 for the actual data.

In a histogram, the *areas* of the bars represent either counts or proportions of the observations. In Figure 1.20(a) we have shaded the bars that represent students with vocabulary scores 6.0 or lower. There are 287 such students, who make up the proportion  $287/947 = 0.303$  of all Gary seventh-graders. The shaded bars in Figure 1.20(a) make up *proportion* 0.303 of the total area under all the bars. If we adjust the scale so that the total area of the bars is 1, the *area of the shaded bars* will also be 0.303.

In Figure 1.20(b), we have shaded the *area under the curve* to the left of 6.0. If we adjust the scale so that the total area under the curve is exactly 1, areas under the curve will then represent proportions of the observations. That is, *area = proportion*. The curve is then a density curve. The shaded area under the density curve in Figure 1.20(b) represents the proportion of students with score 6.0 or lower. This area is 0.293, only 0.010 away from the histogram result. You can see that areas under the density curve give quite good approximations of areas given by the histogram.

## DENSITY CURVE

A **density curve** is a curve that

- is always on or above the horizontal axis and
- has area exactly 1 underneath it.

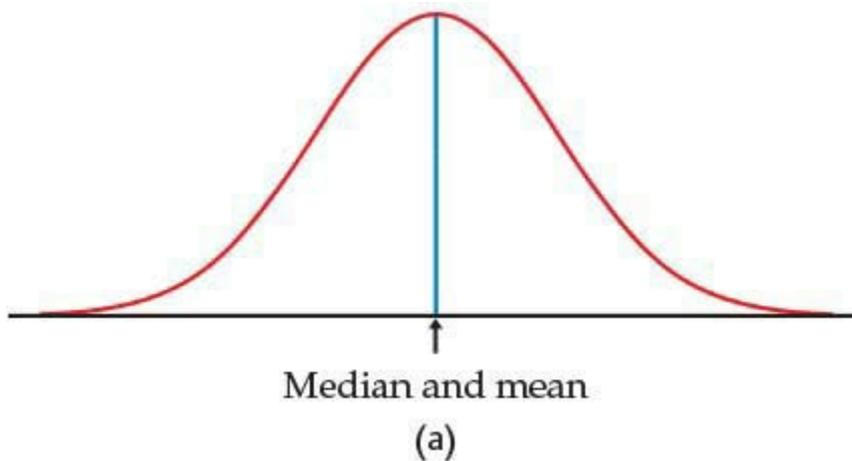
A density curve describes the overall pattern of a distribution. The area under the curve and above any range of values is the proportion of all observations that fall in that range.

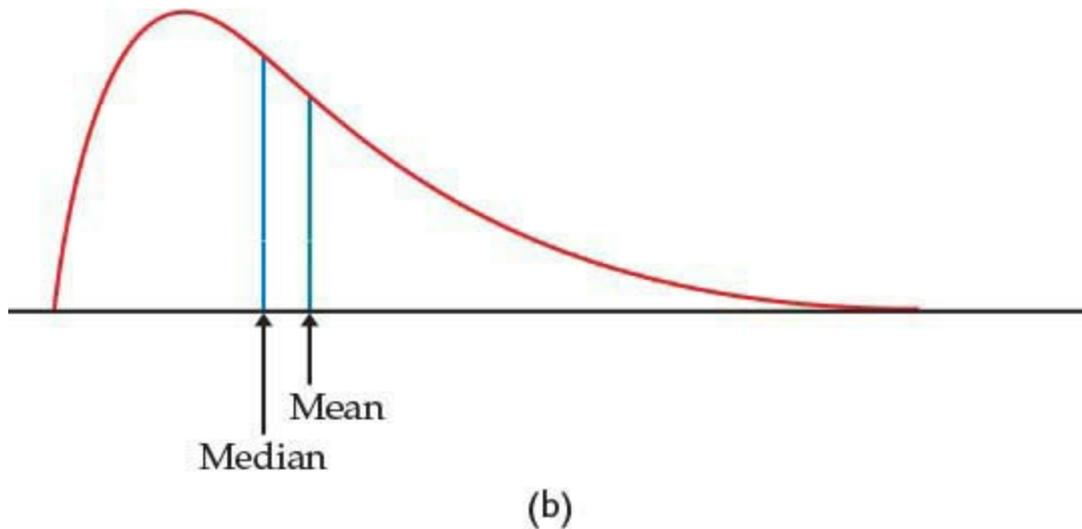
The density curve in Figure 1.20 is a *Normal curve*. Density curves, like distributions, come in many shapes. Figure 1.21 shows two density curves, a symmetric Normal density curve and a right-skewed curve.

We will discuss Normal density curves in detail in this section because of the important role that they play in statistics. There are, however, many applications where the use of other families of density curves are essential.

A density curve of an appropriate shape is often an adequate description of the overall pattern of a distribution. Outliers, which are deviations from the overall pattern, are not described by the curve.

### Measuring center and spread for density curves

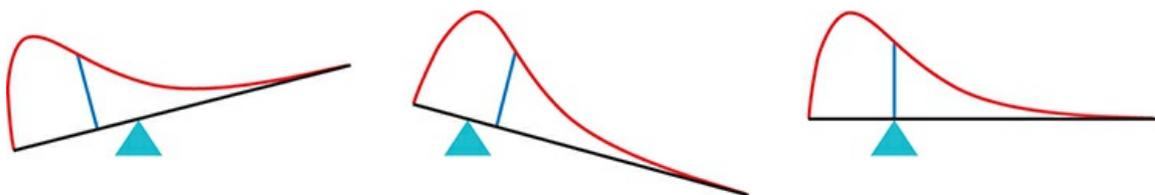




**FIGURE 1.21**

(a) A symmetric Normal density curve with its mean and median marked. (b) A right-skewed density curve with its mean and median marked.

Our measures of center and spread apply to density curves as well as to actual sets of observations, but only some of these measures are easily seen from the curve. A *mode* of a distribution described by a density curve is a peak point of the curve, the location where the curve is highest. Because areas under a density curve represent proportions of the observations, the *median* is the point with half the total area on each side. You can roughly locate the *quartiles* by dividing the area under the curve into quarters as accurately as possible by eye. The *IQR* is the distance between the first and third quartiles. There are mathematical ways of calculating areas under curves. These allow us to locate the median and quartiles exactly on any density curve.



**FIGURE 1.22**

The mean of a density curve is the point at which it would balance.

What about the mean and standard deviation? The mean of a set of observations is their arithmetic average. If we think of the observations as weights strung out along a thin rod, the mean is the point at which the rod would balance. This fact is also true of density curves. The mean is the point at which the curve would balance if it were made out of solid material. Figure 1.22 illustrates this interpretation of the mean.

A symmetric curve, such as the Normal curve in Figure 1.21(a), balances at its center of symmetry. Half the area under a symmetric curve lies on either side of its center, so this is also the median.

For a right-skewed curve, such as those shown in Figures 1.21(b) and 1.22, the small area in the long right tail tips the curve more than the same area near the center. The mean (the balance point) therefore lies to the right of the median. It is hard to locate the balance point by eye on a skewed curve. There are mathematical ways of calculating the mean for any density curve, so we are able to mark the mean as well as the median in Figure 1.21(b). The standard deviation can also be calculated mathematically, but it can't be located by eye on most density curves.

## MEDIAN AND MEAN OF A DENSITY CURVE

The **median** of a density curve is the equal-areas point, the point that divides the area under the curve in half.

The **mean** of a density curve is the balance point, at which the curve would balance if made of solid material.

The median and mean are the same for a symmetric density curve. They both lie at the center of the curve. The mean of a skewed curve is pulled away from the median in the direction of the long tail.

A density curve is an idealized description of a distribution of data. For example, the density curve in Figure 1.20 is exactly symmetric, but the histogram of vocabulary scores is only approximately symmetric. We therefore need to distinguish between the mean and standard deviation of the density curve and the numbers  $\bar{x}$  and  $s$  computed from the actual observations. The usual notation for the mean of an idealized distribution is  $\mu$  (the Greek letter mu). We write the standard deviation of a density curve as  $\sigma$  (the Greek letter sigma).

mean  $\mu$

standard deviation  $\sigma$

## Normal distributions

One particularly important class of density curves has already appeared in Figures 1.20 and 1.21(a). These density curves are symmetric, unimodal, and bell-shaped. They are called **Normal curves**, and they describe **Normal distributions**. All Normal distributions have the same overall shape.

Normal curves

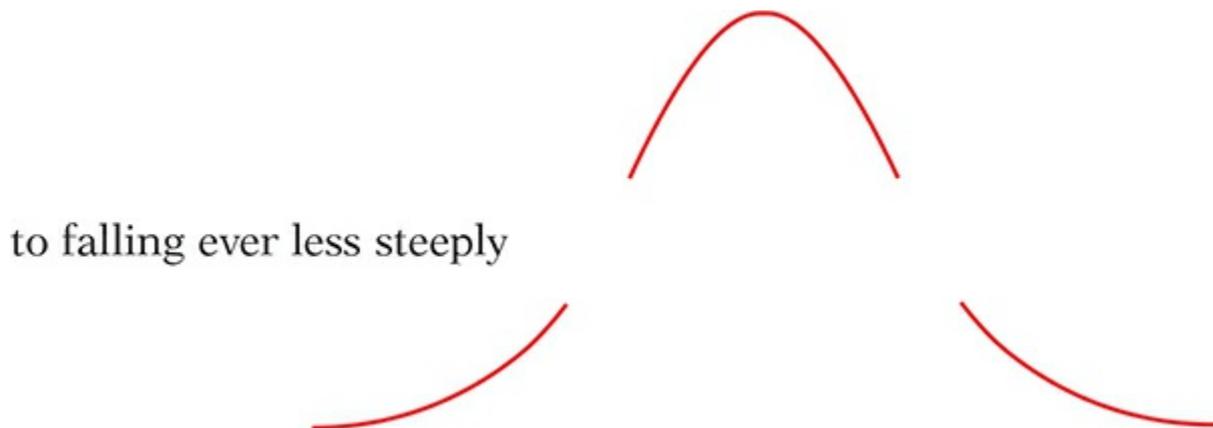
Normal distributions

The exact density curve for a particular Normal distribution is specified by giving the distribution's mean  $\mu$  and its standard deviation  $\sigma$ . The mean is located at the center of the symmetric curve and is the same as the median. Changing  $\mu$

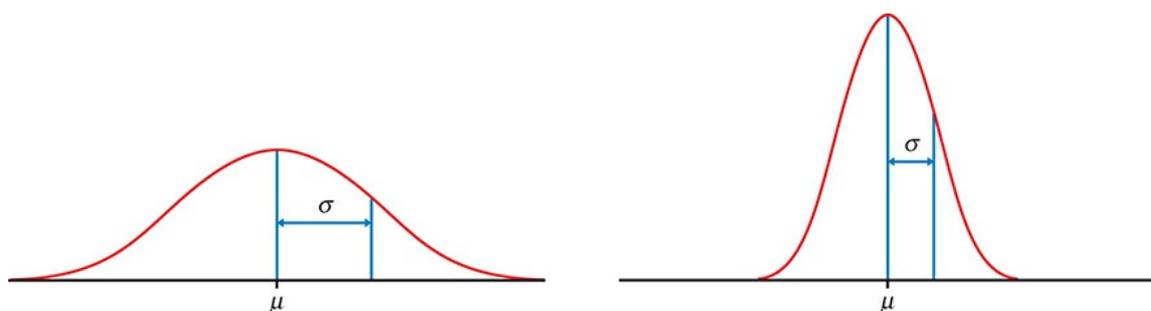
without changing  $\sigma$  moves the Normal curve along the horizontal axis without changing its spread.

The standard deviation  $\sigma$  controls the spread of a Normal curve. Figure 1.23 shows two Normal curves with different values of  $\sigma$ . The curve with the larger standard deviation is more spread out.

The standard deviation  $\sigma$  is the natural measure of spread for Normal distributions. Not only do  $\mu$  and  $\sigma$  completely determine the shape of a Normal curve, but we can locate  $\sigma$  by eye on the curve. Here's how. As we move out in either direction from the center  $\mu$ , the curve changes from falling ever more steeply



*The points at which this change of curvature takes place are located at distance  $\sigma$  on either side of the mean  $\mu$ . You can feel the change as you run your finger along a Normal curve, and so find the standard deviation. Remember that  $\mu$  and  $\sigma$  alone do not specify the shape of most distributions, and that the shape of density curves in general does not reveal  $\sigma$ . These are special properties of Normal distributions.*



**FIGURE 1.23**

Two Normal curves, showing the mean  $\mu$  and the standard deviation  $\sigma$ .

There are other symmetric bell-shaped density curves that are not Normal. The Normal density curves are specified by a particular equation. The height of the density curve at any point  $x$  is given by

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

We will not make direct use of this fact, although it is the basis of mathematical work with Normal distributions. Notice that the equation of the curve is completely

determined by the mean  $\mu$  and the standard deviation  $\sigma$ .

Why are the Normal distributions important in statistics? Here are three reasons.

1. Normal distributions are good descriptions for some distributions of *real data*. Distributions that are often close to Normal include scores on tests taken by many people (such as the Iowa Test of Figure 1.20, page 55), repeated careful measurements of the same quantity, and characteristics of biological populations (such as lengths of baby pythons and yields of corn).
2. Normal distributions are good approximations to the results of many kinds of *chance outcomes*, such as tossing a coin many times.
3. Many *statistical inference* procedures based on Normal distributions work well for other roughly symmetric distributions.

However, *even though many sets of data follow a Normal distribution, many do not*. Most income distributions, for example, are skewed to the right and so are not Normal. Non-Normal data, like nonnormal people, not only are common but are also sometimes more interesting than their Normal counterparts.



## The 68–95–99.7 rule

Although there are many Normal curves, they all have common properties. Here is one of the most important.

### THE 68–95–99.7 RULE

In the Normal distribution with mean  $\mu$  and standard deviation  $\sigma$ :

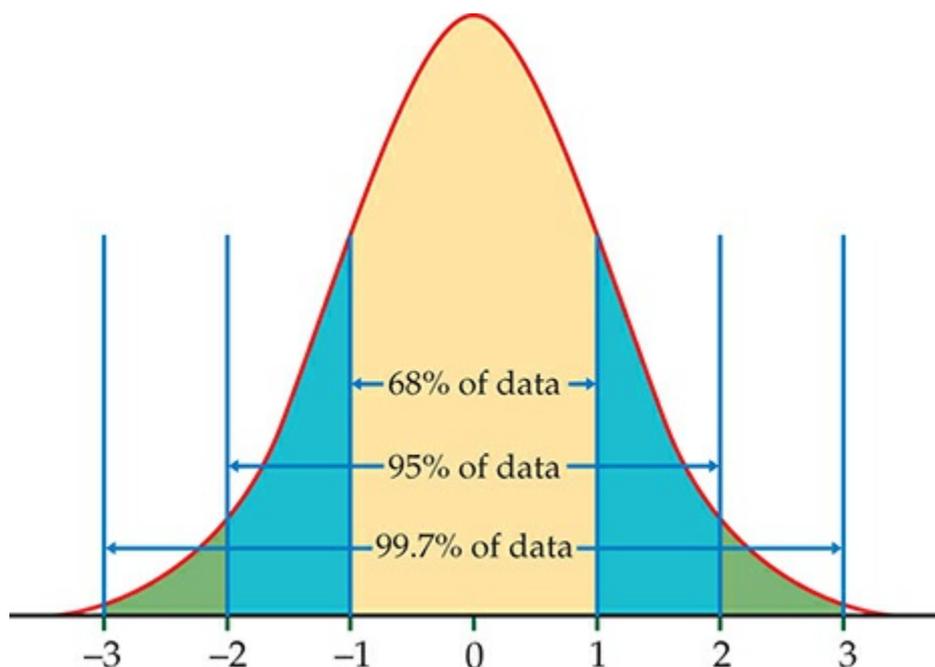
- Approximately **68%** of the observations fall within  $\sigma$  of the mean  $\mu$ .
- Approximately **95%** of the observations fall within  $2\sigma$  of  $\mu$ .
- Approximately **99.7%** of the observations fall within  $3\sigma$  of  $\mu$ .

Figure 1.24 illustrates the 68–95–99.7 rule. By remembering these three numbers, you can think about Normal distributions without constantly making detailed calculations.

## EXAMPLE

### 1.38 Heights of young women.

The distribution of heights of young women aged 18 to 24 is approximately Normal with mean  $\mu = 64.5$  inches and standard deviation  $\sigma = 2.5$  inches. Figure 1.25 shows what the 68–95–99.7 rule says about this distribution.



**FIGURE 1.24**

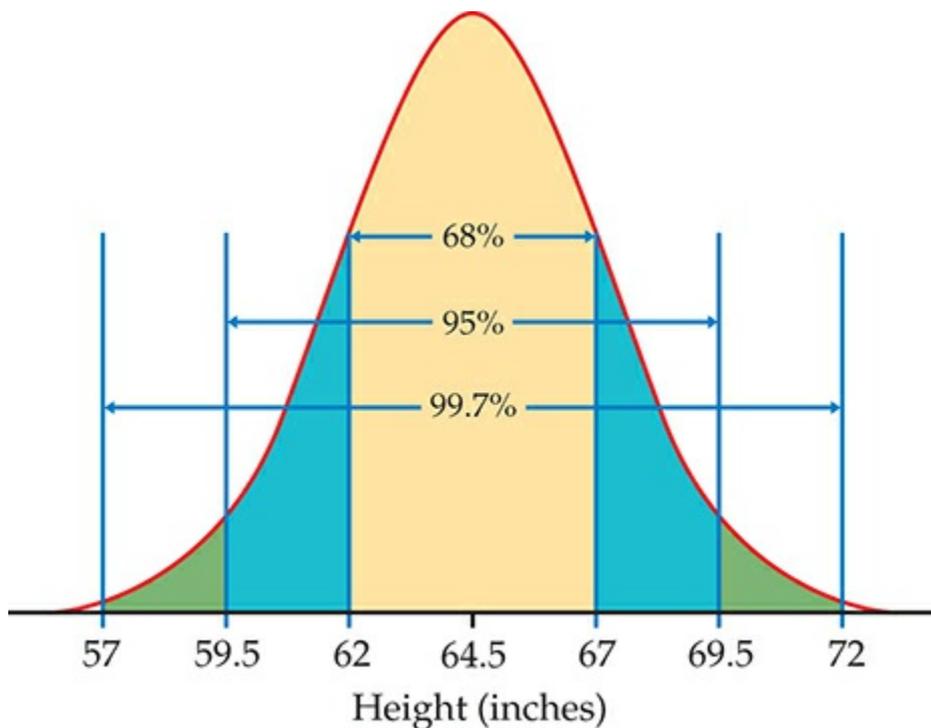
The 68–95–99.7 rule for Normal distributions.

Two standard deviations equals 5 inches for this distribution. The 95 part of the 68–95–99.7 rule says that the middle 95% of young women are between  $64.5 - 5$  and  $64.5 + 5$  inches tall, that is, between 59.5 and 69.5 inches. This fact is exactly true for an exactly Normal distribution. It is approximately true for the heights of young women because the distribution of heights is approximately Normal.

The other 5% of young women have heights outside the range from 59.5 to 69.5 inches. Because the Normal distributions are symmetric, half of these women are on the tall side. So the tallest 2.5% of young women are taller than 69.5 inches.

Because we will mention Normal distributions often, a short notation is helpful. We abbreviate the Normal distribution with mean  $\mu$  and standard deviation  $\sigma$  as  $N(\mu, \sigma)$ . For example, the distribution of young women's heights is  $N(64.5, 2.5)$ .

$N(\mu, \sigma)$



**FIGURE 1.25**

The 68–95–99.7 rule applied to the heights of young women, for Example 1.38.

## USE YOUR KNOWLEDGE

### 1.101 Test scores.

Many states assess the skills of their students in various grades. One program that is available for this purpose is the National Assessment of Educational Progress (NAEP).<sup>34</sup> One of the tests provided by the NAEP assesses the reading skills of twelfth-grade students. In a recent year, the national mean score was 288 and the standard deviation was 38. Assuming that these scores are approximately Normally distributed,  $N(288, 38)$ , use the 68–95–99.7 rule to give a range of scores that includes 95% of these students.

### 1.102 Use the 68–95–99.7 rule.

Refer to the previous exercise. Use the 68–95–99.7 rule to give a range of scores that includes 99.7% of these students.

## Standardizing observations

As the 68–95–99.7 rule suggests, all Normal distributions share many properties. In fact, all Normal distributions are the same if we measure in units of size  $\sigma$  about the mean  $\mu$  as center. Changing to these units is called *standardizing*. To standardize a value, subtract the mean of the distribution and then divide by the standard deviation.

### STANDARDIZING AND z-SCORES

If  $x$  is an observation from a distribution that has mean  $\mu$  and standard deviation  $\sigma$ , the **standardized value** of  $x$  is

$$z = \frac{x - \mu}{\sigma}$$

A standardized value is often called a **z-score**.

A  $z$ -score tells us how many standard deviations the original observation falls away from the mean, and in which direction. Observations larger than the mean are positive when standardized, and observations smaller than the mean are negative.

To compare scores based on different measures,  $z$ -scores can be very useful. For example, see Exercise 1.134 (page 75), where you are asked to compare an SAT score with an ACT score.

### EXAMPLE

#### 1.39 Find some $z$ -scores.

The heights of young women are approximately Normal with  $\mu = 64.5$  inches and  $\sigma = 2.5$  inches. The  $z$ -score for height is

$$z = \frac{\text{height} - 64.5}{2.5}$$

A woman's standardized height is the number of standard deviations by which her height differs from the mean height of all young women. A woman 68 inches tall, for example, has  $z$ -score

$$z = \frac{68 - 64.5}{2.5} = 1.4$$

or 1.4 standard deviations above the mean. Similarly, a woman 5 feet (60

inches) tall has  $z$ -score

$$z = 60 - 64.525 = 1.8$$

or 1.8 standard deviations less than the mean height.

## USE YOUR KNOWLEDGE

### 1.103 Find the $z$ -score.

Consider the NAEP scores (see Exercise 1.101), which we assume are approximately Normal,  $N(288, 38)$ . Give the  $z$ -score for a student who received a score of 365.

### 1.104 Find another $z$ -score.

Consider the NAEP scores, which we assume are approximately Normal,  $N(288, 38)$ . Give the  $z$ -score for a student who received a score of 250. Explain why your answer is negative even though all the test scores are positive.

We need a way to write variables, such as “height” in Example 1.38, that follow a theoretical distribution such as a Normal distribution. We use capital letters near the end of the alphabet for such variables. If  $X$  is the height of a young woman, we can then shorten “the height of a young woman is less than 68 inches” to “ $X < 68$ .” We will use lowercase  $x$  to stand for any specific value of the variable  $X$ .

We often standardize observations from symmetric distributions to express them in a common scale. We might, for example, compare the heights of two children of different ages by calculating their  $z$ -scores. The standardized heights tell us where each child stands in the distribution for his or her age group.

Standardizing is a linear transformation that transforms the data into the standard scale of  $z$ -scores. We know that a linear transformation does not change the shape of a distribution, and that the mean and standard deviation change in a simple manner. In particular, *the standardized values for any distribution always have mean 0 and standard deviation 1.*

If the variable we standardize has a Normal distribution, standardizing does more than give a common scale. It makes all Normal distributions into a single distribution, and this distribution is still Normal. Standardizing a variable that has any Normal distribution produces a new variable that has the *standard Normal distribution.*

## THE STANDARD NORMAL DISTRIBUTION

The **standard Normal distribution** is the Normal distribution  $N(0, 1)$  with mean 0 and standard deviation 1.

If a variable  $X$  has any Normal distribution  $N(\mu, \sigma)$  with mean  $\mu$  and standard deviation  $\sigma$ , then the standardized variable

$$z = \frac{X - \mu}{\sigma}$$

has the standard Normal distribution.

### Normal distribution calculations

Areas under a Normal curve represent proportions of observations from that Normal distribution. There is no formula for areas under a Normal curve. Calculations use either software that calculates areas or a table of areas. The table and most software calculate one kind of area: ***cumulative proportions***. A cumulative proportion is the proportion of observations in a distribution that lie at or below a given value. When the distribution is given by a density curve, the cumulative proportion is the area under the curve to the left of a given value. Figure 1.26 shows the idea more clearly than words do.

#### cumulative proportion

The key to calculating Normal proportions is to match the area you want with areas that represent cumulative proportions. Then get areas for cumulative proportions either from software or (with an extra step) from a table. The following examples show the method in pictures.

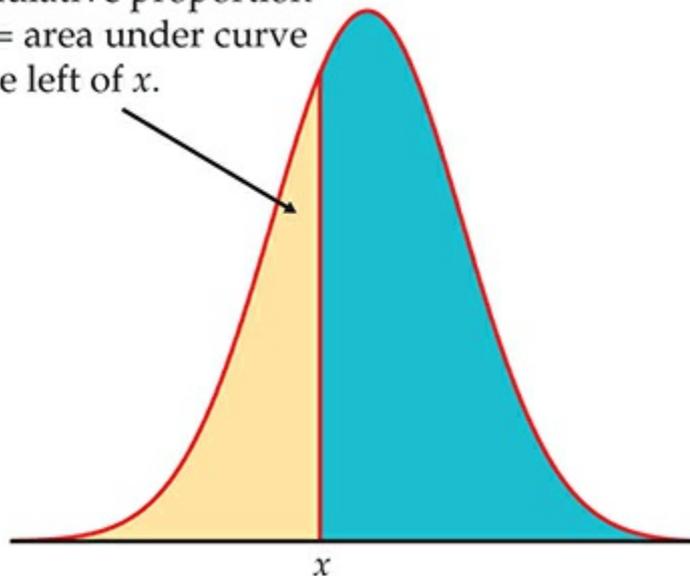
### EXAMPLE

#### 1.40 NCAA eligibility for competition.



To be eligible to compete in their first year of college, the National Collegiate Athletic Association (NCAA) requires Division I athletes to meet certain academic standards. These are based on their grade point average (GPA) in certain courses and combined scores on the SAT Critical Reading and Mathematics sections or the ACT composite score.<sup>35</sup>

Cumulative proportion  
at  $x$  = area under curve  
to the left of  $x$ .



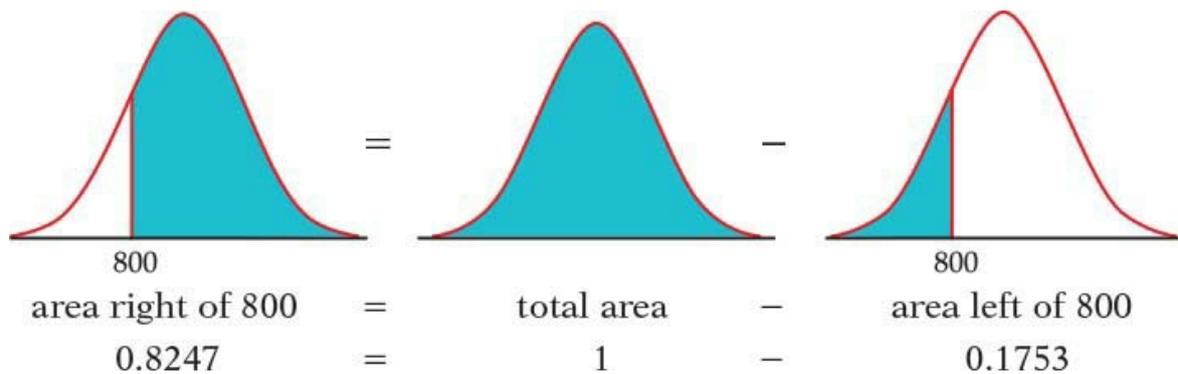
**FIGURE 1.26**

The *cumulative proportion* for a value  $x$  is the proportion of all observations from the distribution that are less than or equal to  $x$ . This is the area to the left of  $x$  under the Normal curve.

For a student with a 3.0 GPA, the combined SAT score must be 800 or

higher. Based on the distribution of SAT scores for college-bound students, we assume that the distribution of the combined Critical Reading and Mathematics scores is approximately Normal with mean 1010 and standard deviation 225.<sup>36</sup> What proportion of college-bound students have SAT scores of 800 or more?

Here is the calculation in pictures: the proportion of scores above 800 is the area under the curve to the right of 800. That's the total area under the curve (which is always 1) minus the cumulative proportion up to 800.



That is, the proportion of college-bound SAT takers with a 3.0 GPA who are eligible to compete is 0.8247, or about 82%.

There is *no* area under a smooth curve that is exactly over the point 800. Consequently, the area to the right of 800 (the proportion of scores  $> 800$ ) is the same as the area at or to the right of this point (the proportion of scores  $\geq 800$ ). The actual data may contain a student who scored exactly 800 on the SAT. That the proportion of scores exactly equal to 800 is 0 for a Normal distribution is a consequence of the idealized smoothing of Normal distributions for data.

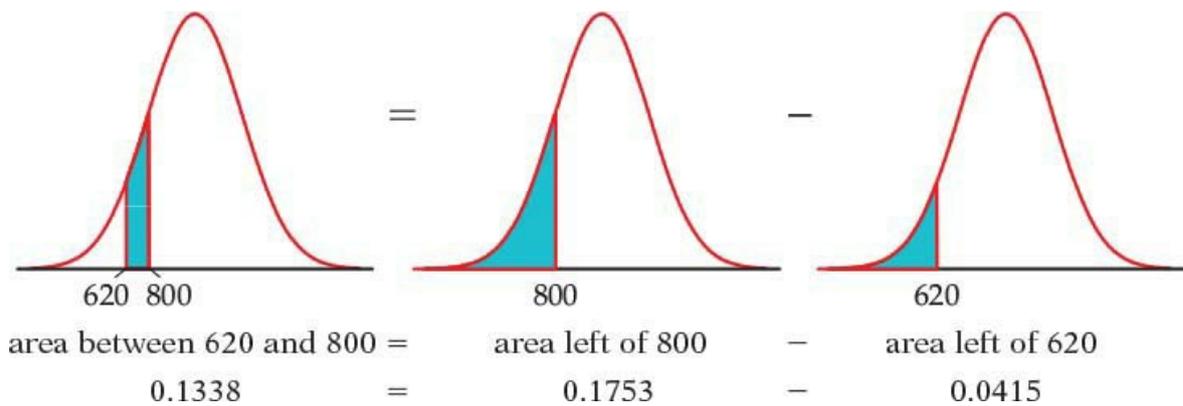
## EXAMPLE

### 1.41 NCAA eligibility for aid and practice.

The NCAA has a category of eligibility in which a first-year student may not compete but is still eligible to receive an athletic scholarship and to practice with the team. The requirements for this category are a 3.0 GPA and combined SAT Critical Reading and Mathematics scores of at least 620.

What proportion of college-bound students who take the SAT would be eligible to receive an athletic scholarship and to practice with the team but would not be eligible to compete? That is, what proportion have scores

between 620 and 800? Here are the pictures:



About 13% of college-bound students with a 3.0 GPA have SAT scores between 620 and 800.

How do we find the numerical values of the areas in Examples 1.40 and 1.41? If you use software, just plug in mean 1010 and standard deviation 225. Then ask for the cumulative proportions for 800 and for 620. (Your software will probably refer to these as “cumulative probabilities.” We will learn in Chapter 4 why the language of probability fits.) Sketches of the areas that you want similar to the ones in Examples 1.40 and 1.41 are very helpful in making sure that you are doing the correct calculations.

You can use the *Normal Curve* applet on the text website, [whfreeman.com/ips8e](http://whfreeman.com/ips8e), to find Normal proportions. The applet is more flexible than most software—it will find any Normal proportion, not just cumulative proportions. The applet is an excellent way to understand Normal curves. But, because of the limitations of web browsers, the applet is not as accurate as statistical software.



If you are not using software, you can find cumulative proportions for Normal curves from a table. That requires an extra step, as we now explain.

### Using the standard Normal table

The extra step in finding cumulative proportions from a table is that we must first standardize to express the problem in the standard scale of *z*-scores. This allows us to get by with just one table, a table of *standard Normal cumulative proportions*. Table A in the back of the book gives standard Normal probabilities. Table A also appears on the last two pages of the text. The picture at the top of the table reminds us that the entries are cumulative proportions, areas under the curve to the left of a value *z*.

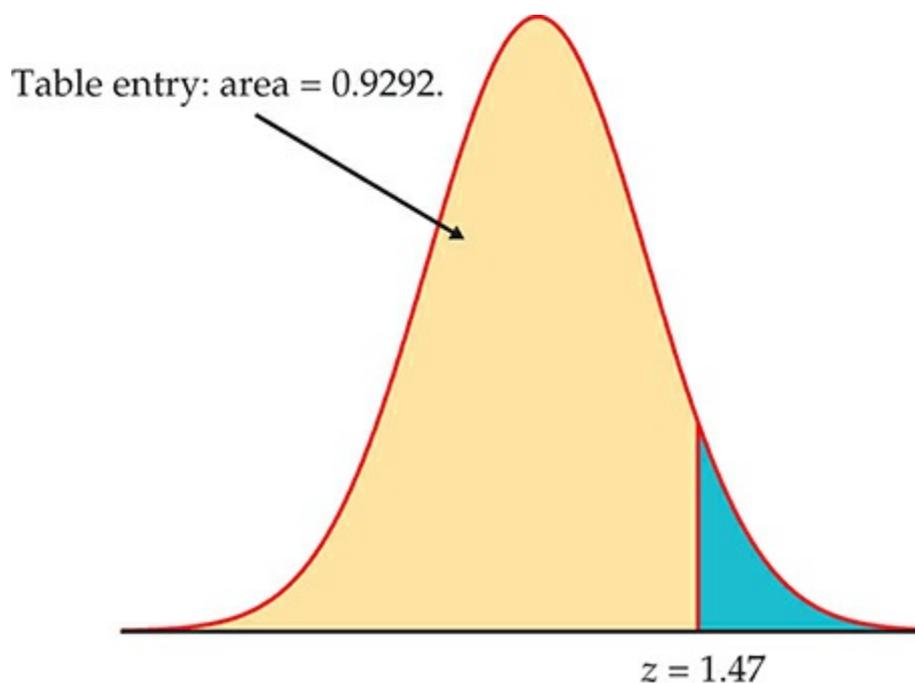
## EXAMPLE

### 1.42 Find the proportion from $z$ .

What proportion of observations on a standard Normal variable  $Z$  take values less than 1.47?

*Solution:* To find the area to the left of 1.47, locate 1.4 in the left-hand column of Table A and then locate the remaining digit 7 as .07 in the top row. The entry opposite 1.4 and under .07 is 0.9292. This is the cumulative proportion we seek. Figure 1.27 illustrates this area.

Now that you see how Table A works, let's redo the NCAA Examples 1.40 and 1.41 using the table.



**FIGURE 1.27**

The area under a standard Normal curve to the left of the point  $z = 1.47$  is 0.9292, for Example 1.42.

## EXAMPLE

### 1.43 Find the proportion from $x$ .

What proportion of college-bound students who take the SAT have scores of at least 800? The picture that leads to the answer is exactly the same as in Example 1.40. The extra step is that we first standardize to read cumulative proportions from Table A. If  $X$  is SAT score, we want the proportion of students for whom  $X \geq x$ , where  $x = 800$ .

1. *Standardize.* Subtract the mean, then divide by the standard deviation, to transform the problem about  $X$  into a problem about a standard Normal  $Z$ :

$$X \geq 800$$

$$X - 1010225 \geq 800 - 1010225$$

$$Z \geq -0.93$$

2. *Use the table.* Look at the pictures in Example 1.40. From Table A, we see that the proportion of observations less than  $-0.93$  is 0.1762. The area to the right of  $-0.93$  is therefore  $1 - 0.1762 = 0.8238$ . This is about 82%.

The area from the table in Example 1.43 (0.8238) is slightly less accurate than the area from software in Example 1.40 (0.8247) because we must round  $z$  to two places when we use Table A. The difference is rarely important in practice.

## EXAMPLE

### 1.44 Eligibility for aid and practice.

What proportion of all students who take the SAT would be eligible to receive athletic scholarships and to practice with the team but would not be eligible to compete in the eyes of the NCAA? That is, what proportion of students have SAT scores between 620 and 800? First, sketch the areas, exactly as in Example 1.41. We again use  $X$  as shorthand for an SAT score.

1. *Standardize.*

$$620 \leq X < 800$$

$$620 - 1010225 \leq X - 1010225 < 800 - 1010225$$

$$-1.73 \leq Z < -0.93$$

2. *Use the table.*

area between  $-1.73$  and  $-0.93 = (\text{area left of } -0.93) - (\text{area left of } -1.73)$

$$= 0.1762 - 0.0418 = 0.1344$$

As in Example 1.41, about 13% of students would be eligible to receive athletic scholarships and to practice with the team.

Sometimes we encounter a value of  $z$  more extreme than those appearing in Table A. For example, the area to the left of  $z = -4$  is not given in the table. The  $z$ -values in Table A leave only area 0.0002 in each tail unaccounted for. For practical purposes, we can act as if there is zero area outside the range of Table A.

## USE YOUR KNOWLEDGE

### 1.105 Find the proportion.

Consider the NAEP scores, which are approximately Normal,  $N(288, 38)$ . Find the proportion of students who have scores less than 340. Find the proportion of students who have scores greater than or equal to 340. Sketch the relationship between these two calculations using pictures of Normal curves similar to the ones given in Example 1.40 (page 63).

### 1.106 Find another proportion.

Consider the NAEP scores, which are approximately Normal,  $N(288, 38)$ . Find the proportion of students who have scores between 340 and 370. Use pictures of Normal curves similar to the ones given in Example 1.41 (page 64) to illustrate your calculations.

## Inverse Normal calculations

Examples 1.40 to 1.44 illustrate the use of Normal distributions to find the proportion of observations in a given event, such as “SAT score between 620 and 800.” We may instead want to find the observed value corresponding to a given proportion.

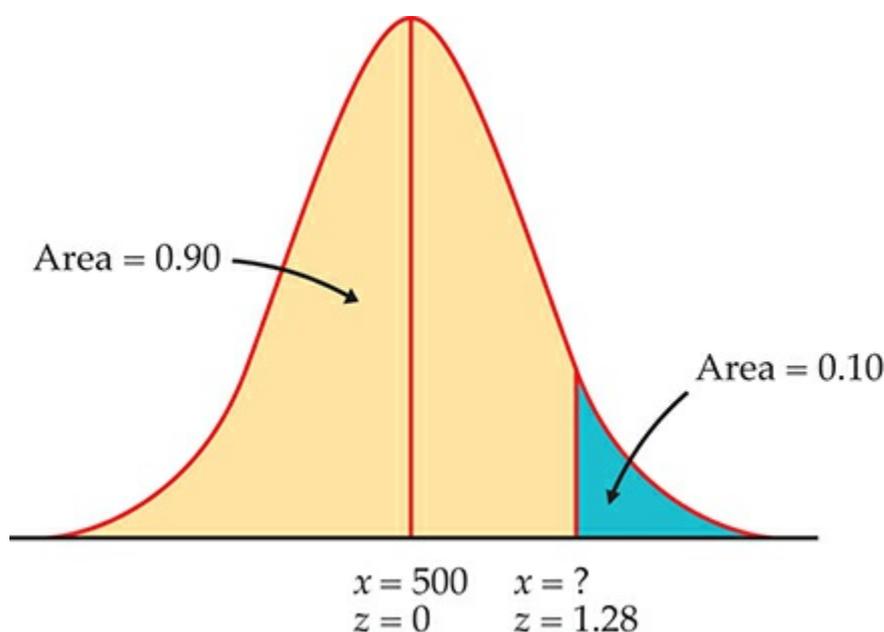
Statistical software will do this directly. Without software, use Table A backward, finding the desired proportion in the body of the table and then reading the corresponding  $z$  from the left column and top row.

## EXAMPLE

### 1.45 How high for the top 10%?

Scores for college-bound students on the SAT Critical Reading test in recent years follow approximately the  $N(500,120)$  distribution.<sup>37</sup> How high must a student score to place in the top 10% of all students taking the SAT?

Again, the key to the problem is to draw a picture. Figure 1.28 shows that we want the score  $x$  with an area of 0.10 above it. That's the same as area below  $x$  equal to 0.90.



**FIGURE 1.28**

Locating the point on a Normal curve with area 0.10 to its right, for Example 1.45.

Statistical software has a function that will give you the  $x$  for any cumulative proportion you specify. The function often has a name such as “inverse cumulative probability.” Plug in mean 500, standard deviation 120, and cumulative proportion 0.9. The software tells you that  $x = 653.786$ . We see that a student must score at least 654 to place in the highest 10%.

Without software, first find the standard score  $z$  with cumulative proportion 0.9, then “unstandardize” to find  $x$ . Here is the two-step process:

1. *Use the table.* Look in the body of Table A for the entry closest to 0.9. It is 0.8997. This is the entry corresponding to  $z = 1.28$ . So  $z = 1.28$  is the standardized value with area 0.9 to its left.
2. *Unstandardize* to transform the solution from  $z$  back to the original  $x$  scale. We know that the standardized value of the unknown  $x$  is  $z = 1.28$ . So  $x$

itself satisfies

$$x - 500 = 1.28(120)$$

Solving this equation for  $x$  gives

$$x = 500 + (1.28)(120) = 653.6$$

This equation should make sense: it finds the  $x$  that lies 1.28 standard deviations above the mean on this particular Normal curve. That is the “unstandardized” meaning of  $z = 1.28$ . The general rule for unstandardizing a  $z$ -score is

$$x = \mu + z\sigma$$

## USE YOUR KNOWLEDGE

### 1.107 What score is needed to be in the top 25%?

Consider the NAEP scores, which are approximately Normal,  $N(288, 38)$ . How high a score is needed to be in the top 25% of students who take this exam?

### 1.108 Find the score that 80% of students will exceed.

Consider the NAEP scores, which are approximately Normal,  $N(288, 38)$ . Eighty percent of the students will score above  $x$  on this exam. Find  $x$ .

## Normal quantile plots

The Normal distributions provide good descriptions of some distributions of real data, such as the Iowa Test vocabulary scores. The distributions of some other common variables are usually skewed and therefore distinctly non-Normal. Examples include economic variables such as personal income and gross sales of business firms, the survival times of cancer patients after treatment, and the service lifetime of mechanical or electronic components. While experience can suggest whether or not a Normal distribution is plausible in a particular case, it is risky to assume that a distribution is Normal without actually inspecting the data.

A histogram or stemplot can reveal distinctly non-Normal features of a distribution, such as outliers, pronounced skewness, or gaps and clusters. If the

stemplot or histogram appears roughly symmetric and unimodal, however, we need a more sensitive way to judge the adequacy of a Normal model. The most useful tool for assessing Normality is another graph, the *Normal quantile plot*.

### Normal quantile plot

Here is the basic idea of a Normal quantile plot. The graphs produced by software use more sophisticated versions of this idea. It is not practical to make Normal quantile plots by hand.

1. Arrange the observed data values from smallest to largest. Record what percentile of the data each value occupies. For example, the smallest observation in a set of 20 is at the 5% point, the second smallest is at the 10% point, and so on.
2. Do Normal distribution calculations to find the values of  $z$  corresponding to these same percentiles. For example,  $z = -1.645$  is the 5% point of the standard Normal distribution, and  $z = -1.282$  is the 10% point. We call these values of  $Z$  *Normal scores*.

### Normal scores

3. Plot each data point  $x$  against the corresponding Normal score. If the data distribution is close to any Normal distribution, the plotted points will lie close to a straight line.

Any Normal distribution produces a straight line on the plot because standardizing turns any Normal distribution into a standard Normal distribution. Standardizing is a linear transformation that can change the slope and intercept of the line in our plot but cannot turn a line into a curved pattern.

## USE OF NORMAL QUANTILE PLOTS

If the points on a **Normal quantile plot** lie close to a straight line, the plot indicates that the data are Normal. Systematic deviations from a straight line indicate a non-Normal distribution. Outliers appear as points that are far away from the overall pattern of the plot. An optional line can be drawn on the plot that corresponds to the Normal distribution with mean equal to the mean of the data and standard deviation equal to the standard deviation of the data.

Figures 1.29 and 1.30 are Normal quantile plots for data we have met earlier. The data  $x$  are plotted vertically against the corresponding standard Normal  $z$ -score plotted horizontally. The  $z$ -score scale generally extends from  $-3$  to  $3$  because

almost all of a standard Normal curve lies between these values. These figures show how Normal quantile plots behave.

## EXAMPLE

### 1.46 IQ scores are approximately Normal.



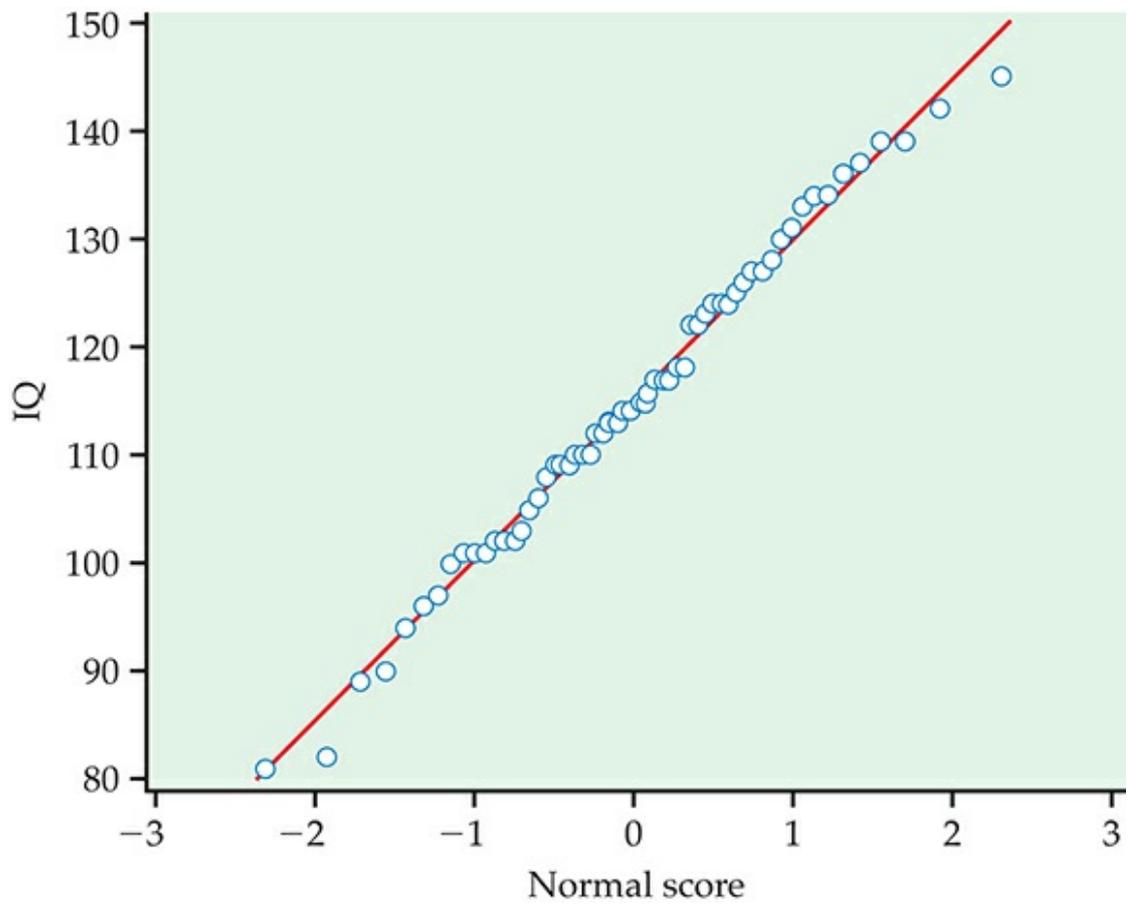
Figure 1.29 is a Normal quantile plot of the 60 fifth-grade IQ scores from Table 1.1 (page 16). The points lie very close to the straight line drawn on the plot. We conclude that the distribution of IQ data is approximately Normal.

## EXAMPLE

### 1.47 Times to start a business are skewed.



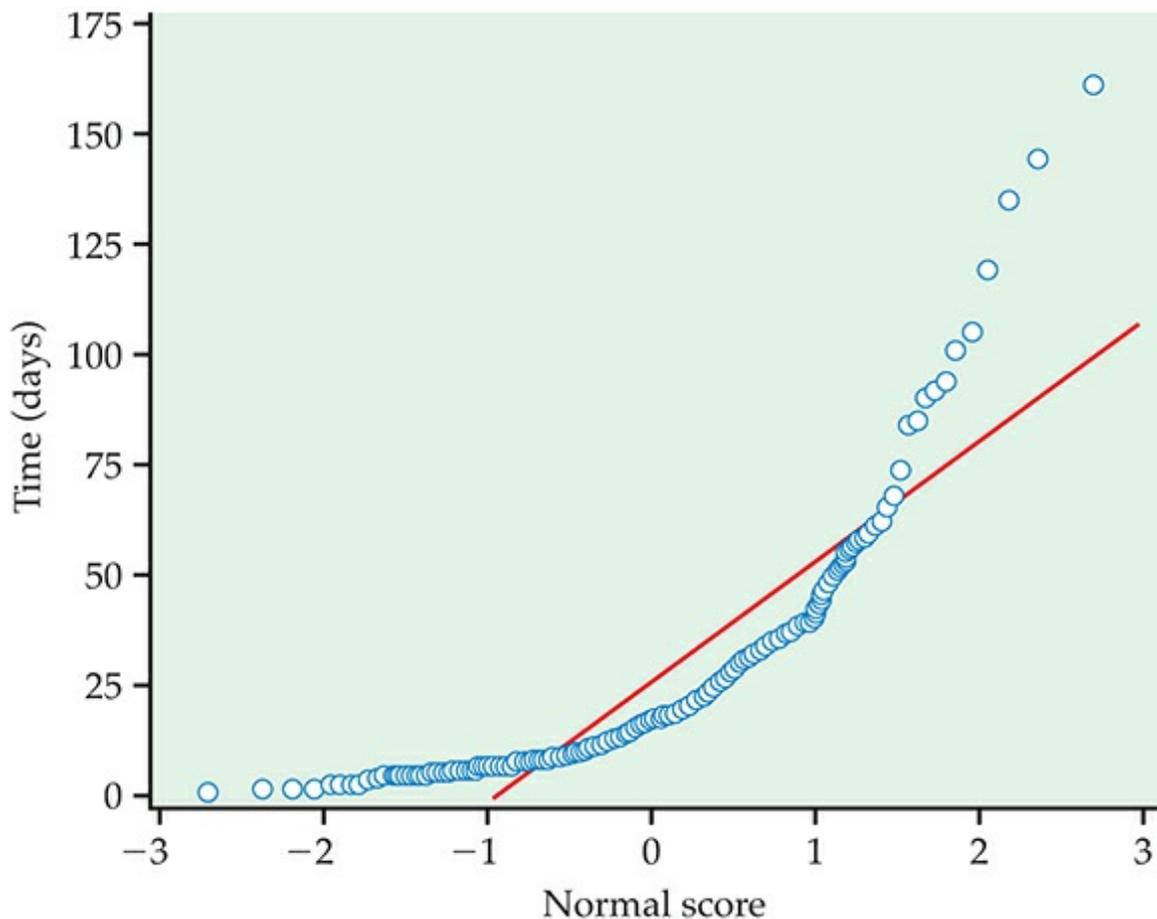
Figure 1.30 is a Normal quantile plot of the data on times to start a business from Example 1.23. We have excluded Suriname, the outlier that you examined in Exercise 1.47. The line drawn on the plot shows clearly that the plot of the data is curved. We conclude that these data are not Normally distributed. The shape of the curve is what we typically see with a distribution that is strongly skewed to the right.



**FIGURE 1.29**

Normal quantile plot of IQ scores, for Example 1.46. This distribution is approximately Normal.





**FIGURE 1.30**

Normal quantile plot of 184 times to start a business, with the outlier, Suriname, excluded, for Example 1.47. This distribution is highly skewed.

Real data often show some departure from the theoretical Normal model. *When you examine a Normal quantile plot, look for shapes that show clear departures from Normality. Don't overreact to minor wiggles in the plot.* When we discuss statistical methods that are based on the Normal model, we are interested in whether or not the data are sufficiently Normal for these procedures to work properly. We are not concerned about minor deviations from Normality. Many common methods work well as long as the data are approximately Normal and outliers are not present.

## BEYOND THE BASICS

### Density Estimation

A density curve gives a compact summary of the overall shape of a

distribution. Many distributions do not have the Normal shape. There are other families of density curves that are used as mathematical models for various distribution shapes. Modern software offers more flexible options. A *density estimator* does not start with any specific shape, such as the Normal shape. It looks at the data and draws a density curve that describes the overall shape of the data. Density estimators join stemplots and histograms as useful graphical tools for exploratory data analysis.

### density estimator

Density estimates can capture other unusual features of a distribution. Here is an example.

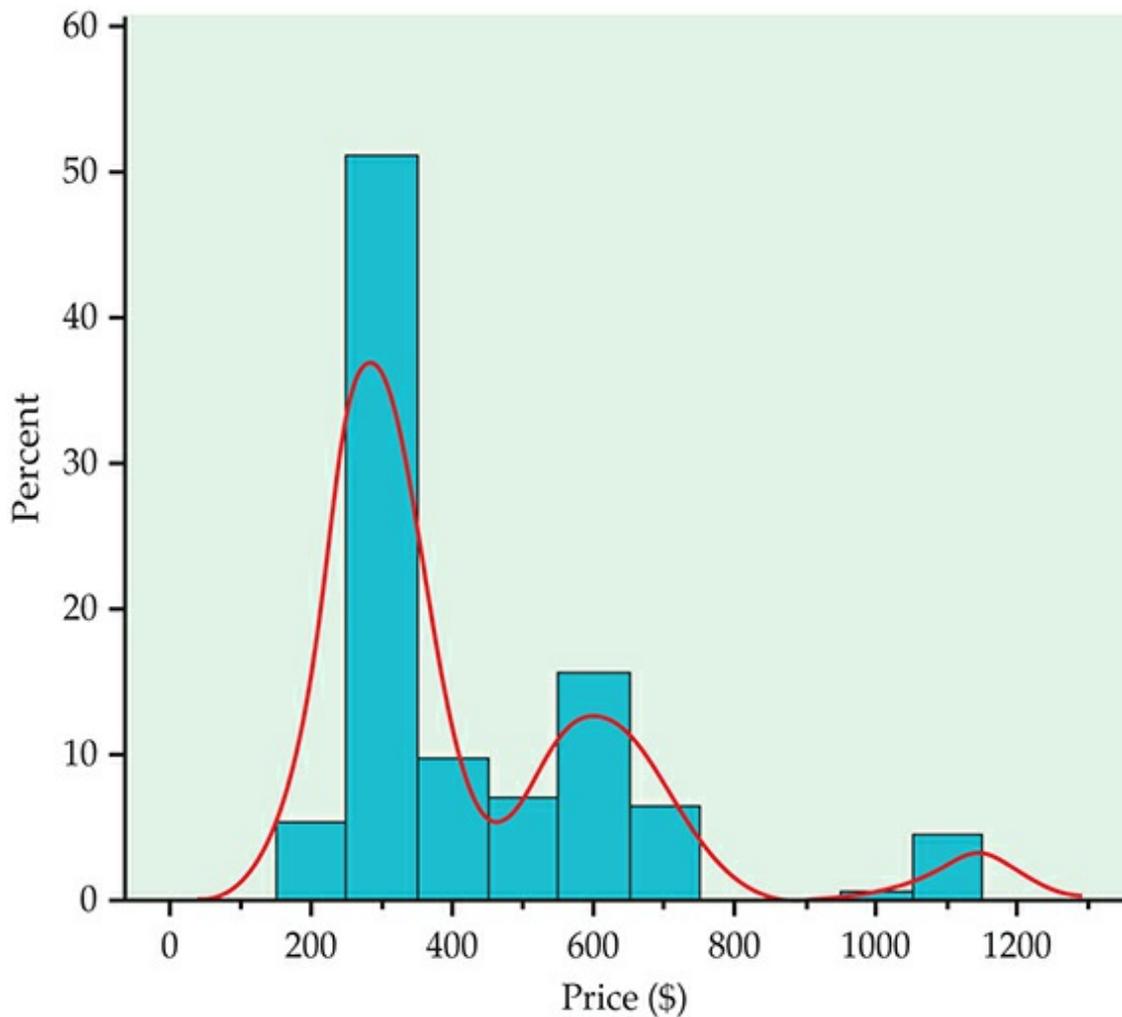
## EXAMPLE

### 1.48 StubHub!



STUBHUB

StubHub! is a website where fans can buy and sell tickets to sporting events. Ticket holders wanting to sell their tickets provide the location of their seats and the selling price. People wanting to buy tickets can choose from among the tickets offered for a given event.<sup>38</sup>



**FIGURE 1.31**

Histogram of StubHub! price per seat for tickets to the 2013 NCAA Women's Final Four Basketball Championship in New Orleans, with a density estimate, for Example 1.48.

There were 186 tickets for the NCAA Women's Final Four Basketball Championship in New Orleans posted for sale on StubHub! on January 2, 2013. A histogram with a density estimate is given in Figure 1.31. The distribution has three peaks, one around \$300, another around \$600, and a third around \$1100. Inspection of the data suggests that these correspond roughly to three different types of seats: lower-level seats, club seats, and special luxury seats.

Many distributions that we have met have a single peak, or mode. The distribution described in Example 1.48 has three modes and is called a *trimodal distribution*. A distribution that has two modes is called a *bimodal distribution*.

trimodal distribution

bimodal distribution

The previous example reminds of a continuing theme for data analysis. We looked at a histogram and a density estimate and saw something interesting. This led us to speculation. Additional data on the type and location of the seats may explain more about the prices than we see in Figure 1.31.

## SECTION 1.4 Summary

The overall pattern of a distribution can often be described compactly by a **density curve**. A density curve has total area 1 underneath it. Areas under a density curve give proportions of observations for the distribution.

The **mean**  $\mu$  (balance point), the **median** (equal-areas point), and the **quartiles** can be approximately located by eye on a density curve. The **standard deviation**  $\sigma$  cannot be located by eye on most density curves. The mean and median are equal for symmetric density curves, but the mean of a skewed curve is located farther toward the long tail than is the median.

The **Normal distributions** are described by bell-shaped, symmetric, unimodal density curves. The mean  $\mu$  and standard deviation  $\sigma$  completely specify the Normal distribution  $N(\mu, \sigma)$ . The mean is the center of symmetry, and  $\sigma$  is the distance from  $\mu$  to the change-of-curvature points on either side. All Normal distributions satisfy the **68–95–99.7 rule**.

To **standardize** any observation  $x$ , subtract the mean of the distribution and then divide by the standard deviation. The resulting **z-score**  $z = (x - \mu)/\sigma$  says how many standard deviations  $x$  lies from the distribution mean. All Normal distributions are the same when measurements are transformed to the standardized scale.

If  $X$  has the  $N(\mu, \sigma)$  distribution, then the standardized variable  $Z = (X - \mu)/\sigma$  has the **standard Normal distribution**  $N(0, 1)$ . Proportions for any Normal distribution can be calculated by software or from the **standard Normal table** (Table A), which gives the **cumulative proportions** of  $Z < z$  for many values of  $z$ .

The adequacy of a Normal model for describing a distribution of data is best assessed by a **Normal quantile plot**, which is available in most statistical software packages. A pattern on such a plot that deviates substantially from a straight line indicates that the data are not Normal.

## SECTION 1.4 Exercises

*For Exercises 1.101 and 1.102, see page 61; for Exercises 1.103 and 1.104, see page 62; for Exercises 1.105 and 1.106, see page 67; and for Exercises 1.107 and 1.108, see page 68.*

### 1.109 Means and medians.

- Sketch a symmetric distribution that is *not* Normal. Mark the location of the mean and the median.
- Sketch a distribution that is skewed to the left. Mark the location of the mean and the median.

### 1.110 The effect of changing the standard deviation.

- (a) Sketch a Normal curve that has mean 20 and standard deviation 5.
- (b) On the same  $x$  axis, sketch a Normal curve that has mean 20 and standard deviation 10.
- (c) How does the Normal curve change when the standard deviation is varied but the mean stays the same?

### 1.111 The effect of changing the mean.

- (a) Sketch a Normal curve that has mean 20 and standard deviation 5.
- (b) On the same  $x$  axis, sketch a Normal curve that has mean 30 and standard deviation 5.
- (c) How does the Normal curve change when the mean is varied but the standard deviation stays the same?

### 1.112 NAEP music scores.

In Exercise 1.101 (page 61) we examined the distribution of NAEP scores for the twelfth-grade reading skills assessment. For eighth-grade students the average music score is approximately Normal with mean 150 and standard deviation 35.

- (a) Sketch this Normal distribution.
- (b) Make a table that includes values of the scores corresponding to plus or minus one, two, and three standard deviations from the mean. Mark these points on your sketch along with the mean.
- (c) Apply the 68–95–99.7 rule to this distribution. Give the ranges of reading score values that are within one, two, and three standard deviations of the mean.

### 1.113 NAEP U.S. history scores.

Refer to the previous exercise. The scores for twelfth-grade students on the U.S. history assessment are approximately  $N(288, 32)$ . Answer the questions in the previous exercise for this assessment.

### 1.114 Standardize some NAEP music scores.

The NAEP music assessment scores for eighth-grade students are approximately  $N(150, 35)$ . Find  $z$ -scores by standardizing the following scores: 150, 140, 100, 180, 230.

### 1.115 Compute the percentile scores.

Refer to the previous exercise. When scores such as the NAEP assessment scores are reported for individual students, the actual values of the scores are not particularly meaningful. Usually, they are transformed into percentile scores. The percentile score is the proportion of students who would score less than or equal to the score for the individual student. Compute the percentile scores for the five scores in the previous exercise. State whether you used software or Table A for these computations.

### 1.116 Are the NAEP U.S. history scores approximately Normal?

In Exercise 1.113, we assumed that the NAEP U.S. history scores for twelfth-grade students are approximately Normal with the reported mean and standard deviation,  $N(288, 32)$ . Let's check that assumption. In addition to means and standard deviations, you can find selected percentiles for the NAEP assessments (see previous exercise). For the twelfth-grade U.S. history scores, the following percentiles are reported:

Percentile	Score
10%	246
25%	276
50%	290
75%	311
90%	328

Use these percentiles to assess whether or not the NAEP U.S. history scores for twelfth-grade students are approximately Normal. Write a short report describing your methods and conclusions.

### 1.117 Are the NAEP mathematics scores approximately Normal?

Refer to the previous exercise. For the NAEP mathematics scores for twelfth-graders the mean is 153 and the standard deviation is 34. Here are the reported percentiles:

Percentile	Score
10%	110
25%	130
50%	154
75%	177
90%	197

Is the  $N(153, 34)$  distribution a good approximation for the NAEP mathematics scores? Write a short report describing your methods and conclusions.

### 1.118 Do women talk more?

Conventional wisdom suggests that women are more talkative than men. One study designed to examine this stereotype collected data on the speech of 42 women and 37 men in the United States.<sup>39</sup>  TALK

- The mean number of words spoken per day by the women was 14,297 with a standard deviation of 6441. Use the 68–95–99.7 rule to describe this distribution.
- Do you think that applying the rule in this situation is reasonable? Explain your answer.
- The men averaged 14,060 words per day with a standard deviation of 9056. Answer the questions in parts (a) and (b) for the men.
- Do you think that the data support the conventional wisdom? Explain your answer. Note that in Section 7.2 we will learn formal statistical methods to answer this type of question.

### 1.119 Data from Mexico.

Refer to the previous exercise. A similar study in Mexico was conducted with 31 women and 20 men. The women averaged 14,704 words per day with a standard deviation of 6215. For men the mean was 15,022

and the standard deviation was 7864.  TALKM

- (a) Answer the questions from the previous exercise for the Mexican study.
- (b) The means for both men and women are higher for the Mexican study than for the U.S. study. What conclusions can you draw from this observation?

### 1.120 Total scores.

Here are the total scores of 10 students in an introductory statistics course:

62 93 54 76 73 98 64 55 80 71

Previous experience with this course suggests that these scores should come from a distribution that is approximately Normal with mean 72 and standard deviation 10.

- (a) Using these values for  $\mu$  and  $\sigma$ , standardize the scores of these 10 students.
- (b) If the grading policy is to give a grade of A to the top 15% of scores based on the Normal distribution with mean 72 and standard deviation 10, what is the cutoff for an A in terms of a standardized score?
- (c) Which of the 10 students earned a grade of A in the course? Show your work.

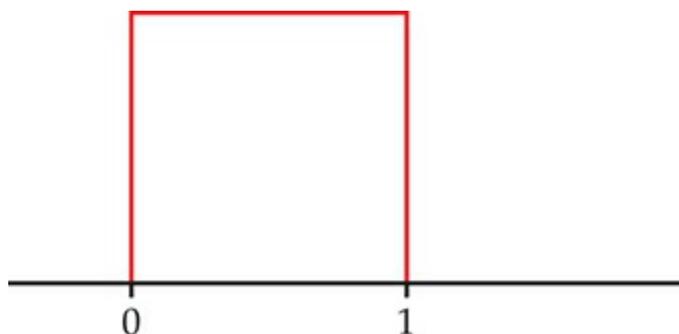
### 1.121 Assign more grades.

Refer to the previous exercise. The grading policy says that the cutoffs for the other grades correspond to the following: bottom 5% receive F, next 10% receive D, next 40% receive C, and next 30% receive B. These cutoffs are based on the  $N(72, 10)$  distribution.

- (a) Give the cutoffs for the grades in this course in terms of standardized scores.
- (b) Give the cutoffs in terms of actual total scores.
- (c) Do you think that this method of assigning grades is a good one? Give reasons for your answer.

### 1.122 A uniform distribution.

If you ask a computer to generate “random numbers” between 0 and 1, you will get observations from a **uniform distribution**. Figure 1.32 graphs the density curve for a uniform distribution. Use areas under this density curve to answer the following questions.



**FIGURE 1.32**

The density curve of a uniform distribution, for Exercise 1.122.

- (a) Why is the total area under this curve equal to 1?
- (b) What proportion of the observations lie below 0.34?
- (c) What proportion of the observations lie between 0.34 and 0.60?

**1.123 Use a different range for the uniform distribution.**

Many random number generators allow users to specify the range of the random numbers to be produced. Suppose that you specify that the outcomes are to be distributed uniformly between 0 and 5. Then the density curve of the outcomes has constant height between 0 and 5, and height 0 elsewhere.

- (a) What is the height of the density curve between 0 and 5? Draw a graph of the density curve.
- (b) Use your graph from (a) and the fact that areas under the curve are proportions of outcomes to find the proportion of outcomes that are less than 1.
- (c) Find the proportion of outcomes that lie between 0.5 and 2.5.

**1.124 Find the mean, the median, and the quartiles.**

What are the mean and the median of the uniform distribution in Figure 1.32? What are the quartiles?

**1.125 Three density curves.**

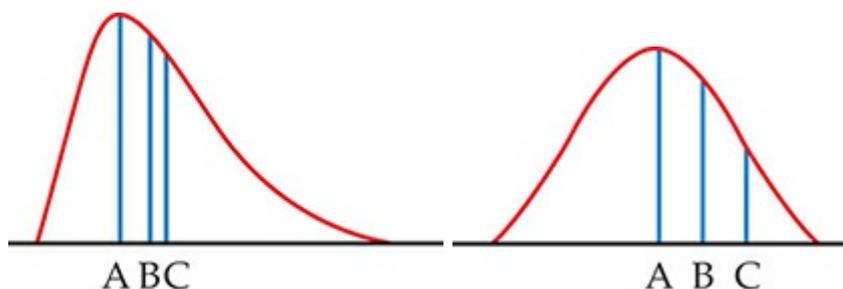
Figure 1.33 displays three density curves, each with three points marked on it. At which of these points on each curve do the mean and the median fall?

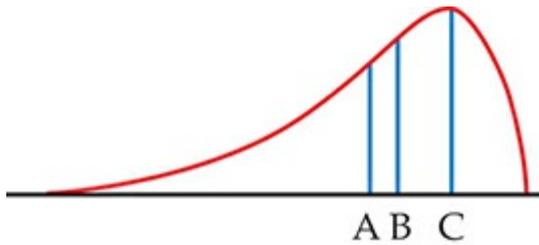
 **1.126 Use the *Normal Curve* applet.**

Use the *Normal Curve* applet for the standard Normal distribution to say how many standard deviations above and below the mean the quartiles of any Normal distribution lie.

 **1.127 Use the *Normal Curve* applet.**

The 68–95–99.7 rule for Normal distributions is a useful approximation. You can use the *Normal Curve* applet on the text website, [whfreeman.com/ips8e](http://whfreeman.com/ips8e), to see how accurate the rule is. Drag one flag across the other so that the applet shows the area under the curve between the two flags.





**FIGURE 1.33**

Three density curves, for Exercise 1.125.

- (a) Place the flags one standard deviation on either side of the mean. What is the area between these two values? What does the 68–95–99.7 rule say this area is?
- (b) Repeat for locations two and three standard deviations on either side of the mean. Again compare the 68–95–99.7 rule with the area given by the applet.

### 1.128 Find some proportions.

Using either Table A or your calculator or software, find the proportion of observations from a standard Normal distribution that satisfies each of the following statements. In each case, sketch a standard Normal curve and shade the area under the curve that is the answer to the question.

- (a)  $Z > 1.55$
- (b)  $Z < 1.55$
- (c)  $Z > -0.70$
- (d)  $-0.70 < Z < 1.55$

### 1.129 Find more proportions.

Using either Table A or your calculator or software, find the proportion of observations from a standard Normal distribution for each of the following events. In each case, sketch a standard Normal curve and shade the area representing the proportion.

- (a)  $Z \leq -1.7$
- (b)  $Z \geq -1.7$
- (c)  $Z > 1.9$
- (d)  $-1.7 < Z < 1.9$

### 1.130 Find some values of $z$ .

Find the value  $z$  of a standard Normal variable  $Z$  that satisfies each of the following conditions. (If you use Table A, report the value of  $z$  that comes closest to satisfying the condition.) In each case, sketch a standard Normal curve with your value of  $z$  marked on the axis.

- (a) 28% of the observations fall below  $z$ .
- (b) 60% of the observations fall above  $z$ .

### 1.131 Find more values of $z$ .

The variable  $Z$  has a standard Normal distribution.

- (a) Find the number  $z$  that has cumulative proportion 0.78.
- (b) Find the number  $z$  such that the event  $Z > z$  has proportion 0.22.

### 1.132 Find some values of $z$ .

The Wechsler Adult Intelligence Scale (WAIS) is the most common IQ test. The scale of scores is set separately for each age group, and the scores are approximately Normal with mean 100 and standard deviation 15. People with WAIS scores below 70 are considered developmentally disabled when, for example, applying for Social Security disability benefits. What percent of adults are developmentally disabled by this criterion?

### 1.133 High IQ scores.

The Wechsler Adult Intelligence Scale (WAIS) is the most common IQ test. The scale of scores is set separately for each age group, and the scores are approximately Normal with mean 100 and standard deviation 15. The organization MENSA, which calls itself “the high-IQ society,” requires a WAIS score of 130 or higher for membership. What percent of adults would qualify for membership?

*There are two major tests of readiness for college, the ACT and the SAT. ACT scores are reported on a scale from 1 to 36. The distribution of ACT scores is approximately Normal with mean  $\mu = 21.5$  and standard deviation  $\sigma = 5.4$ . SAT scores are reported on a scale from 600 to 2400. The distribution of SAT scores is approximately Normal with mean  $\mu = 1498$  and standard deviation  $\sigma = 316$ . Exercises 1.134 to 1.143 are based on this information.*

### 1.134 Compare an SAT score with an ACT score.

Jessica scores 1825 on the SAT. Ashley scores 28 on the ACT. Assuming that both tests measure the same thing, who has the higher score? Report the  $z$ -scores for both students.

### 1.135 Make another comparison.

Joshua scores 17 on the ACT. Anthony scores 1030 on the SAT. Assuming that both tests measure the same thing, who has the higher score? Report the  $z$ -scores for both students.

### 1.136 Find the ACT equivalent.

Jorge scores 2060 on the SAT. Assuming that both tests measure the same thing, what score on the ACT is equivalent to Jorge’s SAT score?

### 1.137 Find the SAT equivalent.

Alyssa scores 32 on the ACT. Assuming that both tests measure the same thing, what score on the SAT is equivalent to Alyssa’s ACT score?

### **1.138 Find an SAT percentile.**

Reports on a student's ACT or SAT results usually give the percentile as well as the actual score. The percentile is just the cumulative proportion stated as a percent: the percent of all scores that were lower than or equal to this one. Renee scores 2040 on the SAT. What is her percentile?

### **1.139 Find an ACT percentile.**

Reports on a student's ACT or SAT results usually give the percentile as well as the actual score. The percentile is just the cumulative proportion stated as a percent: the percent of all scores that were lower than or equal to this one. Joshua scores 17 on the ACT. What is his percentile?

### **1.140 How high is the top 15%?**

What SAT scores make up the top 15% of all scores?

### **1.141 How low is the bottom 10%?**

What SAT scores make up the bottom 10% of all scores?

### **1.142 Find the ACT quintiles.**

The quintiles of any distribution are the values with cumulative proportions 0.20, 0.40, 0.60, and 0.80. What are the quintiles of the distribution of ACT scores?

### **1.143 Find the SAT quartiles.**

The quartiles of any distribution are the values with cumulative proportions 0.25 and 0.75. What are the quartiles of the distribution of SAT scores?

### **1.144 Do you have enough “good cholesterol?”**

High-density lipoprotein (HDL) is sometimes called the “good cholesterol” because low values are associated with a higher risk of heart disease. According to the American Heart Association, people over the age of 20 years should have at least 40 milligrams per deciliter (mg/dl) of HDL cholesterol.<sup>40</sup> U.S. women aged 20 and over have a mean HDL of 55 mg/dl with a standard deviation of 15.5 mg/dl. Assume that the distribution is Normal.

- (a) What percent of women have low values of HDL (40 mg/dl or less)?
- (b) HDL levels of 60 mg/dl and higher are believed to protect people from heart disease. What percent of women have protective levels of HDL?
- (c) Women with more than 40 mg/dl but less than 60 mg/dl of HDL are in the intermediate range, neither very good or very bad. What proportion are in this category?

### **1.145 Men and HDL cholesterol.**

HDL cholesterol levels for men have a mean of 46 mg/dl with a standard deviation of 13.6 mg/dl. Answer the questions given in the previous exercise for the population of men.

### 1.146 Diagnosing osteoporosis.

Osteoporosis is a condition in which the bones become brittle due to loss of minerals. To diagnose osteoporosis, an elaborate apparatus measures bone mineral density (BMD). BMD is usually reported in standardized form. The standardization is based on a population of healthy young adults. The World Health Organization (WHO) criterion for osteoporosis is a BMD 2.5 standard deviations below the mean for young adults. BMD measurements in a population of people similar in age and sex roughly follow a Normal distribution.

- (a) What percent of healthy young adults have osteoporosis by the WHO criterion?
- (b) Women aged 70 to 79 are of course not young adults. The mean BMD in this age is about  $-2$  on the standard scale for young adults. Suppose that the standard deviation is the same as for young adults. What percent of this older population has osteoporosis?

### 1.147 Deciles of Normal distributions.

The **deciles** of any distribution are the 10th, 20th, . . . , 90th percentiles. The first and last deciles are the 10th and 90th percentiles, respectively.

- (a) What are the first and last deciles of the standard Normal distribution?
- (b) The weights of 9-ounce potato chip bags are approximately Normal with mean 9.12 ounces and standard deviation 0.15 ounce. What are the first and last deciles of this distribution?

### 1.148 Quartiles for Normal distributions.

The quartiles of any distribution are the values with cumulative proportions 0.25 and 0.75.

- (a) What are the quartiles of the standard Normal distribution?
- (b) Using your numerical values from (a), write an equation that gives the quartiles of the  $N(\mu, \sigma)$  distribution in terms of  $\mu$  and  $\sigma$

### 1.149 *IQR* for Normal distributions.

Continue your work from the previous exercise. The interquartile range *IQR* is the distance between the first and third quartiles of a distribution.

- (a) What is the value of the *IQR* for the standard Normal distribution?
- (b) There is a constant  $c$  such that  $IQR = c\sigma$  for any Normal distribution  $N(\mu, \sigma)$ . What is the value of  $c$ ?

### 1.150 Outliers for Normal distributions.

Continue your work from the previous two exercises. The percent of the observations that are suspected outliers according to the  $1.5 \times IQR$  rule is the same for any Normal distribution. What is this percent?

### 1.151 Deciles of HDL cholesterol.

The **deciles** of any distribution are the 10th, 20th, . . . , 90th percentiles. Refer to Exercise 1.144 where we assumed that the distribution of HDL cholesterol in U.S. women aged 20 and over is Normal with mean 55 mg/dl and standard deviation 15.5 mg/dl. Find the deciles for this distribution.

*The remaining exercises for this section require the use of software that will make Normal quantile plots.*

### 1.152 Longleaf pine trees.

Exercise 1.72 (page 50) gives the diameter at breast height (DBH) for 40 longleaf pine trees from the Wade Tract in Thomas County, Georgia. Make a Normal quantile plot for these data and write a short paragraph interpreting what it describes.  PINES

### 1.153 Three varieties of flowers.

The study of tropical flowers and their hummingbird pollinators (Exercise 1.88, page 52) measured the lengths of three varieties of *Heliconia* flowers. We expect that such biological measurements will have roughly Normal distributions.  HELICON

- Make Normal quantile plots for each of the three flower varieties. Which distribution is closest to Normal?
- The other two distributions show the same kind of mild deviation from Normality. In what way are these distributions non-Normal?
- Compute the mean for each variety. For each flower, subtract the mean for its variety. Make a single data set for all varieties that contains the deviations from the means. Use this data set to create a Normal quantile plot. Examine the plot and summarize your conclusions.

### 1.154 Use software to generate some data.

Use software to generate 200 observations from the standard Normal distribution. Make a histogram of these observations. How does the shape of the histogram compare with a Normal density curve? Make a Normal quantile plot of the data. Does the plot suggest any important deviations from Normality? (Repeating this exercise several times is a good way to become familiar with how histograms and Normal quantile plots look when data actually are close to Normal.)

### 1.155 Use software to generate more data.

Use software to generate 200 observations from the uniform distribution described in Exercise 1.122. Make a histogram of these observations. How does the histogram compare with the density curve in Figure 1.32? Make a Normal quantile plot of your data. According to this plot, how does the uniform distribution deviate from Normality?

## CHAPTER 1 Exercises

### 1.156 Comparing fuel efficiency.

Let's compare the fuel efficiencies (mpg) of small cars and sporty cars for model year 2013.<sup>41</sup> Here are the data:

Small Cars										
50	45	37	37	37	36	35	34	34	34	
34	34	34	34	33	33	33	33	33		
Sporty Cars										
33	32	32	32	32	31	31	31	31	31	
31	30	30	30	30	30	30	30	29	29	
29	29	29	29	29						

Give graphical and numerical descriptions of the fuel efficiencies for these two types of vehicles. What are the main features of the distributions? Compare the two distributions and summarize your results in a short paragraph.  **MPGSS**

### 1.157 Smoking.

The Behavioral Risk Factor Surveillance System (BRFSS) conducts a large survey of health conditions and risk behaviors in the United States.<sup>42</sup> The BRFSS data file contains data on 23 demographic factors and risk factors for each state. Use the percent of smokers (SmokeEveryDay) for this exercise.  **BRFSS**

- Prepare a graphical display of the distribution and use your display to describe the major features of the distribution.
- Calculate numerical summaries. Give reasons for your choices.
- Write a short paragraph summarizing what the data tell us about smoking in the United States.

### 1.158 Eat your fruits and vegetables.

Nutrition experts recommend that we eat five servings of fruits and vegetables each day. The BRFSS data file described in the previous exercise includes a variable that gives the percent of people who regularly eat five or more servings of fruits and vegetables (FruitVeg5). Answer the questions given in the previous exercise for this variable.  **BRFSS**

### 1.159 Vehicle colors.

Vehicle colors differ among types of vehicle in different regions. Here are data on the most popular

colors in 2011 for several different regions of the world:<sup>43</sup>

Color	North America percent	South America percent	Europe percent	China percent	South Korea percent	Japan percent
Silver	16	30	15	26	30	19
White	23	17	20	15	25	26
Gray	13	15	18	10	12	9
Black	18	19	25	21	15	20
Blue	9	1	7	9	4	9
Red	10	11	6	7	4	5
Brown	5	5	5	4	4	4
Yellow	3	1	1	2	1	1
Green	2	1	1	1	1	1
Other	1	0	2	5	4	6

Use the methods you learned in this chapter to compare the vehicle color preferences for the regions of the world presented in this table. Write a report summarizing your findings with an emphasis on similarities and differences across regions. Include recommendations related to marketing and advertising of vehicles in these regions.  **VCOLORS**

### 1.160 Canadian international trade.

The government organization Statistics Canada provides data on many topics related to Canada's population, resources, economy, society, and culture. Go to the web page [statcan.gc.ca/start-debut-eng.html](http://statcan.gc.ca/start-debut-eng.html). Under the "Subject" tab, choose "International trade." Pick some data from the resources listed and use the methods that you learned in this chapter to create graphical and numerical summaries. Write a report summarizing your findings that includes supporting evidence from your analyses.

### 1.161 Travel and tourism in Canada.

Refer to the previous exercise. Under the "Subject" tab, choose "Travel and tourism." Pick some data from the resources listed and use the methods that you learned in this chapter to create graphical and numerical summaries. Write a report summarizing your findings that includes supporting evidence from your analyses.

### 1.162 Internet use.

The World Bank collects data on many variables related to development for countries throughout the world.<sup>44</sup> One of these is Internet use, expressed as the number of users per 100 people. The data file for this exercise gives 2011 values of this variable for 185 countries. Use graphical and numerical methods to describe this distribution. Write a short report summarizing what the data tell about worldwide Internet use.  **INETUSE**

### 1.163 Change Internet use.

Refer to the previous exercise. The data file also contains the numbers of users per 100 people for 2010.  **INETUSE**

- (a) Analyze the 2010 data.
- (b) Compute the change in the number of users per 100 people from 2010 to 2011. Analyze the changes.
- (c) Compute the percent change in the number of users per 100 people from 2010 to 2011. Analyze the percent changes.
- (d) Write a summary of your analyses in parts (a) to (c). Include a comparison of the changes versus the percent changes.

### 1.164 Leisure time for college students.

You want to measure the amount of “leisure time” that college students enjoy. Write a brief discussion of two issues:

- (a) How will you define “leisure time”?
- (b) Once you have defined leisure time, how will you measure Sally’s leisure time this week?

### 1.165 Internet service.

Providing Internet service is a very competitive business in the United States. The numbers of subscribers claimed by the top 10 providers of service were as follows.<sup>45</sup>

Service provider	Subscribers (millions)	Service provider	Subscribers (millions)
Comcast	17.0	Charter	5.5
Time Warner	9.7	Verizon	4.3
AT&T	17.8	CenturyLink	6.4
Cox	3.9	SuddenLink	1.4
Optimum	3.3	EarthLink	1.6

Display these data in a graph. Write a short summary describing the distribution of subscribers for these 10 providers. Business people looking at this graph see an industry that offers opportunities for larger companies to take over. 

### 1.166 Internet service provider ratings.

Refer to the previous exercise. The following table gives overall ratings, on a 10-point scale, for these providers. These were posted on the TopTenREVIEWS website.<sup>46</sup> 

Service provider	Rating	Service provider	Rating
Comcast	9.25	Charter	7.88
Time Warner	8.60	Verizon	7.63
AT&T	8.53	CenturyLink	7.58
Cox	8.38	SuddenLink	7.38
Optimum	8.20	EarthLink	7.20

Display these data in a graph. Write a short summary describing the distribution of ratings for these 10 providers. 

### 1.167 What graph would you use?

What type of graph or graphs would you plan to make in a study of each of the following issues?

- (a) What makes of cars do students drive? How old are their cars?
- (b) How many hours per week do students study? How does the number of study hours change during a semester?
- (c) Which radio stations are most popular with students?
- (d) When many students measure the concentration of the same solution for a chemistry course laboratory assignment, do their measurements follow a Normal distribution?

### 1.168 Spam filters.

A university department installed a spam filter on its computer system. During a 21-day period, 6693 messages were tagged as spam. How much spam you get depends on what your online habits are. Here are the counts for some students and faculty in this department (with log-in IDs changed, of course):

ID	Count	ID	Count	ID	Count	ID	Count
AA	1818	BB	1358	CC	442	DD	416
EE	399	FF	389	GG	304	HH	251
II	251	JJ	178	KK	158	LL	103

All other department members received fewer than 100 spam messages. How many did the others receive in total? Make a graph and comment on what you learn from these data.  **SPAM**

### 1.169 How much vitamin C do you need?

The Food and Nutrition Board of the Institute of Medicine working in cooperation with scientists from Canada have used scientific data to answer this question for a variety of vitamins and minerals.<sup>47</sup> Their methodology assumes that needs, or requirements, follow a distribution. They have produced guidelines called dietary reference intakes for different gender-by-age combinations. For vitamin C, there are three dietary reference intakes: the estimated average requirement (EAR), which is the mean of the requirement distribution; the recommended dietary allowance (RDA), which is the intake that would be sufficient for 97% to 98% of the population; and the tolerable upper level (UL), the intake that is unlikely to pose health risks. For women aged 19 to 30 years, the EAR is 60 milligrams per day (mg/d), the RDA is 75 mg/d, and the UL is 2000 mg/d.<sup>48</sup>

- (a) The researchers assumed that the distribution of requirements for vitamin C is Normal. The EAR gives the mean. From the definition of the RDA, let's assume that its value is the 97.72 percentile. Use this information to determine the standard deviation of the requirement distribution.
- (b) Sketch the distribution of vitamin C requirements for 19- to 30-year-old women. Mark the EAR, the RDA, and the UL on your plot.

### 1.170 How much vitamin C do men need?

Refer to the previous exercise. For men aged 19 to 30 years, the EAR is 75 milligrams per day (mg/d), the RDA is 90 mg/d, and the UL is 2000 mg/d. Answer the questions in the previous

exercise for this population.

### 1.171 How much vitamin C do women consume?

To evaluate whether or not the intake of a vitamin or mineral is adequate, comparisons are made between the intake distribution and the requirement distribution. Here is some information about the distribution of vitamin C intake, in milligrams per day, for women aged 19 to 30 years:<sup>49</sup>

Percentile (mg/d)									
Mean	1st	5th	19th	25th	50th	75th	90th	95th	99th
84.1	31	42	48	61	79	102	126	142	179

- (a) Use the 5th, the 50th, and the 95th percentiles of this distribution to estimate the mean and standard deviation of this distribution assuming that the distribution is Normal. Explain your method for doing this.
- (b) Sketch your Normal intake distribution on the same graph with a sketch of the requirement distribution that you produced in part (b) of Exercise 1.69.
- (c) Do you think that many women aged 19 to 30 years are getting the amount of vitamin C that they need? Explain your answer.

### 1.172 How much vitamin C do men consume?

To evaluate whether or not the intake of a vitamin or mineral is adequate, comparisons are made between the intake distribution and the requirement distribution. Here is some information about the distribution of vitamin C intake, in milligrams per day, for men aged 19 to 30 years:

Percentile (mg/d)									
Mean	1st	5th	19th	25th	50th	75th	90th	95th	99th
122.2	39	55	65	85	114	150	190	217	278

- (a) Use the 5th, the 50th, and the 95th percentiles of this distribution to estimate the mean and standard deviation of this distribution assuming that the distribution is Normal. Explain your method for doing this.
- (b) Sketch your Normal intake distribution on the same graph with a sketch of the requirement distribution that you produced in Exercise 1.70.
- (c) Do you think that many men aged 19 to 30 years are getting the amount of vitamin C that they need? Explain your answer.

### 1.173 Time spent studying.

Do women study more than men? We asked the students in a large first-year college class how many minutes they studied on a typical weeknight. Here are the responses of random samples of 30 women and 30 men from the class:  STUDY

Women					Men				
170	120	180	360	240	80	120	30	90	200
120	180	120	240	170	90	45	30	120	75

150	120	180	180	150	150	120	60	240	300
200	150	180	150	180	240	60	120	60	30
120	60	120	180	180	30	230	120	95	150
90	240	180	115	120	0	200	120	120	180

- (a) Examine the data. Why are you not surprised that most responses are multiples of 10 minutes? We eliminated one student who claimed to study 30,000 minutes per night. Are there any other responses that you consider suspicious?
- (b) Make a back-to-back stemplot of these data. Report the approximate midpoints of both groups. Does it appear that women study more than men (or at least claim that they do)?
- (c) Make side-by-side boxplots of these data. Compare the boxplots with the stemplot you made in part (b). Which do you prefer? Give reasons for your answer.

### 1.174 Product preference.

Product preference depends in part on the age, income, and gender of the consumer. A market researcher selects a large sample of potential car buyers. For each consumer, she records gender, age, household income, and automobile preference. Which of these variables are categorical and which are quantitative?

### 1.175 Two distributions.

If two distributions have exactly the same mean and standard deviation, must their histograms have the same shape? If they have the same five-number summary, must their histograms have the same shape? Explain.

### 1.176 Norms for reading scores.

Raw scores on behavioral tests are often transformed for easier comparison. A test of reading ability has mean 70 and standard deviation 10 when given to third-graders. Sixth-graders have mean score 80 and standard deviation 11 on the same test. To provide separate “norms” for each grade, we want scores in each grade to have mean 100 and standard deviation 20.

- (a) What linear transformation will change third-grade scores  $x$  into new scores  $x_{\text{new}} = a + bx$  that have the desired mean and standard deviation? (Use  $b > 0$  to preserve the order of the scores.)
- (b) Do the same for the sixth-grade scores.
- (c) David is a third-grade student who scores 72 on the test. Find David’s transformed score. Nancy is a sixth-grade student who scores 78. What is her transformed score? Who scores higher within his or her grade?
- (d) Suppose that the distribution of scores in each grade is Normal. Then both sets of transformed scores have the  $N(100, 20)$  distribution. What percent of third-graders have scores less than 75? What percent of sixth-graders have scores less than 75?

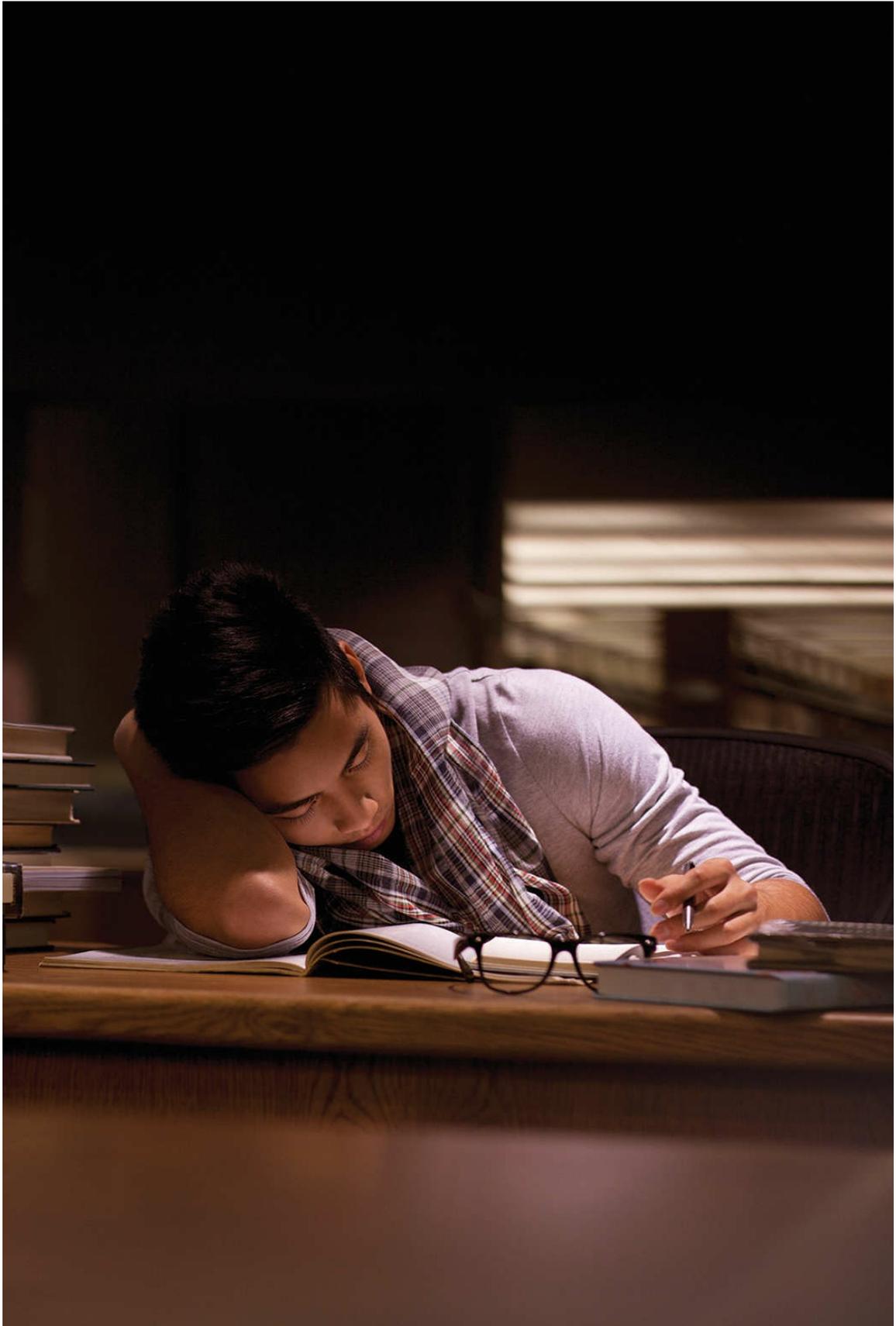
### 1.177 Use software to generate some data.

Most statistical software packages have routines for generating values of variables having specified distributions. Use your statistical software to generate 30 observations from the  $N(25, 8)$  distribution.

Compute the mean and standard deviation  $\bar{x}$  and  $s$  of the 30 values you obtain. How close are  $\bar{x}$  and  $s$  to the  $\mu$  and  $\sigma$  of the distribution from which the observations were drawn? Repeat 19 more times the process of generating 30 observations from the  $N(25, 8)$  distribution and recording  $\bar{x}$  and  $s$ . Make a stemplot of the 20 values of  $\bar{x}$  and another stemplot of the 20 values of  $s$ . Make Normal quantile plots of both sets of data. Briefly describe each of these distributions. Are they symmetric or skewed? Are they roughly Normal? Where are their centers? (The distributions of measures like  $\bar{x}$  and  $s$  when repeated sets of observations are made from the same theoretical distribution will be very important in later chapters.)

# 2 Looking at Data— Relationships

CHAPTER



**2.1 Relationships**

**2.2 Scatterplots**

- 2.3 Correlation**
- 2.4 Least-Squares Regression**
- 2.5 Cautions about Correlation and Regression**
- 2.6 Data Analysis for Two-Way Tables**
- 2.7 The Question of Causation**

## Introduction

In Chapter 1 we learned to use graphical and numerical methods to describe the distribution of a single variable. Many of the interesting examples of the use of statistics involve relationships between pairs of variables. Learning ways to describe relationships with graphical and numerical methods is the focus of this chapter.

In Section 2.2 we focus on graphical descriptions. The scatterplot is our fundamental graphical tool for displaying the relationship between two quantitative variables. Sections 2.3 and 2.4 move on to numerical summaries for these relationships. Cautions about the use of these methods are discussed in Section 2.5. Graphical and numerical methods for describing the relationship between two categorical variables are presented in Section 2.6. We conclude with Section 2.7, a brief overview of issues related to the distinction between associations and causation.

## 2.1 Relationships

When you complete this section, you will be able to

- Identify the key characteristics of a data set to be used to explore a relationship between two variables.
- Categorize variables as response variables or explanatory variables.

In Chapter 1 (page 4) we discussed the key characteristics of a data set. Cases are the objects described by a set of data, and a variable is a characteristic of a case. We also learned to categorize variables as categorical or quantitative. For Chapter 2, we focus on data sets that have two variables.

### Example

#### 2.1 Stress and lack of sleep.

Stress is a common problem for college students. Exploring factors that are associated with stress may lead to strategies that will help students to relieve some of the stress that they experience. Recent studies have suggested that a lack of sleep is associated with stress.<sup>1</sup> The two variables involved in the relationship here are lack of sleep and stress. The cases are the students who are the subjects for a particular study.

When we study relationships between two variables, it is not sufficient to collect data on the two variables. *A key idea for this chapter is that both variables must be measured on the same cases.*



### USE YOUR KNOWLEDGE

## 2.1 Relationship between attendance at class and final exam.

You want to study the relationship between the attendance at class and the score on the final for the 30 students enrolled in an elementary statistics class.

- (a) Who are the cases for your study?
- (b) What are the variables?
- (c) Are the variables quantitative or categorical? Explain your answer.

We use the term *associated* to describe the relationship between two variables, such as stress and lack of sleep in Example 2.1. Here is another example where two variables are associated.

### Example

## 2.2 Size and price of a coffee beverage.

You visit a local Starbucks to buy a Mocha Frappuccino<sup>®</sup>. The barista explains that this blended coffee beverage comes in three sizes and asks if you want a Tall, a Grande, or a Venti. The prices are \$3.75, \$4.35, and \$4.85, respectively. There is a clear association between the size of the Mocha Frappuccino and its price.

## ASSOCIATION BETWEEN VARIABLES

Two variables measured on the same cases are **associated** if knowing the values of one of the variables tells you something about the values of the other variable that you would not know without this information.



In the Mocha Frappuccino example, knowing the size tells you the exact price, so the association here is very strong. Many statistical associations, however, are simply overall tendencies that allow exceptions. Some people get adequate sleep and are highly stressed. Others get little sleep and do not experience much stress. The association here is much weaker than the one in the Mocha Frappuccino example.

### **Examining relationships**

To examine the relationship between two or more variables, we first need to know some basic characteristics of the data. Here is an example.

## Example

### 2.3 Stress and lack of sleep.

A study of stress and lack of sleep collected data on 1125 students from an urban midwestern university. Two of the variables measured were the Pittsburgh Sleep Quality Index (PSQI) and the Subjective Units of Distress Scale (SUDS). In this study the cases are the 1125 students studied.<sup>2</sup> The PSQI is based on responses to a large number of questions that are summarized in a single variable that has a value between 0 and 21 for each subject. Therefore, we will treat the PSQI as a quantitative variable. The SUDS is a similar scale with values between 0 and 100 for each subject. We will treat the SUDS as a quantitative variable also.

In many situations, we measure a collection of categorical variables and then combine them in a scale that can be viewed as a quantitative variable. The PSQI is an example. We can also turn the tables in the other direction. Here is an example.

## Example

### 2.4 Hemoglobin and anemia.

Hemoglobin is a measure of iron in the blood. The units are grams of hemoglobin per deciliter of blood (g/dl). Typical values depend on age and gender. Adult women typically have values between 12 and 16 g/dl.

Anemia is a major problem in developing countries, and many studies have been designed to address the problem. In these studies, computing the mean hemoglobin is not particularly useful. For studies like these, it is more appropriate to use a definition of severe anemia (a hemoglobin level of less than 8 g/dl). Thus, for example, researchers can compare the proportions of subjects who are severely anemic for two treatments rather than the difference in the mean hemoglobin levels. In this situation, the categorical variable, severely anemic or not, is much more useful than the quantitative variable, hemoglobin.



When analyzing data to draw conclusions it is important to carefully consider the best way to summarize the data. *Just because a variable is measured as a quantitative variable, it does not necessarily follow that the best summary is based on the mean (or the median).* As the previous example illustrates, converting a quantitative variable to a categorical variable is a very useful option to keep in mind.

## USE YOUR KNOWLEDGE

### 2.2 Create a categorical variable from a quantitative variable.

Consider the study described in Example 2.3. Some analyses compared three groups of students. The students were classified as having optimal sleep quality (a PSQI of 5 or less), borderline sleep quality (a PSQI of 6 or 7), or poor sleep quality (a PSQI of 8 or more). When the three groups of students are compared, is the PSQI being used as a quantitative variable or as a categorical variable? Explain your answer and describe some advantages to using the optimal, borderline, and poor categories in explaining the results of a study such as this.

### 2.3 Replace names by ounces.

In the Mocha Frappuccino example, the variable size is categorical, with Tall, Grande, and Venti as the possible values. Suppose that you converted these values to the number of ounces: Tall is 12 ounces, Grande is 16 ounces, and Venti is 24 ounces. For studying the relationship between ounces and price, describe the cases and the variables, and state whether each is quantitative or categorical.

When you examine the relationship between two variables, a new question becomes important:

- Is your purpose simply to explore the nature of the relationship, or do you hope to show that one of the variables can explain variation in the other? Is one of the variables a *response variable* and the other an *explanatory variable*?

## RESPONSE VARIABLE, EXPLANATORY VARIABLE

A **response variable** measures an outcome of a study. An **explanatory variable** explains or causes changes in the response variable.

### Example

#### 2.5 Stress and lack of sleep.

Refer to the study of stress and lack of sleep in Example 2.3. Here, the explanatory variable is the Pittsburgh Sleep Quality Index, and the response variable is the Subjective Units of Distress Scale.

## USE YOUR KNOWLEDGE

#### 2.4 Sleep and stress or stress and sleep?

Consider the scenario described in the previous example. Make an argument for treating the Subjective Units of Distress Scale as the explanatory variable and the Pittsburgh Sleep Quality Index as the response variable.

In some studies it is easy to identify explanatory and response variables. The following example illustrates one situation where this is true: when we actually set values of one variable to see how it affects another variable.

### Example

#### 2.6 How much calcium do you need?

Adolescence is a time when bones are growing very actively. If young people do not have enough calcium, their bones will not grow properly. How much calcium is enough? Research designed to answer this question has been performed for many years at events called “Camp Calcium.”<sup>3</sup> At these camps, subjects eat controlled diets that are identical except for the amount of calcium. The amount of calcium retained by the body is the major response variable of interest. Since the amount of calcium consumed is controlled by the researchers, this variable is the explanatory variable.

When you don’t set the values of either variable but just observe both variables, there may or may not be explanatory and response variables. Whether there are depends on how you plan to use the data.

## Example

### 2.7 Student loans.

A college student aid officer looks at the findings of the National Student Loan Survey. She notes data on the amount of debt of recent graduates, their current income, and how stressful they feel about college debt. She isn’t interested in predictions but is simply trying to understand the situation of recent college graduates.

A sociologist looks at the same data with an eye to using amount of debt and income, along with other variables, to explain the stress caused by college debt. Now, amount of debt and income are explanatory variables, and stress level is the response variable.

In many studies, the goal is to show that changes in one or more explanatory variables actually *cause* changes in a response variable. But many explanatory-response relationships do not involve direct causation. The SAT scores of high school students help predict the students’ future college grades, but high SAT scores certainly don’t cause high college grades.

## KEY CHARACTERISTICS OF DATA FOR RELATIONSHIPS

A description of the key characteristics of a data set that will be used to explore a relationship between two variables should include

- **Cases.** Identify the cases and how many there are in the data set.

- **Label.** Identify what is used as a label variable if one is present.
- **Categorical or quantitative.** Classify each variable as categorical or quantitative.
- **Values.** Identify the possible values for each variable.
- **Explanatory or response.** If appropriate, classify each variable as explanatory or response.

Some of the statistical techniques in this chapter require us to distinguish explanatory from response variables; others make no use of this distinction. You will often see explanatory variables called *independent variables* and response variables called *dependent variables*. These terms express mathematical ideas; they are not statistical terms. The concept that underlies this language is that response variables *depend* on explanatory variables. Because the words “independent” and “dependent” have other meanings in statistics that are unrelated to the explanatory-response distinction, we prefer to avoid those words.

independent variable

dependent variable

Most statistical studies examine data on more than one variable. Fortunately, statistical analysis of several-variable data builds on the tools used for examining individual variables. The principles that guide our work also remain the same:

- Start with a graphical display of the data.
- Look for overall patterns and deviations from those patterns.
- Based on what you see, use numerical summaries to describe specific aspects of the data.

## SECTION 2.1 Summary

To study relationships between variables, we must measure the variables on the same cases.

If we think that a variable  $x$  may explain or even cause changes in another variable  $y$ , we call  $x$  an **explanatory variable** and  $y$  a **response variable**.

## SECTION 2.1 Exercises

*For Exercise 2.1, see page 82; for Exercises 2.2 and 2.3, see page 84; and for Exercise 2.4, see page 84.*

## 2.5 High click counts on Twitter.

A study was done to identify variables that might produce high click counts on Twitter. You and 9 of your

friends collect data on all of your tweets for a week. You record the number of click counts, the time of day, the day of the week, the gender of the person posting the tweet, and the length of the tweet.

- (a) What are the cases for this study?
- (b) Classify each of the variables as categorical or quantitative.
- (c) Classify each of the variables as explanatory, response, or neither. Explain your answers.

## **2.6 Explanatory or response?**

For each of the following scenarios, classify each of the pair of variables as explanatory or response or neither. Give reasons for your answers.

- (a) The amount of calcium per day in your diet and the amount of vitamin A per day in your diet.
- (b) The number of bedrooms in an apartment and the monthly rent of the apartment.
- (c) The diameter of an apple and the weight of the apple.
- (d) The length of time that you spend in the sun and the amount of vitamin D that is produced by your skin.

## **2.7 Buy and sell prices of used textbooks.**

Think about a study designed to compare the prices of textbooks for third- and fourth-year college courses in five different majors. For the five majors, you want to examine the relationship between the difference in the price that you pay for a used textbook and the price that the seller gives back to you when you return the textbook. Describe a data set that could be used for this study, and give the key characteristics of the data.

## **2.8 Protein and carbohydrates.**

Think about a study designed to examine the relationship between protein intake and carbohydrate intake in the diets of college sophomores. Describe a data set that could be used for this study, and give the key characteristics of the data.

## **2.9 Can you examine the relationship?**

For each of the following scenarios, determine whether or not the data would allow you to examine a relationship between two variables. If your answer is Yes, give the key characteristics of a data set that could be analyzed. If your answer is No, explain your answer.

- (a) The temperature where you live yesterday and the temperature where you live today.
- (b) The average high school grade point averages of the first-year students at your college and the college grade point averages of the students who will graduate this year.
- (c) A consumer study reported the price per load and an overall quality score for 24 brands of laundry detergents.

## 2.2 Scatterplots

When you complete this section, you will be able to

- Make a scatterplot to examine a relationship between two quantitative variables.
- Describe the overall pattern in a scatterplot and any striking deviations from that pattern.
- Use a scatterplot to describe the form, direction, and strength of a relationship.
- Use a scatterplot to identify outliers.
- Identify a linear pattern in a scatterplot.
- Explain the effect of a change of units on a scatterplot.
- Use a log transformation to change a curved relationship into a linear relationship.
- Use different plotting symbols to include information about a categorical variable in a scatterplot.

### Example

#### 2.8 Laundry detergents.

Consumers Union provides ratings on a large variety of consumer products. They use sophisticated testing methods as well as surveys of their members to create these ratings. The ratings are published in their magazine, *Consumer Reports*.<sup>4</sup>



LAUNDRY

One recent article rated laundry detergents on a scale from 1 to 100. Here are the ratings along with the price per load, in cents, for 24 laundry

detergents:<sup>5</sup>

Rating	Price (cents)						
61	17	59	22	56	22	55	16
55	30	52	23	51	11	50	15
50	9	48	16	48	15	48	18
46	13	46	13	45	17	36	8
35	8	34	12	33	7	32	6
32	5	29	14	26	11	26	13



We will examine the relationship between rating and price per load for these laundry detergents. We expect that the higher-priced detergents will tend to have higher ratings.

## USE YOUR KNOWLEDGE

### 2.10 Examine the spreadsheet.

Examine the spreadsheet that gives the laundry detergent data in the data file LAUNDRY.



LAUNDRY

(a) How many cases are in the data set?

(b) Describe the labels, variables, and values.

(c) Which columns represent quantitative variables? Which columns represent categorical variables?

(d) Is there an explanatory variable? A response variable? Explain your answer.

## 2.11 Use the data set.



### LAUNDRY

Using the data set from the previous exercise, create graphical and numerical summaries for the rating and for the price per load.

The most common way to display the relationship between two quantitative variables is a *scatterplot*.

## SCATTERPLOT

A **scatterplot** shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual.

## Example

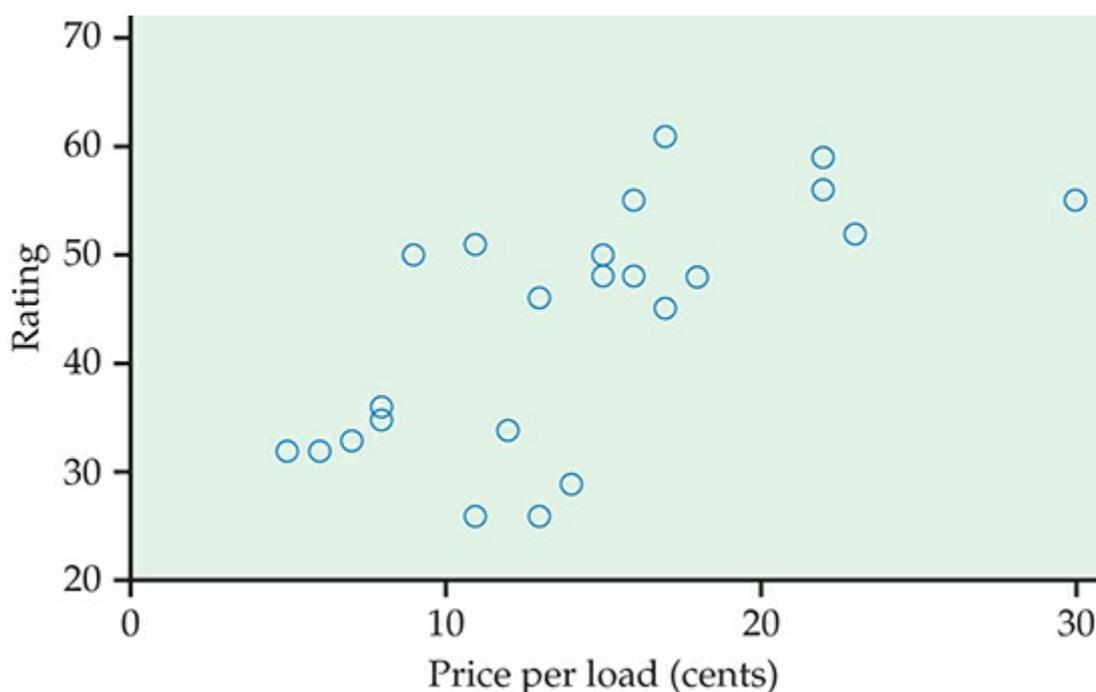
### 2.9 Laundry detergents.

A higher price for a product should be associated with a better product. Therefore, let's treat price per load as the explanatory variable and rating as the response variable in our examination of the relationship between these two variables. We begin with a graphical display.



## LAUNDRY

Figure 2.1 gives a scatterplot that displays the relationship between the response variable, rating, and the explanatory variable, price per load. The plot confirms our idea that a higher price should be associated with a better rating.



**FIGURE 2.1**

Scatterplot of price per load (in cents) versus rating for 24 laundry detergents, for Example 2.9.

Always plot the explanatory variable, if there is one, on the horizontal axis (the  $x$  axis) of a scatterplot. We usually call the explanatory variable  $x$  and the response variable  $y$ . If there is no explanatory-response distinction, either variable can go on the horizontal axis. Time plots, such as the one in Figure 1.13 (page 24), are special scatterplots where the explanatory variable  $x$  is a measure of time.

## USE YOUR KNOWLEDGE

### 2.12 Make a scatterplot.



## LAUNDRY

- (a) Make a scatterplot similar to Figure 2.1 for the laundry detergent data.
- (b) Two of the laundry detergents are gels. These products are made by the same manufacturer, and one of them has an additive for stain removal. The ratings and prices per load are the same; the rating is 46 and the price is 13. Mark the location of these gels on your plot.
- (c) Cases with identical values for both variables are generally indistinguishable in a scatterplot. To what extent do you think that this could give a distorted picture of the relationship between two variables for a data set that has a large number of duplicate values? Explain your answer.

### 2.13 Change the units.



## LAUNDRY

- (a) Create a spreadsheet for the laundry detergent data with the price per load expressed in dollars.
- (b) Make a scatterplot for the data in your spreadsheet.
- (c) Describe how this scatterplot differs from Figure 2.1.

## Interpreting scatterplots

To look more closely at a scatterplot such as Figure 2.1, apply the strategies of exploratory analysis learned in Chapter 1.

### EXAMINING A SCATTERPLOT

In any graph of data, look for the **overall pattern** and for striking **deviations** from that pattern.

You can describe the overall pattern of a scatterplot by the **form**, **direction**, and **strength** of the relationship.

An important kind of deviation is an **outlier**, an individual value that falls outside the overall pattern of the relationship.

Figure 2.1 shows a clear *form*: the data lie in a roughly straight-line, or *linear*, pattern. To help us see this relationship, we can use software to put a straight line through the data. We will see more details about how this is done in Section 2.4.

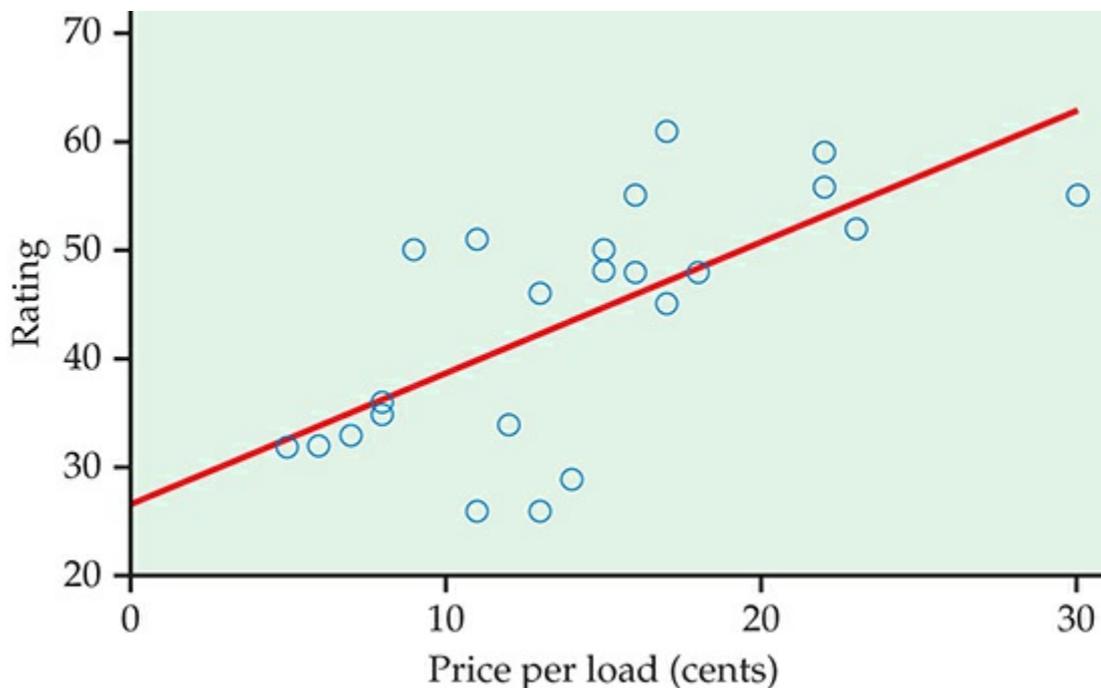
## Example

### 2.10 Scatterplot with a straight line.

Figure 2.2 plots the laundry detergent data with a fitted straight line. The line helps us to see and to evaluate the linear form of the relationship.



There is a large amount of scatter about the line. Referring to the data given in Example 2.8, we see that for 11 cents per load, one detergent has a rating of 26, while another has a rating of 51, almost twice as large. No clear outliers are evident.



**FIGURE 2.2**

Scatterplot of rating versus price per load (in cents) with a fitted straight line, for Example 2.10.

The relationship in Figure 2.2 has a clear *direction*: laundry detergents that cost more generally have higher ratings. This is a *positive association* between the two

variables.

## POSITIVE ASSOCIATION, NEGATIVE ASSOCIATION

Two variables are **positively associated** when above-average values of one tend to accompany above-average values of the other and below-average values also tend to occur together.

Two variables are **negatively associated** when above-average values of one tend to accompany below-average values of the other, and vice versa.

The *strength* of a relationship in a scatterplot is determined by how closely the points follow a clear form. The overall relationship in Figure 2.2 is fairly moderate. Here is an example of a stronger linear relationship.

### Example

#### 2.11 Debt for 33 countries.

The amount of debt owed by a country is a measure of its economic health. The Organisation for Economic Co-operation and Development collects data on the central government debt for many countries. One of their tables gives the debt for countries over several years.<sup>6</sup>

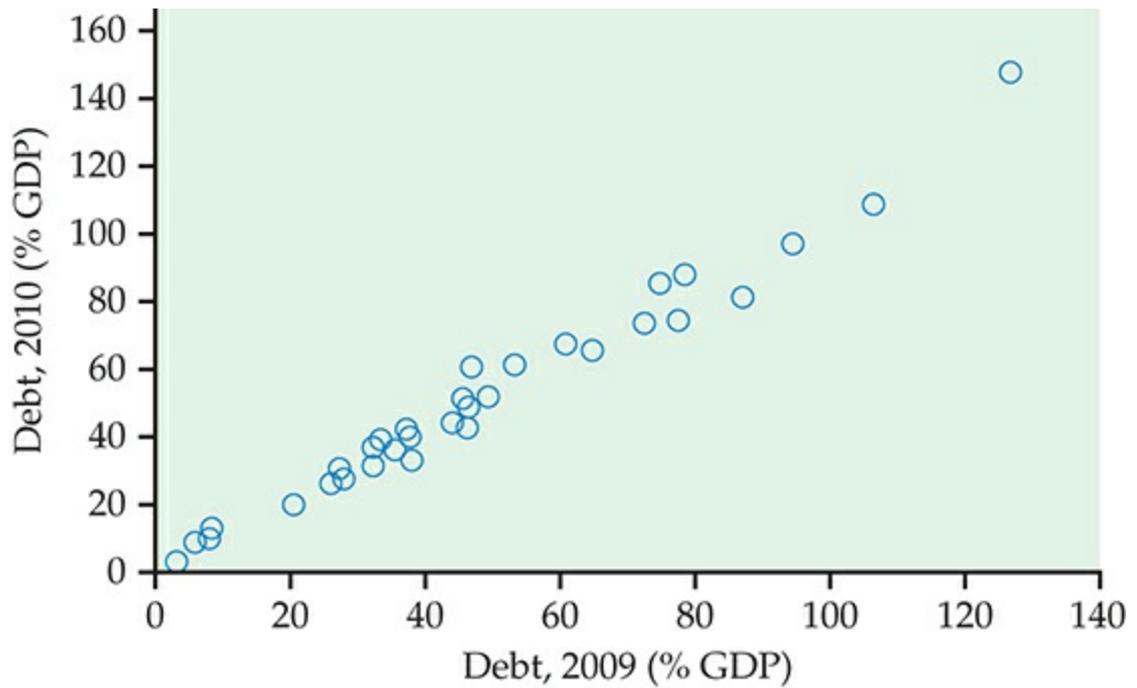


Figure 2.3 is a spreadsheet giving the government debt for 33 countries that have data for the years 2005 to 2010. Since countries that have large economies tend to have large debts, we have chosen a table that expresses the debt as a percent of the gross domestic product (GDP).

	A	B	C	D	E	F	G
1	Country	Debt2005	Debt2006	Debt2007	Debt2008	Debt2009	Debt2010
2	Australia	6.312	5.76	5.181	4.922	8.195	10.966
3	Austria	62.116	60.434	57.829	59.319	64.916	65.754
4	Belgium	91.774	87.568	85.295	90.094	94.893	96.789
5	Canada	30.235	27.934	25.183	28.642	35.716	36.073
6	Chile	7.282	5.264	4.097	5.173	6.228	9.185
7	Czech Republic	23.164	24.904	25.24	27.102	32.496	36.625
8	Denmark	39.292	32.715	27.765	32.318	37.891	39.59
9	Estonia	2.091	1.836	1.319	1.761	3.55	3.227
10	Finland	38.17	35.561	31.201	29.452	37.549	41.683
11	France	53.275	52.131	52.118	53.406	61.231	67.418
12	Germany	40.832	41.232	39.55	39.55	44.205	44.403
13	Greece	110.572	107.675	105.674	110.617	127.022	147.839
14	Hungary	58.103	61.971	61.551	67.668	72.79	73.898
15	Iceland	19.378	24.807	23.237	44.175	87.473	81.257
16	Ireland	23.524	20.253	19.834	28.001	47.074	60.703
17	Israel	92.102	82.659	75.948	75.307	77.693	74.714
18	Italy	97.656	97.454	95.627	98.093	106.778	109.015
19	Korea	27.595	30.065	29.651	29.027	32.558	31.935
20	Luxembourg	0.821	1.458	1.419	8.153	8.489	12.578
21	Mexico	20.295	20.583	20.861	24.369	28.086	27.46
22	Netherlands	42.952	39.169	37.552	50.068	49.719	51.845
23	New Zealand	22.069	21.58	20.343	20.721	27.53	30.45
24	Norway	17.173	12.473	11.681	13.905	26.363	26.077
25	Poland	44.764	45.143	42.62	44.686	47.015	49.679
26	Portugal	66.194	67.732	66.622	68.88	78.73	87.962
27	Slovak Republic	33.103	29.164	28.108	26.342	33.749	39.078
28	Slovenia	26.9	25.782	23.207	21.188	33.628	36.023
29	Spain	36.36	32.965	30.019	33.695	46.026	51.693
30	Sweden	46.232	42.242	36.406	35.56	38.098	33.782
31	Switzerland	28.102	25.195	23.216	22.376	20.723	20.24
32	Turkey	51.087	45.498	39.551	40.011	46.35	42.851
33	United Kingdom	43.523	43.185	42.744	61.059	75.27	85.535
34	United States	36.149	36.039	35.703	40.183	53.573	61.274

**FIGURE 2.3**

Central government debt in the years 2005 to 2010 for 33 countries, in percent of GDP, for Example 2.11.



**FIGURE 2.4**

Scatterplot of debt in 2010 (percent of GDP) versus debt in 2009 (percent of GDP) for 33 countries, for Example 2.11.

Figure 2.4 is a scatterplot of the central government debt in 2010 versus the central government debt in 2009. The scatterplot shows a strong positive relationship between the debt in these two years.

## USE YOUR KNOWLEDGE

### 2.14 Make a scatterplot.

In our Mocha Frappuccino example, the 12-ounce drink costs \$3.75, the 16-ounce drink costs \$4.35, and the 24-ounce drink costs \$4.85. Explain which variable should be used as the explanatory variable, and make a scatterplot. Describe the scatterplot and the association between these two variables.

Can we conclude that the strong linear relationship that we found between the central government debt in 2009 and 2010 is evidence that the debt for each country is approximately the same in the two years? The answer is No. The first exercise below asks you to explore this issue.

## USE YOUR KNOWLEDGE

### 2.15 Are the debts in 2009 and 2010 approximately the same?

Use the methods you learned in Chapter 1 to examine whether or not the central government debts in 2009 and 2010 are approximately the same. (*Hint:* Think about creating a new variable that would help you to answer this question.)



### 2.16 The relationship between debt in 2005 and debt in 2010.

Make a plot similar to Figure 2.4 to examine the relationship between debt in 2010 and debt in 2005.



- Describe the relationship and compare it with the relationship between debt in 2010 and debt in 2009.
- Answer the question posed in the previous exercise for these data.

Of course, not all relationships are linear. Here is an example where the relationship is described by a curve.

## Example

### 2.12 Calcium retention.

Our bodies need calcium to build strong bones. How much calcium do we need? Does the amount that we need depend on our age? Questions like these

are studied by nutrition researchers. One series of studies used the amount of calcium retained by the body as a response variable and the amount of calcium consumed as an explanatory variable.<sup>7</sup>

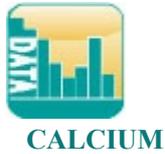
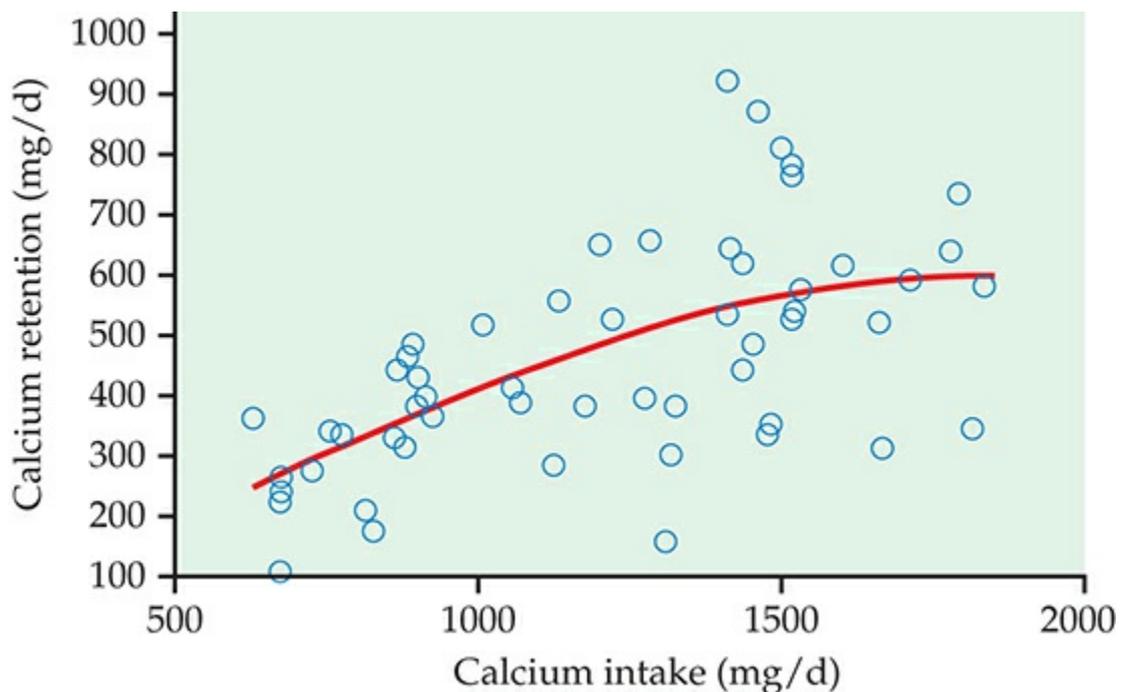


Figure 2.5 is a scatterplot of calcium retention in milligrams per day (mg/d) versus calcium intake (mg/d) for 56 children aged 11 to 15 years. A smooth curve generated by software helps us see the relationship between the two variables.

There is clearly a relationship here. As calcium intake increases, the body retains more calcium. However, the relationship is not linear. The curve is approximately linear for low values of intake, but then the line curves more and becomes almost level.



**FIGURE 2.5**

Scatterplot of calcium retention (mg/d) versus calcium intake (mg/d) for 56 children with a fitted curve, for Example 2.12. There is a positive relationship between these two variables but it is not linear.

There are many kinds of curved relationships like that in Figure 2.5. For some of these, we can apply a *transformation* to the data that will make the relationship approximately linear. To do this, we replace the original values with the transformed values and then use the transformed values for our analysis.

*transformation*

Transforming data is common in statistical practice. There are systematic principles that describe how transformations behave and guide the search for transformations that will, for example, make a distribution more Normal or a curved relationship more linear.

## The log transformation

The most important transformation that we will use is the *log transformation*. This transformation can be used for variables that have positive values only. Occasionally, we use it when there are zeros, but in this case we first replace the zero values by some small value, often one-half of the smallest positive value in the data set.

### log transformation

You have probably encountered logarithms in one of your high school mathematics courses as a way to do certain kinds of arithmetic. Logarithms are a lot more fun when used in statistical analyses. We will use natural logarithms. Statistical software and statistical calculators generally provide easy ways to perform this transformation.

Let's try a log transformation on our calcium retention data. Here are the details.

### Example

#### 2.13 Calcium retention with logarithms.

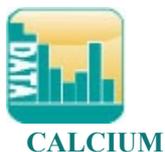
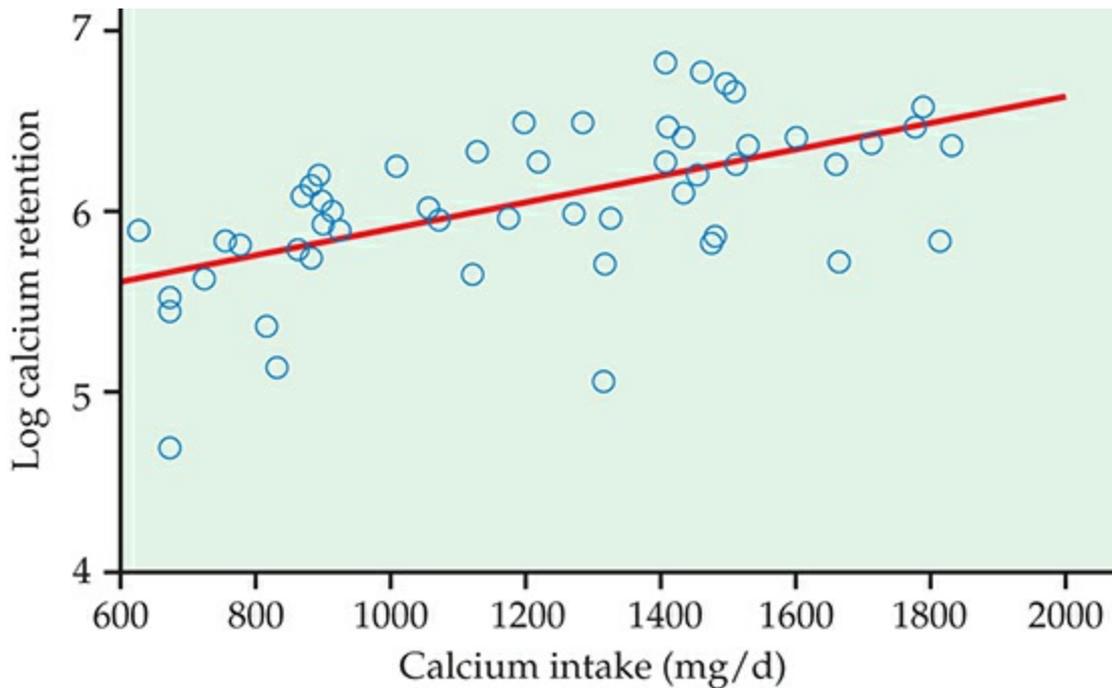


Figure 2.6 is a scatterplot of the log of calcium retention versus calcium intake. The plot includes a fitted straight line to help us see the relationship. We see that the transformation has worked. Our relationship is now approximately linear.



**FIGURE 2.6**

Scatterplot of log calcium retention versus calcium intake, with a fitted line, for 56 children, for Example 2.13. The relationship is approximately linear.

Our analysis of the calcium retention data in Examples 2.12 and 2.13 reminds us of an important issue when describing relationships. In Example 2.12 we noted that the relationship appeared to become approximately flat. Biological processes are consistent with this observation. There is probably a point where additional intake does not result in any additional retention. With our transformed relationship in Figure 2.6, however, there is no leveling off as we saw in Figure 2.5, even though we appear to have a good fit to the data. The relationship and fit apply to the range of data that are analyzed. *We cannot assume that the relationship extends beyond the range of the data.*



Use of transformations and the interpretation of scatterplots are an art that requires judgment and knowledge about the variables that we are studying. *Always ask yourself if the relationship that you see makes sense.* If it does not, then additional analyses are needed to understand the data.



### **Adding categorical variables to scatterplots**

In Example 2.9 (page 88) we looked at the relationship between the rating and the price per load for 24 laundry detergents. A more detailed look at the data shows that there are three different types of laundry detergent included in this data set. In Exercise 2.12 we saw that two of the detergents were gels. The other two types are liquid and powder. Let's examine where these three types of laundry detergents are in our plot.

## CATEGORICAL VARIABLES IN SCATTERPLOTS

To add a categorical variable to a scatterplot, use a different plot color or symbol for each category.

### Example

#### 2.14 Rating versus price and type of laundry detergent.

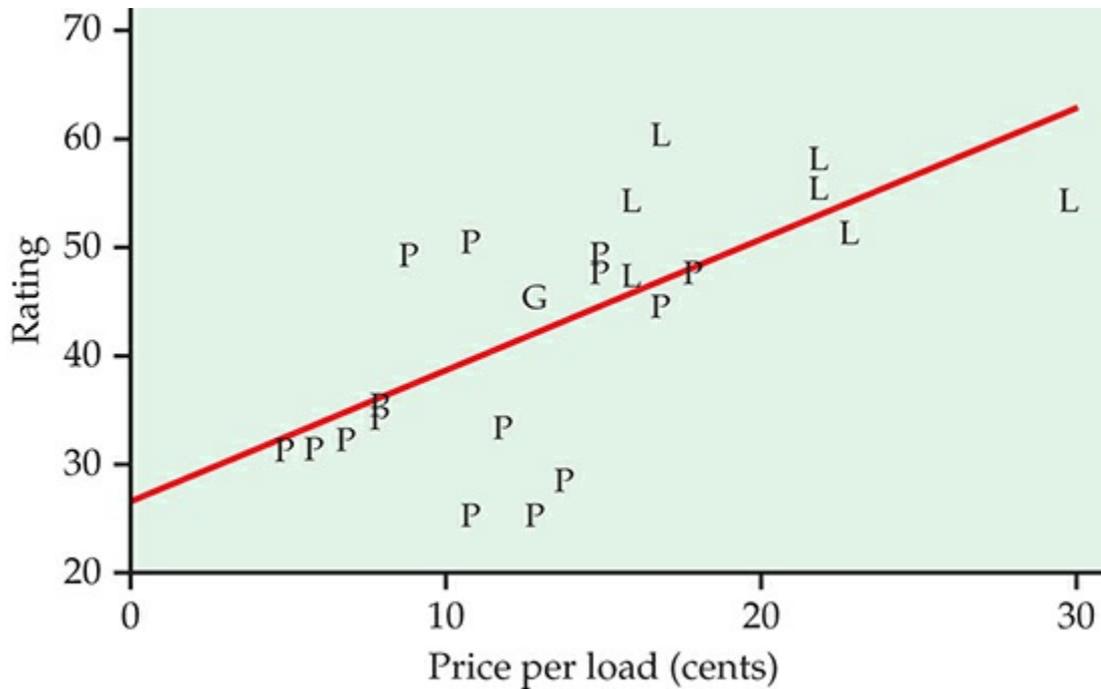
In our scatterplot, we use the symbol “G” for gels, “L” for liquids, and “P” for powders. The scatterplot with these plotting symbols is given in Figure 2.7.



The two gels appear in the middle of the plot as a single point because the ratings and prices are identical. There is a tendency for the liquids to be clustered in the upper right of the plot, with high ratings and high prices. In contrast, the powders tend to be in the left, with low ratings and low prices.

In this example, we used a categorical variable, type, to distinguish the three types of laundry detergents in our plot. Suppose that the additional variable that we want to investigate is quantitative. In this situation, we sometimes can combine the values into ranges of the quantitative variable, such as high, medium, and low, to create a categorical variable.

*Careful judgment is needed in using this graphical method.* Don't be discouraged if your first attempt is not very successful. In performing a good data analysis, you will often produce several plots before you find the one that you believe to be the most effective in describing the data.<sup>8</sup>



**FIGURE 2.7**

Scatterplot of rating versus price per load (in cents), with a fitted straight line, for 24 laundry detergents, for Example 2.14. The type of detergent is indicated by the plotting symbol; “G” for gel, “L” for liquid, and “P” for powder.

## USE YOUR KNOWLEDGE

### 2.17 Is a linear relationship the best description?

Look carefully at the plot in Figure 2.7.



#### LAUNDRY

(a) Do you think that the linear relationship we found between rating and price is mostly due to the difference between liquid and powder detergents? Explain your answer.

(b) In describing the laundry detergent data would you say that (i) there is a linear relationship between rating and price or (ii) powders cost less and have lower ratings; liquids cost more and have higher ratings; and gels are somewhere in the middle? Give reasons for your answer.

## BEYOND THE BASICS

### Scatterplot smoothers

The relationship in Figure 2.4 (page 92) appears to be linear. Some statistical software packages provide a tool to help us make this kind of judgment. These use computer-intensive methods called *algorithms* that calculate a smooth curve that gives an approximate fit to the points in a scatterplot. This is called *smoothing* a scatterplot. Usually, these methods use a smoothing parameter that determines how smooth the fit will be. You can vary it until you have a fit that you judge suitable for your data. Here is an example.

**algorithms**

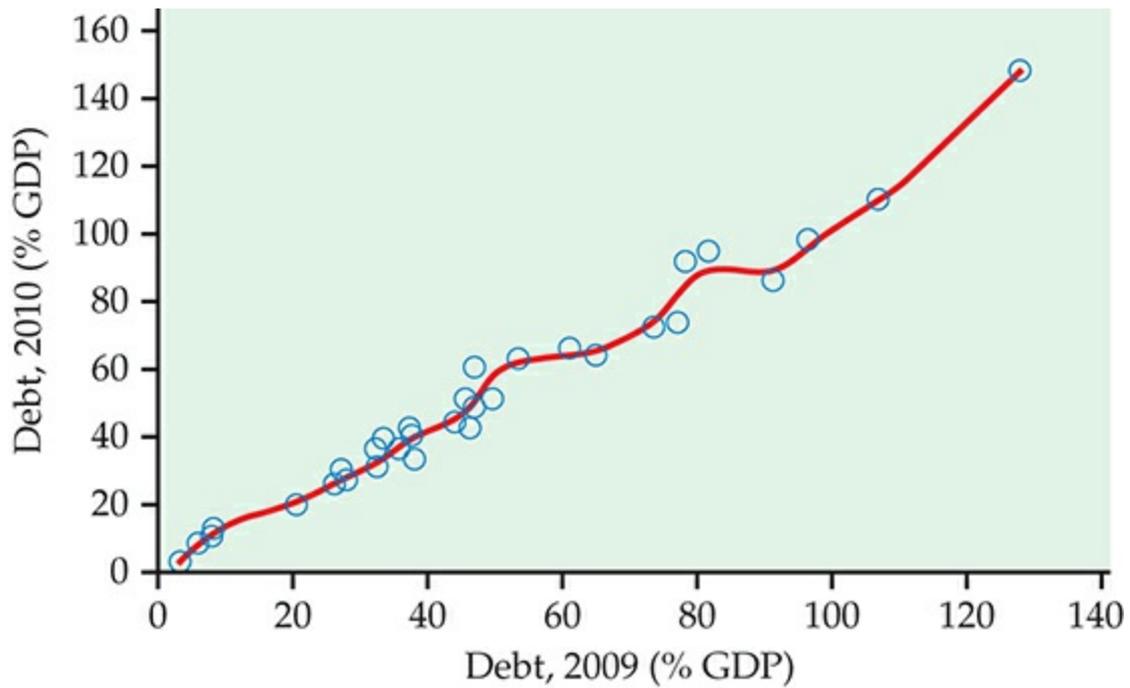
**smoothing**

### Example

#### 2.15 Debt for 33 countries with a smooth fit.

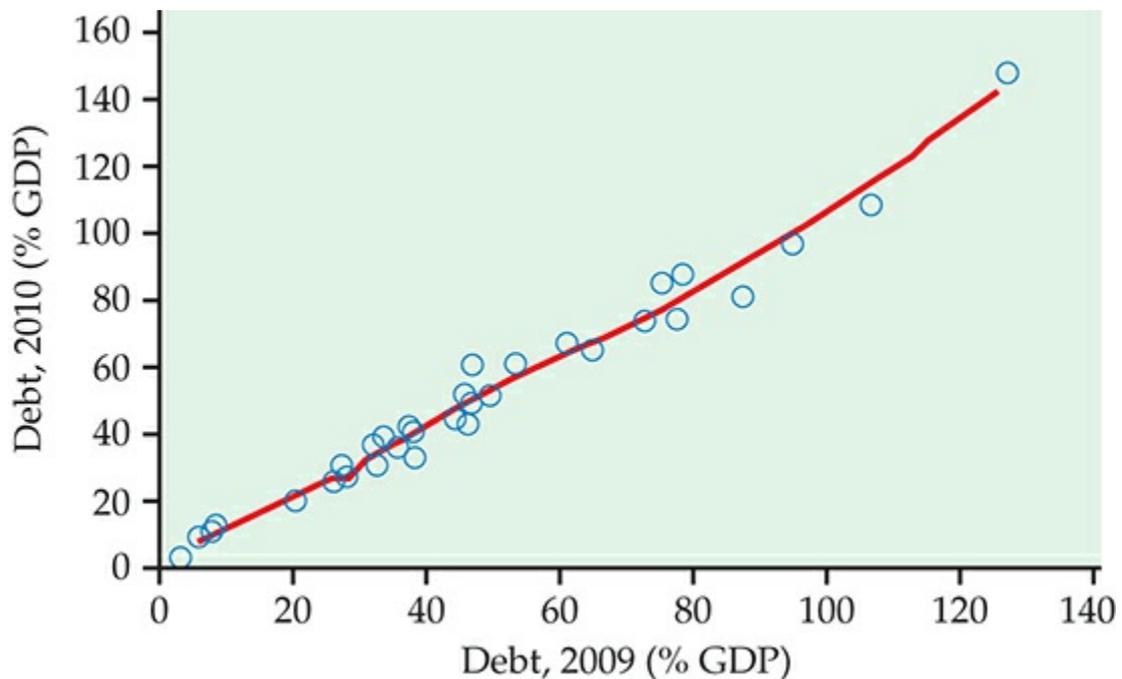


Figure 2.8 gives the scatterplot that we examined in Figure 2.4 with a smooth fit. Notice that the smooth curve fits almost all the points. However, the curve is too wavy and does not provide a good summary of the relationship.



**FIGURE 2.8**

Scatterplot of debt in 2010 (percent of GDP) versus debt in 2009 (percent of GDP), with a smooth curve fitted to the data, for 33 countries, for Example 2.15. This smooth curve fits the data too well and does not provide a good summary of the relationship.



**FIGURE 2.9**

Scatterplot of debt in 2010 (percent of GDP) versus debt in 2009 (percent of GDP), with a smooth curve fitted to the data, for 33 countries, for Example 2.16. This smooth curve gives a good summary of the relationship. It is approximately linear.

Our first attempt at smoothing the data was not very successful. This scenario happens frequently when we use data analysis methods to learn something from

our data. *Don't be discouraged when your first attempt at summarizing data produces unsatisfactory results.* Take what you learn and refine your analysis until you are satisfied that you have found a good summary. It is your last attempt, not your first, that is most important.



## Example

### 2.16 A better smooth fit for the debt data.



By varying the smoothing parameter, we can make the curve more or less smooth. Figure 2.9 gives the same data as in the previous figure but with a better smooth fit. The smooth curve is very close to a straight line. In this way we have confirmed our original impression that the relationship between these two variables is approximately linear.

## Categorical explanatory variables

Scatterplots display the association between two quantitative variables. To display a relationship between a categorical variable and a quantitative variable, make a side-by-side comparison of the distributions of the response for each category. Back-to-back stemplots (page 14) and side-by-side boxplots (page 41) are useful tools for this purpose.

We will study methods for describing the association between two categorical variables in Section 2.6 (page 139).

## SECTION 2.2 Summary

A **scatterplot** displays the relationship between two quantitative variables. Mark values of one variable on the horizontal axis ( $x$  axis) and values of the other variable on the vertical axis ( $y$  axis). Plot each individual's data as a point on the graph.

Always plot the explanatory variable, if there is one, on the  $x$  axis of a scatterplot. Plot the response variable on the  $y$  axis.

Plot points with different colors or symbols to see the effect of a categorical variable in a scatterplot.

In examining a scatterplot, look for an overall pattern showing the **form**, **direction**, and **strength** of the relationship, and then for **outliers** or other deviations from this pattern.

**Form: Linear relationships**, where the points show a straight-line pattern, are an important form of relationship between two variables. Curved relationships are other forms to watch for.

**Direction:** If the relationship has a clear direction, we speak of either **positive association** (high values of the two variables tend to occur together) or **negative association** (high values of one variable tend to occur with low values of the other variable).

**Strength:** The **strength** of a relationship is determined by how close the points in the scatterplot lie to a simple form such as a line.

To display the relationship between a categorical explanatory variable and a quantitative response variable, make a graph that compares the distributions of the response for each category of the explanatory variable.

## SECTION 2.2 Exercises

*For Exercises 2.10 and 2.11, see page 88; for Exercises 2.12 and 2.13, see page 89; for Exercise 2.14, see page 92; for Exercises 2.15 and 2.16, see page 92; and for Exercise 2.17, see page 96.*

### 2.18 Bone strength.

Osteoporosis is a condition where bones become weak. It affects more than 200 million people worldwide. Exercise is one way to produce strong bones and to prevent osteoporosis. Since we use our dominant arm (the right arm for most people) more than our nondominant arm, we expect the bone in our dominant arm to be stronger than the bone in our nondominant arm. By comparing the strengths, we can get an idea of the effect that exercise can have on bone strength. Here are some data on the strength of bones, measured in  $\text{cm}^4/1000$ , for the arms of 15 young men:<sup>9</sup>  **ARMSTR**

ID	Nondominant	Dominant	ID	Nondominant	Dominant
1	15.7	16.3	9	15.9	20.1
2	25.2	26.9	10	13.7	18.7
3	17.9	18.7	11	17.7	18.7
4	19.1	22.0	12	15.5	15.2
5	12.0	14.8	13	14.4	16.2
6	20.0	19.8	14	14.1	15.0
7	12.3	13.1	15	12.3	12.9
8	14.4	17.5			

Before attempting to compare the arm strengths of the dominant and nondominant arms, let's take a careful look at the data for these two variables.

- (a) Make a scatterplot of the data with the nondominant arm strength on the  $x$  axis and the dominant arm strength on the  $y$  axis.
- (b) Describe the overall pattern in the scatterplot and any striking deviations from the pattern.
- (c) Describe the form, direction, and strength of the relationship.
- (d) Identify any outliers.
- (e) Is the relationship approximately linear?

## 2.19 Bone strength for baseball players.

Refer to the previous exercise. The study collected arm bone strength information for two groups of young men. The data in the previous exercise were for a control group. The second group in the study comprised men who played baseball. We know that these baseball players use their dominant arm in throwing (those who throw with their nondominant arm were excluded), so they get more arm exercise than the controls.

Here are the data for the baseball players:  **ARMSTR**

ID	Nondominant	Dominant	ID	Nondominant	Dominant
16	17.0	19.3	24	15.1	19.4
17	16.9	19.0	25	13.5	20.4
18	17.7	25.2	26	13.6	17.1
19	21.2	37.7	27	20.3	26.5
20	21.0	40.3	28	17.3	30.3
21	14.6	20.8	29	14.6	17.4
22	31.5	36.9	30	22.6	35.0
23	14.9	21.2			

Answer the questions in the previous exercise for the baseball players.

## 2.20 Compare the baseball players with the controls.

Refer to the previous two exercises.  **ARMSTR**

- (a) Plot the data for the two groups on the same graph using different symbols for the baseball players and the controls.
- (b) Use your plot to describe and compare the relationships for the two variables. Write a short paragraph summarizing what you have found.

## 2.21 College students by state.

In Example 1.19 (page 21) we examined the distribution of undergraduate college students in the United States and displayed the histogram for these data in Figure 1.11. We noted that we could explain some of the variation in this distribution by considering the populations of the states. In Example 1.20, we transformed the number of undergraduate college students into the number of undergraduates per 1000 population. Let's look at these data a little differently. Let's examine the relationship between two variables: number of college students and population of the state.  **COLLEGE**

- (a) Which variable do you choose to be the explanatory variable? Which variable do you choose to be the

response variable? Give reasons for your choices.

(b) Make a scatterplot of the two variables and write a short paragraph describing the relationship.

## 2.22 Decay of a radioactive element.

Barium-137m is a radioactive form of the element barium that decays very rapidly. It is easy and safe to use for lab experiments in schools and colleges.<sup>10</sup> In a typical experiment, the radioactivity of a sample of barium-137m is measured for one minute. It is then measured for three additional one-minute periods, separated by two minutes. So data are recorded at 1, 3, 5, and 7 minutes after the start of the first counting period. The measurement units are counts. Here are the data for one of these experiments.<sup>11</sup>  DECAy

Time	1	3	5	7
Count	578	317	203	118

(a) Make a scatterplot of the data. Give reasons for the choice of which variables to use on the  $x$  and  $y$  axes.

(b) Describe the overall pattern in the scatterplot and any striking deviations from the pattern.

(c) Describe the form, direction, and strength of the relationship.

(d) Identify any outliers.

(e) Is the relationship approximately linear?

## 2.23 Use a log for the radioactive decay.

Refer to the previous exercise. Transform the counts using a log transformation. Then repeat parts (a) through (e) for the transformed data and compare your results with those from the previous exercise.  DECAy

## 2.24 Make some sketches.

For each of the following situations, make a scatterplot that illustrates the given relationship between two variables.

(a) A weak negative relationship.

(b) No apparent relationship.

(c) A strong positive linear relationship.

(d) A more complicated relationship. Explain the relationship.

## 2.25 What's wrong?

Explain what is wrong with each of the following:

(a) If two variables are negatively associated, then high values of one variable are associated with high values of the other variable.

(b) In a scatterplot we put the response variable on the  $x$  axis and the explanatory variable on the  $y$  axis.

(c) A histogram can be used to examine the relationship between two variables.

## 2.26 What's in the beer?

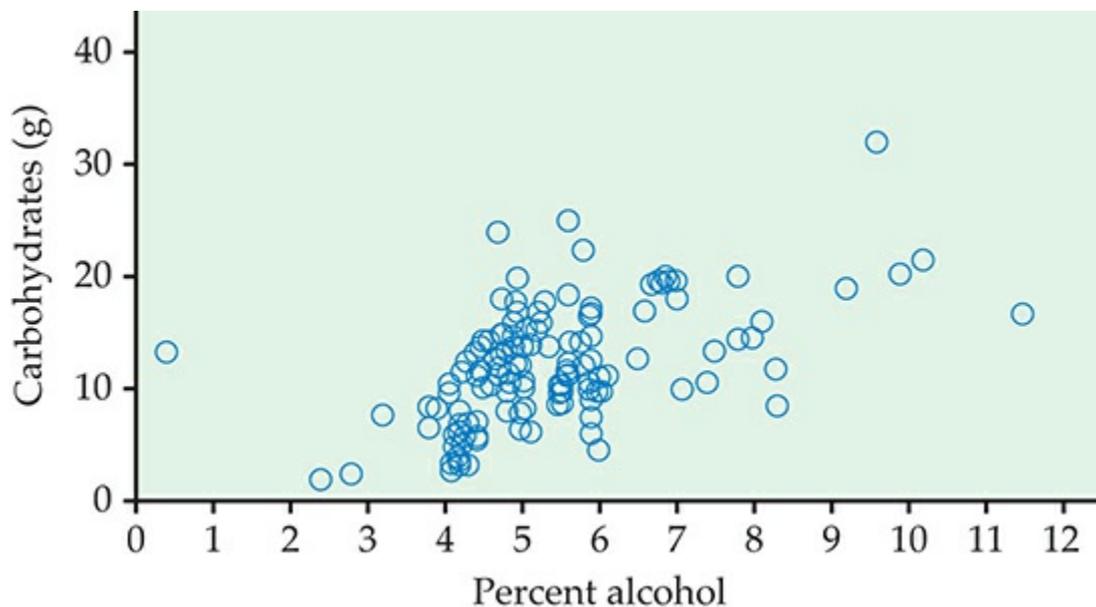
The website **beer100.com** advertises itself as “Your Place for All Things Beer.” One of their “things” is a list of 153 domestic beer brands with the percent alcohol, calories per 12 ounces, and carbohydrates per 12 ounces (in grams).<sup>12</sup>  **BEER**

(a) Figure 2.10 gives a scatterplot of carbohydrates versus percent alcohol. Give a short summary of what can be learned from the plot.

(b) One of the points is an outlier. Use the data file to find the outlier brand of beer. How is this brand of beer marketed compared with the other brands?

(c) Remove the outlier from the data set and generate a scatterplot of the remaining data.

(d) Describe the relationship between carbohydrates and percent alcohol based on what you see in your scatterplot.



**FIGURE 2.10**

Scatterplot of carbohydrates versus percent alcohol for 153 brands of beer, for Exercise 2.26.

## 2.27 More beer.

Refer to the previous exercise.  **BEER**

(a) Make a scatterplot of calories versus percent alcohol using the data set without the outlier.

(b) Describe the relationship between these two variables.

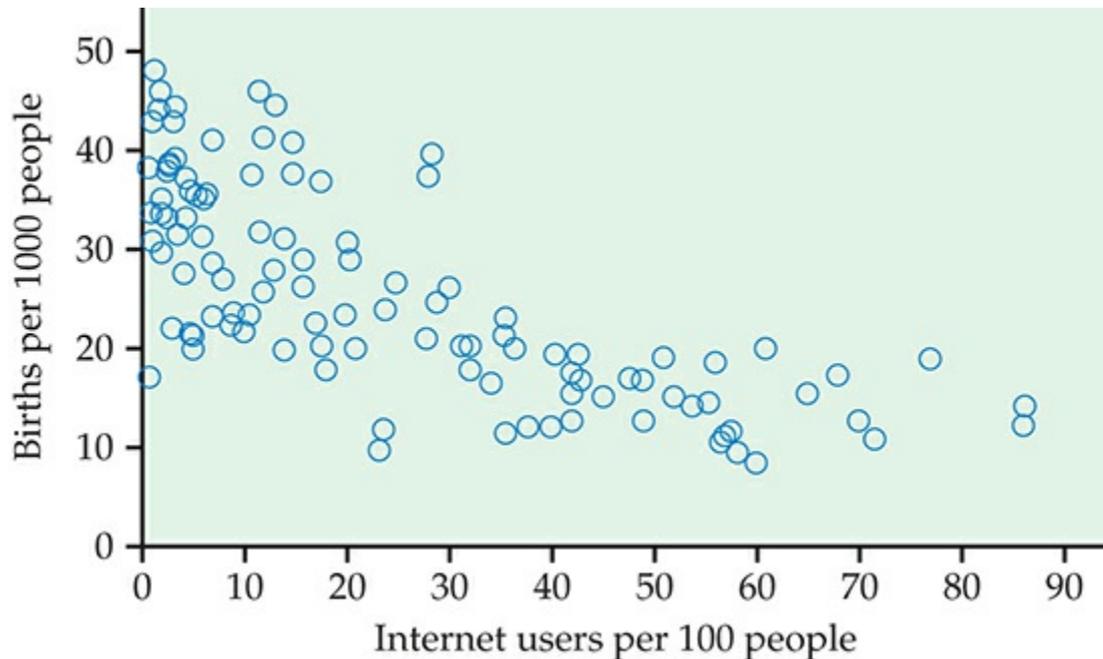
## 2.28 Internet use and babies.

The World Bank collects data on many variables related to world development for countries throughout the world. Two of these are Internet use, in number of users per 100 people, and birthrate, in births per 1000

people.<sup>13</sup> Figure 2.11 is a scatterplot of birthrate versus Internet use for the 106 countries that have data available for both variables.  INBIRTH

(a) Describe the relationship between these two variables.

(b) A friend looks at this plot and concludes that using the Internet will decrease the number of babies born. Write a short paragraph explaining why the association seen in the scatterplot does not provide a reason to draw this conclusion.



**FIGURE 2.11**

Scatterplot of births (per 1000 people) versus Internet users (per 100 people) for 106 countries, for Exercise 2.28.

## 2.29 Try a log.

Refer to the previous exercise.  INBIRTH

(a) Make a scatterplot of the log of births per 1000 people versus Internet users per 100 people.

(b) Describe the relationship that you see in this plot and compare it with Figure 2.11.

(c) Which plot do you prefer? Give reasons for your answer.

## 2.30 Make another plot.

Refer to Exercise 2.28.  INBIRTH

(a) Make a new data set that has Internet users expressed as users per 10,000 people and births as births per 10,000 people.

(b) Explain why these transformations to give new variables are linear transformations. (*Hint:* See linear transformations on page 45.)

- (c) Make a scatterplot using the transformed variables.
- (d) Compare your new plot with the one in Figure 2.11.
- (e) Why do you think that the analysts at the World Bank chose to express births as births per 1000 people and Internet users as users per 100 people?

### **2.31 Explanatory and response variables.**

In each of the following situations, is it more reasonable to simply explore the relationship between the two variables or to view one of the variables as an explanatory variable and the other as a response variable? In the latter case, which is the explanatory variable and which is the response variable?

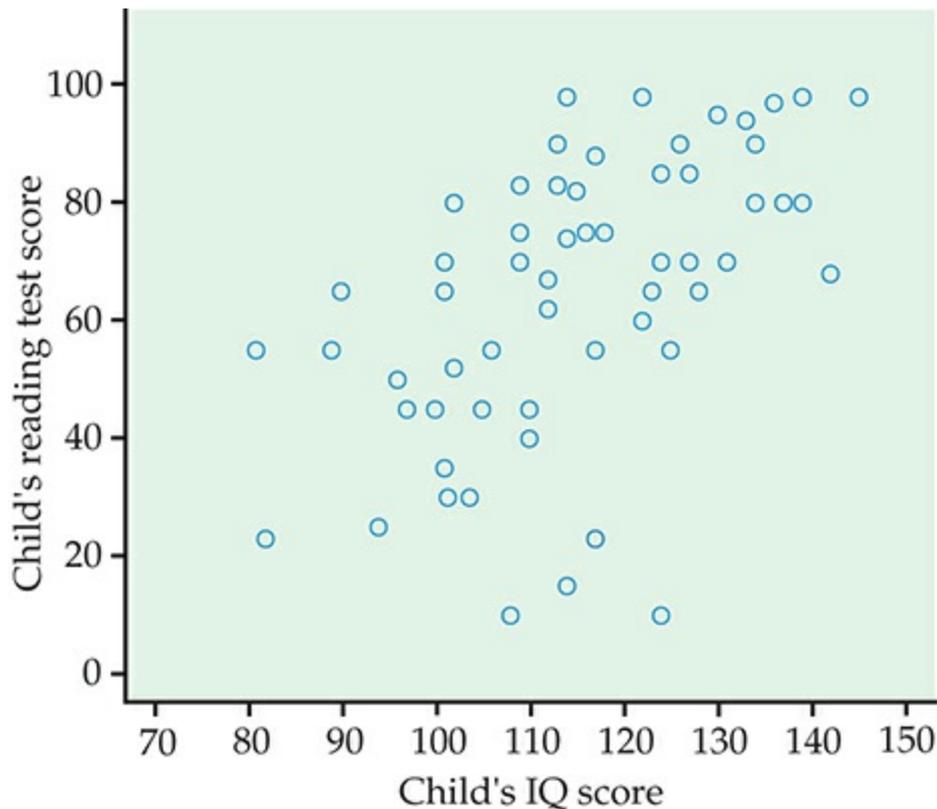
- (a) The reading ability of a child and the shoe size of the child.
- (b) College grade point average and high school grade point average.
- (c) The rental price of an apartment and the number of square feet in the apartment.
- (d) The amount of sugar added to a cup of coffee and how sweet the coffee tastes.
- (e) The temperature outside today at noon and the temperature outside yesterday at noon.

### **2.32 Parents' income and student loans.**

How well does the income of a college student's parents predict how much the student will borrow to pay for college? We have data on parents' income and college debt for a sample of 1200 recent college graduates. What are the explanatory and response variables? Are these variables categorical or quantitative? Do you expect a positive or negative association between these variables? Why?

### **2.33 Reading ability and IQ.**

A study of reading ability in schoolchildren chose 60 fifth-grade children at random from a school. The researchers had the children's scores on an IQ test and on a test of reading ability.<sup>14</sup> Figure 2.12 plots reading test score (response) against IQ score (explanatory).



**FIGURE 2.12**

IQ and reading test scores for 60 fifth-grade children, for Exercise 2.33.

- Explain why we should expect a positive association between IQ and reading score for children in the same grade. Does the scatterplot show a positive association?
- A group of four points appear to be outliers. In what way do these children's IQ and reading scores deviate from the overall pattern?
- Ignoring the outliers, is the association between IQ and reading score roughly linear? Is it very strong? Explain your answers.

### 2.34 Can children estimate their reading ability?

The main purpose of the study cited in Exercise 2.33 was to ask whether schoolchildren can estimate their own reading ability. The researchers had the children's scores on a test of reading ability. They asked each child to estimate his or her reading level, on a scale from 1 (low) to 5 (high). Figure 2.13 is a scatterplot of the children's estimates (response) against their reading scores (explanatory).

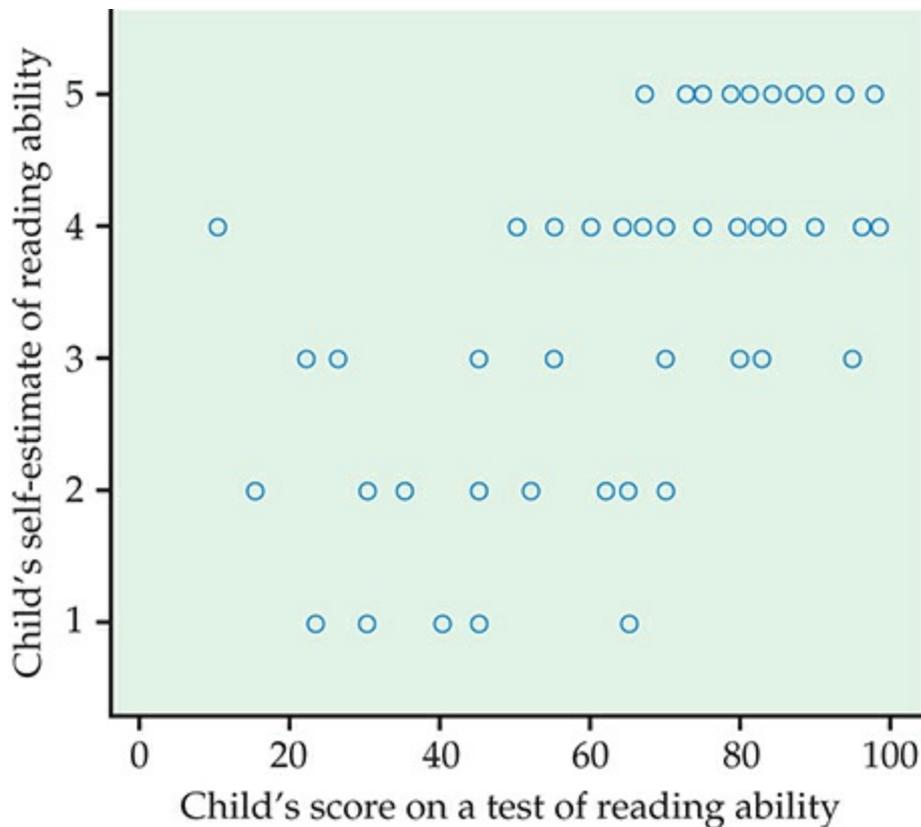
- What explains the "stair-step" pattern in the plot?
- Is there an overall positive association between reading score and self-estimate?
- There is one clear outlier. What is this child's self-estimated reading level? Does this appear to over- or underestimate the level as measured by the test?



### 2.35 Body mass and metabolic rate.

Metabolic rate, the rate at which the body consumes energy, is important in studies of weight gain, dieting, and exercise. The following table gives data on the lean body mass and resting metabolic rate for 12

women and 7 men who are subjects in a study of dieting. Lean body mass, given in kilograms, is a person's weight leaving out all fat. Metabolic rate is measured in calories burned per 24 hours, the same calories used to describe the energy content of foods. The researchers believe that lean body mass is an important influence on metabolic rate.  **BMAS**



**FIGURE 2.13**

Reading test scores for 60 fifth-grade children and the children's estimates of their own reading levels, for Exercise 2.34.

Subject	Sex	Mass	Rate	Subject	Sex	Mass	Rate
1	M	62.0	1792	11	F	40.3	1189
2	M	62.9	1666	12	F	33.1	913
3	F	36.1	995	13	M	51.9	1460
4	F	54.6	1425	14	F	42.4	1124
5	F	48.5	1396	15	F	34.5	1052
6	F	42.0	1418	16	F	51.1	1347
7	M	47.4	1362	17	F	41.2	1204
8	F	50.6	1502	18	M	51.9	1867
9	F	42.0	1256	19	M	46.9	1439
10	M	48.7	1614				

(a) Make a scatterplot of the data, using different symbols or colors for men and women.

(b) Is the association between these variables positive or negative? What is the form of the relationship? How strong is the relationship? Does the pattern of the relationship differ for women and men? How do the male subjects as a group differ from the female subjects as a group?

### 2.36 Team value in the NFL.

Management theory says that the value of a business should depend on its operating income, the income produced by the business after taxes. (Operating income excludes income from sales of assets and investments, which don't reflect the actual business.) Total revenue, which ignores costs, should be less important. Debt includes borrowing for the construction of a new arena. The data file NFL gives the value (in millions of dollars), debt (as percent of value), revenue (in millions of dollars), and operating income (in millions of dollars) of the 32 teams in the National Football League (NFL).<sup>15</sup>  NFL

- Plot team value against revenue. Describe the relationship.
- Plot team value against debt. Describe the relationship.
- Plot team value against operating income. Describe the relationship.
- Write a short summary comparing the relationships that you described in parts (a), (b), and (c) of this exercise.

### 2.37 Records for men and women in the 10K.

Table 2.1 shows the progress of world record times (in seconds) for the 10,000-meter run for both men and women.<sup>16</sup>  TENK

- Make a scatterplot of world record time against year, using separate symbols for men and women. Describe the pattern for each sex. Then compare the progress of men and women.
- Women began running this long distance later than men, so we might expect their improvement to be more rapid. Moreover, it is often said that men have little advantage over women in distance running as opposed to sprints, where muscular strength plays a greater role. Do the data appear to support these claims?

Men		Women	
Record year	Time (seconds)	Record year	Time (seconds)
1912	1880.8	1963	1695.6
1921	1840.2	1965	1659.3
1924	1835.4	1972	1658.4
1924	1823.2	1973	1650.8
1924	1806.2	1977	1650.5
1937	1805.6	1978	1642.4
1938	1802.0	1984	1633.8
1939	1792.6	1989	1628.2
1944	1775.4	1993	1627.9
1949	1768.2	1993	1618.4
1949	1767.2	1994	1612.2
1949	1761.2	1995	1603.5
1950	1742.6	1996	1598.1
1953	1741.6	1997	1591.3
		1967	2286.4
		1970	2130.5
		1975	2100.4
		1975	2041.4
		1977	1995.1
		1979	1972.5
		1981	1950.8
		1981	1937.2
		1982	1895.3
		1983	1895.0
		1983	1887.6
		1984	1873.8
		1985	1859.4
		1986	1813.7

1954	1734.2	1997	1587.8	1993	1771.8
1956	1722.8	1998	1582.7		
1956	1710.4	2004	1580.3		
1960	1698.8	2005	1577.3		
1962	1698.2				

## 2.3 Correlation

When you complete this section, you will be able to

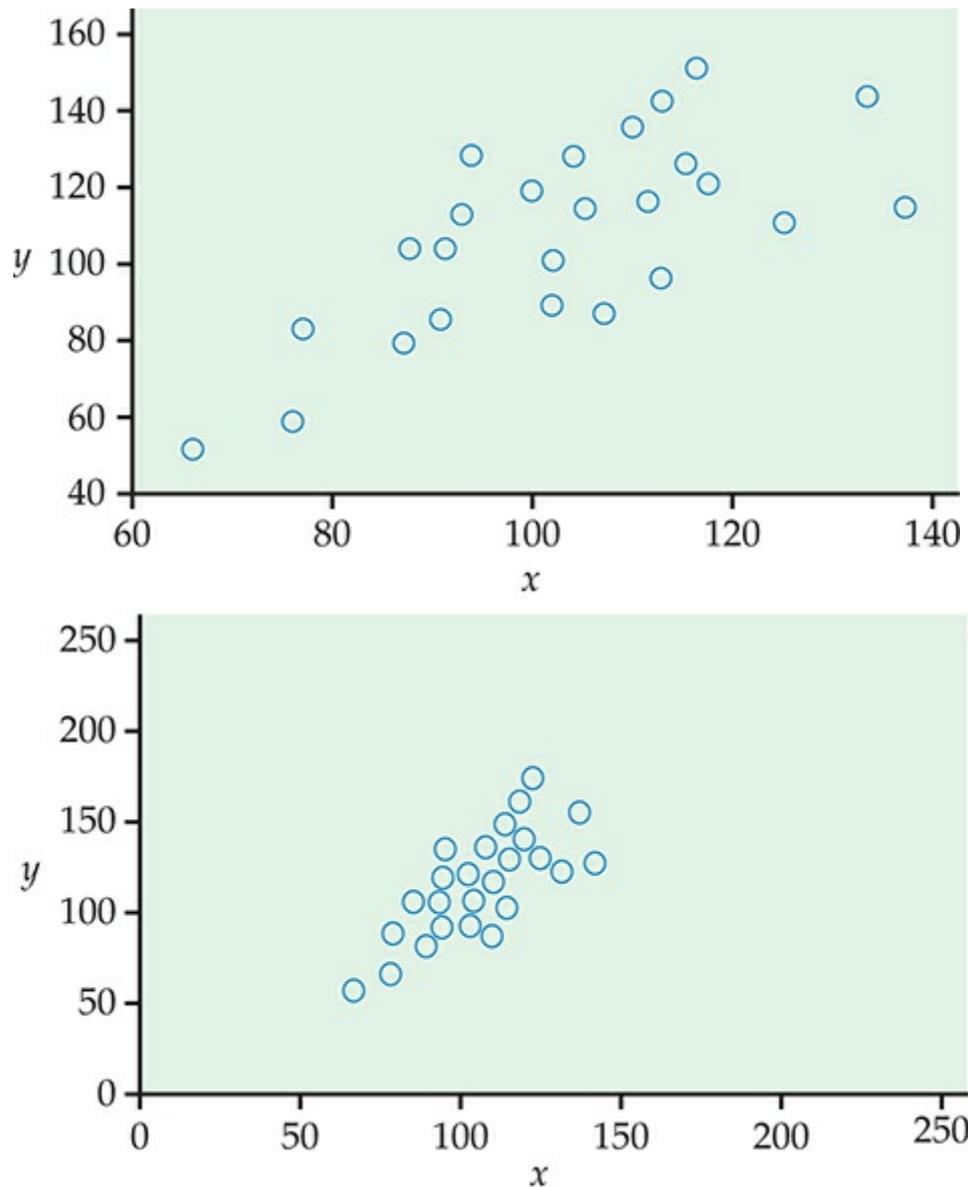
- Use a correlation to describe the direction and strength of a linear relationship between two quantitative variables.
- Interpret the sign of a correlation.
- Identify situations where the correlation is not a good measure of association between two quantitative variables.
- Identify a linear pattern in a scatterplot.
- For describing the relationship between two quantitative variables, identify the roles of the correlation, a numerical summary, and the scatterplot (a graphical summary).

A scatterplot displays the form, direction, and strength of the relationship between two quantitative variables. Linear (straight-line) relations are particularly important because a straight line is a simple pattern that is quite common. We say a linear relationship is strong if the points lie close to a straight line, and weak if they are widely scattered about a line. Our eyes are not good judges of how strong a relationship is. The two scatterplots in Figure 2.14 depict exactly the same data, but the plot on the right is drawn smaller in a large field. The plot on the right seems to show a stronger relationship.

Our eyes can be fooled by changing the plotting scales or the amount of white space around the cloud of points in a scatterplot.<sup>17</sup> We need to follow our strategy for data analysis by using a numerical measure to supplement the graph. *Correlation* is the measure we use.

### The correlation $r$

We have data on variables  $x$  and  $y$  for  $n$  individuals. Think, for example, of measuring height and weight for  $n$  people. Then  $x_1$  and  $y_1$  are your height and your weight,  $x_2$  and  $y_2$  are my height and my weight, and so on. For the  $i$ th individual, height  $x_i$  goes with weight  $y_i$ . Here is the definition of correlation.



**FIGURE 2.14**

Two scatterplots of the same data. The linear pattern in the plot on the right appears stronger because of the surrounding space.

## CORRELATION

The **correlation** measures the direction and strength of the linear relationship between two quantitative variables. Correlation is usually written as  $r$ .

Suppose that we have data on variables  $x$  and  $y$  for  $n$  individuals. The means and standard deviations of the two variables are  $\bar{x}$  and  $s_x$  for the  $x$ -values, and  $\bar{y}$  and  $s_y$  for the  $y$ -values. The correlation  $r$  between  $x$  and  $y$  is

$$r = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / (s_x s_y)$$

As always, the summation sign  $\Sigma$  means “add these terms for all the individuals.” The formula for the correlation  $r$  is a bit complex. It helps us see what correlation is but is not convenient for actually calculating  $r$ . In practice you should use software or a calculator that finds  $r$  from keyed-in values of two variables  $x$  and  $y$ .

The formula for  $r$  begins by standardizing the observations. Suppose, for example, that  $x$  is height in centimeters and  $y$  is weight in kilograms and that we have height and weight measurements for  $n$  people. Then  $\bar{x}$  and  $s_x$  are the mean and standard deviation of the  $n$  heights, both in centimeters. The value

$$\frac{x_i - \bar{x}}{s_x}$$

is the standardized height of the  $i$ th person. The standardized height says how many standard deviations above or below the mean a person’s height lies. Standardized values have no units—in this example, they are no longer measured in centimeters. Standardize the weights also. The correlation  $r$  is an average of the products of the standardized height and the standardized weight for the  $n$  people.

## USE YOUR KNOWLEDGE

### 2.38 Laundry detergents.

Example 2.8 describes data on the rating and price per load for 24 laundry detergents. Use these data to compute the correlation between rating and the price per load.



LAUNDRY

### 2.39 Change the units.

Refer to the previous exercise. Express the price per load in dollars.



LAUNDRY

(a) Is the transformation from cents to dollars a linear transformation? Explain your answer.

- (b) Compute the correlation between rating and price per load expressed in dollars.
- (c) How does the correlation that you computed in part (b) compare with the one you computed in the previous exercise?
- (d) What can you say in general about the effect of changing units using linear transformations on the size of the correlation?

## Properties of correlation

The formula for correlation helps us see that  $r$  is positive when there is a positive association between the variables. Height and weight, for example, have a positive association. People who are above average in height tend to also be above average in weight. Both the standardized height and the standardized weight for such a person are positive. People who are below average in height tend also to have below-average weight. Then both standardized height and standardized weight are negative. In both cases, the products in the formula for  $r$  are mostly positive and so  $r$  is positive. In the same way, we can see that  $r$  is negative when the association between  $x$  and  $y$  is negative. More detailed study of the formula gives more detailed properties of  $r$ . Here is what you need to know in order to interpret correlation:

- Correlation makes no use of the distinction between explanatory and response variables. It makes no difference which variable you call  $x$  and which you call  $y$  in calculating the correlation.
- *Correlation requires that both variables be quantitative.* For example, we cannot calculate a correlation between the incomes of a group of people and what city they live in, because city is a categorical variable.



- Because  $r$  uses the standardized values of the observations,  $r$  does not change when we change the units of measurement (a linear transformation) of  $x$ ,  $y$ , or both. Measuring height in inches rather than centimeters and weight in pounds rather than kilograms does not change the correlation between height and weight. The correlation  $r$  itself has no unit of measurement; it is just a number.
- Positive  $r$  indicates positive association between the variables, and negative  $r$  indicates negative association.
- The correlation  $r$  is always a number between  $-1$  and  $1$ . Values of  $r$  near  $0$  indicate a very weak linear relationship. The strength of the relationship increases as  $r$  moves away from  $0$  toward either  $-1$  or  $1$ . Values of  $r$  close to  $-1$  or  $1$  indicate that the points lie close to a straight line. The extreme values  $r = -1$  and  $r = 1$  occur only when the points in a scatterplot lie exactly along a straight line.

- Correlation measures the strength of only the linear relationship between two variables. *Correlation does not describe curved relationships between variables, no matter how strong they are.*

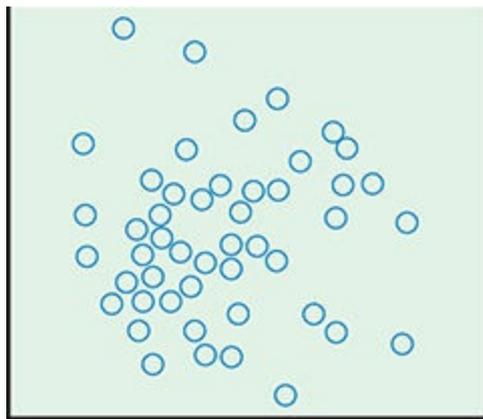


- *Like the mean and standard deviation, the correlation is not resistant:  $r$  is strongly affected by a few outlying observations. Use  $r$  with caution when outliers appear in the scatterplot.*

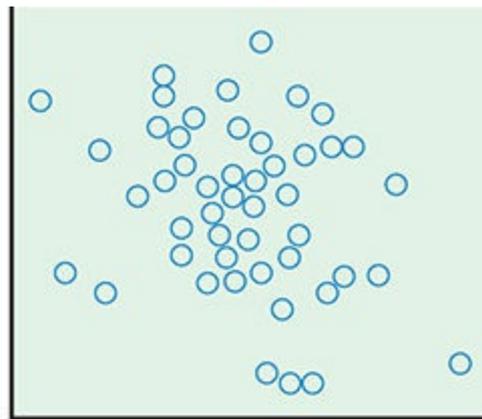


The scatterplots in Figure 2.15 illustrate how values of  $r$  closer to 1 or  $-1$  correspond to stronger linear relationships. To make the essential meaning of  $r$  clear, the standard deviations of both variables in these plots are equal and the horizontal and vertical scales are the same. In general, it is not so easy to guess the value of  $r$  from the appearance of a scatterplot. Remember that changing the plotting scales in a scatterplot may mislead our eyes, but it does not change the standardized values of the variables and therefore cannot change the correlation. To explore how extreme observations can influence  $r$ , use the *Correlation and Regression* applet available on the text website.

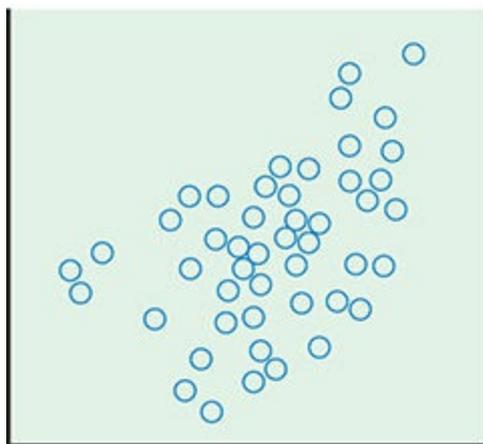




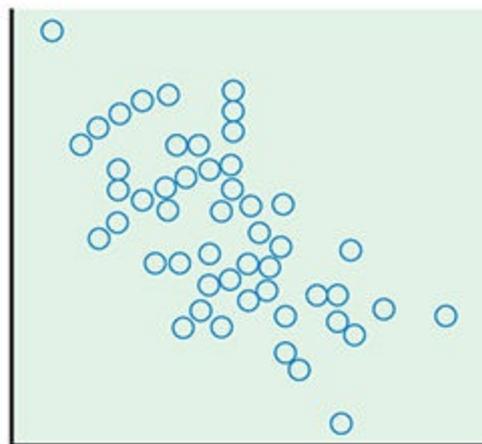
Correlation  $r = 0$



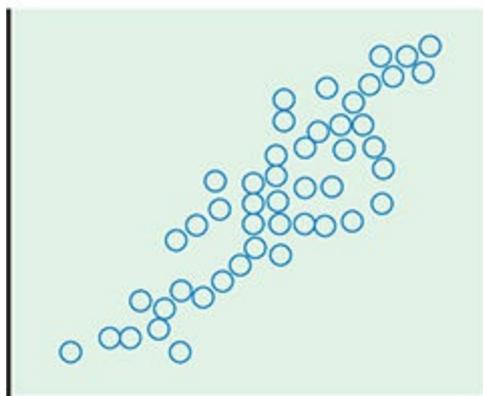
Correlation  $r = -0.3$



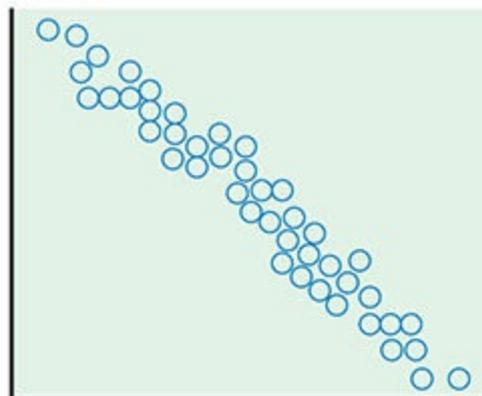
Correlation  $r = 0.5$



Correlation  $r = -0.7$



Correlation  $r = 0.9$



Correlation  $r = -0.99$

**FIGURE 2.15**

How the correlation  $r$  measures the direction and strength of a linear association.

Finally, remember that **correlation is not a complete description of two-variable data**, even when the relationship between the variables is linear. You should give the means and standard deviations of both  $x$  and  $y$  along with the correlation. (Because the formula for correlation uses the means and standard deviations, these measures are the proper choices to accompany a correlation.)

Conclusions based on correlations alone may require rethinking in the light of a more complete description of the data.

## Example

### 2.17 Scoring of figure skating in the Olympics.

Until a scandal at the 2002 Olympics brought change, figure skating was scored by judges on a scale from 0.0 to 6.0. The scores were often controversial. We have the scores awarded by two judges, Pierre and Elena, to many skaters. How well do they agree? We calculate that the correlation between their scores is  $r = 0.9$ . But the mean of Pierre's scores is 0.8 point lower than Elena's mean.

These facts in the example above do not contradict each other. They are simply different kinds of information. The mean scores show that Pierre awards lower scores than Elena. But because Pierre gives *every* skater a score about 0.8 point lower than Elena, the correlation remains high. Adding the same number to all values of either  $x$  or  $y$  does not change the correlation. If both judges score the same skaters, the competition is scored consistently because Pierre and Elena agree on which performances are better than others. The high  $r$  shows their agreement. But if Pierre scores some skaters and Elena others, we must add 0.8 point to Pierre's scores to arrive at a fair comparison.

## SECTION 2.3 Summary

The **correlation**  $r$  measures the direction and strength of the linear (straight line) association between two quantitative variables  $x$  and  $y$ . Although you can calculate a correlation for any scatterplot,  $r$  measures only linear relationships.

Correlation indicates the direction of a linear relationship by its sign:  $r > 0$  for a positive association and  $r < 0$  for a negative association.

Correlation always satisfies  $-1 \leq r \leq 1$  and indicates the strength of a relationship by how close it is to  $-1$  or  $1$ . Perfect correlation,  $r = \pm 1$ , occurs only when the points lie exactly on a straight line.

Correlation ignores the distinction between explanatory and response variables. The value of  $r$  is not affected by changes in the unit of measurement of either variable. Correlation is not resistant, so outliers can greatly change the value of  $r$ .

## SECTION 2.3 Exercises

For Exercises 2.38 and 2.39, see page 104.

## 2.40 Correlations and scatterplots.

Explain why you should always look at a scatterplot when you want to use a correlation to describe the relationship between two quantitative variables.

## 2.41 Interpret some correlations.

For each of the following correlations, describe the relationship between the two quantitative variables in terms of the direction and the strength of the linear relationship.

(a)  $r = 0.0$

(b)  $r = -0.9$

(c)  $r = 0.3$

(d)  $r = 0.8$

## 2.42 When should you not use a correlation?

Describe two situations where a correlation would not give a good numerical summary of the relationship between two quantitative variables. Illustrate each situation with a scatterplot and write a short paragraph explaining why the correlation would not be appropriate in each of these situations.

## 2.43 Bone strength.

Exercise 2.18 (page 98) gives the bone strengths of the dominant and the nondominant arms for 15 men who were controls in a study.  **ARMSTR**

(a) Find the correlation between the bone strength of the dominant arm and the bone strength of the nondominant arm.

(b) Look at the scatterplot for these data that you made in part (a) of Exercise 2.18 (or make one if you did not do that exercise). Is the correlation a good numerical summary of the graphical display in the scatterplot? Explain your answer.

## 2.44 Bone strength for baseball players.

Refer to the previous exercise. Similar data for baseball players is given in Exercise 2.19 (page 98). Answer parts (a) and (b) of the previous exercise for these data.  **ARMSTR**

## 2.45 College students by state.

In Exercise 2.21 (page 99) you used a scatterplot to display the relationship between the number of undergraduates and the populations of the states.  **COLLEGE, COL46**

(a) What is the correlation between these two variables?

(b) Does the correlation give a good numerical summary of the relationship between these two variables? Explain your answer.

(c) Eliminate the four states with populations greater than 15 million and find the correlation for the other 46 states. How does this correlation differ from the one that you found in part (a)? What does this tell you about how the range of the values of the variables in a data set can affect the magnitude of a correlation?

## 2.46 Decay of a radioactive element.

Data for an experiment on the decay of barium-137m is given in Exercise 2.22 (page 99).  **DECAY**

(a) Find the correlation between the radioactive counts and the time after the start of the first counting period.

(b) Does the correlation give a good numerical summary of the relationship between these two variables? Explain your answer.

## 2.47 Decay in the log scale.

Refer to the previous exercise and to Exercise 2.23 (page 99), where the counts were transformed by a log.  **DECAY**

(a) Find the correlation between the log counts and the time after the start of the first counting period.

(b) Does the correlation give a good numerical summary of the relationship between these two variables? Explain your answer.

(c) Compare your results for this exercise with those from the previous exercise.

## 2.48 Thinking about correlation.

Figure 2.9 (page 97) is a scatterplot of 2010 debt versus 2009 debt for 33 countries. Is the correlation  $r$  for these data near  $-1$ , clearly negative but not near  $-1$ , near  $0$ , clearly positive but not near  $1$ , or near  $1$ ?

Explain your answer. Verify your answer by doing the calculation.  **DEBT**

## 2.49 Brand names and generic products.

(a) If a store always prices its generic “store brand” products at 80% of the brand name products’ prices, what would be the correlation between the prices of the brand name products and the store brand products? (*Hint*: Draw a scatterplot for several prices.)

(b) If the store always prices its generic products \$2 less than the corresponding brand name products, then what would be the correlation between the prices of the brand name products and the store brand products?

## 2.50 Strong association but no correlation.

Here is a data set that illustrates an important point about correlation:

$X$	25	35	45	55	65
$Y$	10	30	50	30	10

- (a) Make a scatterplot of  $Y$  versus  $X$ .
- (b) Describe the relationship between  $Y$  and  $X$ . Is it weak or strong? Is it linear?
- (c) Find the correlation between  $Y$  and  $X$ .
- (d) What important point about correlation does this exercise illustrate?

### 2.51 Alcohol and carbohydrates in beer.

Figure 2.10 (page 100) gives a scatterplot of the percent alcohol versus carbohydrates in 153 brands of beer. Compute the correlation for these data.  **BEER**

### 2.52 Alcohol and carbohydrates in beer revisited.

Refer to the previous exercise. The data that you used to compute the correlation includes an outlier.  **BEER**

- (a) Remove the outlier and recompute the correlation.
- (b) Write a short paragraph about the possible effects of outliers on a correlation using this example to illustrate your ideas.

### 2.53 Internet use and babies.

Figure 2.11 (page 100) is a scatterplot of the number of births per 1000 people versus Internet users per 100 people for 106 countries. In Exercise 2.28 (page 100) you described this relationship.  **INBIRTH**

- (a) Make a plot of the data similar to Figure 2.11 and report the correlation.
- (b) Is the correlation a good numerical summary for this relationship? Explain your answer.

### 2.54 NFL teams.

In Exercise 2.36 (page 102) you used graphical summaries to examine the relationship between team value and three possible explanatory variables for 32 National Football League teams. Find the correlations for these variables. Do you think that these correlations provide good numerical summaries for the relationships? Explain your answers.  **NFL**

### 2.55 Use the applet.

You are going to use the *Correlation and Regression* applet to make different scatterplots with 10 points that have correlation close to 0.8. *Many patterns can have the same correlation. Always plot your data before you trust a correlation.*

- (a) Stop after adding the first 2 points. What is the value of the correlation? Why does it have this value no matter where the 2 points are located?
- (b) Make a lower-left to upper-right pattern of 10 points with correlation about  $r = 0.8$ . (You can drag points up or down to adjust  $r$  after you have 10 points.) Make a rough sketch of your scatterplot.

(c) Make another scatterplot, this time with 9 points in a vertical stack at the left of the plot. Add one point far to the right and move it until the correlation is close to 0.8. Make a rough sketch of your scatterplot.

(d) Make yet another scatterplot, this time with 10 points in a curved pattern that starts at the lower left, rises to the right, then falls again at the far right. Adjust the points up or down until you have a quite smooth curve with correlation close to 0.8. Make a rough sketch of this scatterplot also.



## 2.56 Use the applet.

Go to the *Correlation and Regression* applet. Click on the scatterplot to create a group of 10 points in the lower-right corner of the scatterplot with a strong straight-line negative pattern (correlation about  $-0.9$ ).

(a) Add one point at the upper left that is in line with the first 10. How does the correlation change?

(b) Drag this last point down until it is opposite the group of 10 points. How small can you make the correlation? Can you make the correlation positive? *A single outlier can greatly strengthen or weaken a correlation. Always plot your data to check for outlying points.*

## 2.57 An interesting set of data.

Make a scatterplot of the following data:  INTER

$x$	1	2	3	4	10	10
$y$	1	3	3	5	1	11

Use your calculator to show that the correlation is about 0.5. What feature of the data is responsible for reducing the correlation to this value despite a strong straight-line association between  $x$  and  $y$  in most of the observations?



## 2.58 High correlation does not mean that the values are the same.

Investment reports often include correlations. Following a table of correlations among mutual funds, a report adds, “Two funds can have perfect correlation, yet different levels of risk. For example, Fund A and Fund B may be perfectly correlated, yet Fund A moves 20% whenever Fund B moves 10%.” Write a brief explanation, for someone who knows no statistics, of how this can happen. Include a sketch to illustrate your explanation.

## 2.59 Student ratings of teachers.

A college newspaper interviews a psychologist about student ratings of the teaching of faculty members. The psychologist says, “The evidence indicates that the correlation between the research productivity and teaching rating of faculty members is close to zero.” The paper reports this as “Professor McDaniel said that good researchers tend to be poor teachers, and vice versa.” Explain why the paper’s report is wrong. Write a statement in plain language (don’t use the word “correlation”) to explain the psychologist’s meaning.

## 2.60 What’s wrong?

Each of the following statements contains a blunder. Explain in each case what is wrong.

- (a) “There is a high correlation between the age of American workers and their occupation.”
- (b) “We found a high correlation ( $r = 1.19$ ) between students’ ratings of faculty teaching and ratings made by other faculty members.”
- (c) “The correlation between the gender of a group of students and the color of their cell phone was  $r = 0.23$ .”

### 2.61 IQ and GPA.

Table 1.3 (page 29) reports data on 78 seventh-grade students. We expect a positive association between IQ and GPA. Moreover, some people think that self-concept is related to school performance. Examine in detail the relationships between GPA and the two explanatory variables IQ and self-concept. Are the relationships roughly linear? How strong are they? Are there unusual points? What is the effect of removing these points?  SEVENGR

## 2.4 Least-Squares Regression

When you complete this section, you will be able to

- Draw a straight line on a scatterplot of a set of data, given the equation of the line.
- Predict a value of the response variable  $y$  for a given value of the explanatory variable  $x$  using a regression equation.
- Explain the meaning of the term “least squares.”
- Calculate the equation of a least-squares regression line from the means and standard deviations of the explanatory and response variables and their correlation.
- Read the output of statistical software to find the equation of the least-squares regression line and the value of  $r^2$ .
- Explain the meaning of  $r^2$  in the regression setting.

Correlation measures the direction and strength of the linear (straight-line) relationship between two quantitative variables. If a scatterplot shows a linear relationship, we would like to summarize this overall pattern by drawing a line on the scatterplot. A *regression line* summarizes the relationship between two variables, but only in a specific setting: when one of the variables helps explain or predict the other. That is, regression describes a relationship between an explanatory variable and a response variable.

### REGRESSION LINE

A **regression line** is a straight line that describes how a response variable  $y$  changes as an explanatory variable  $x$  changes. We often use a regression line to **predict** the value of  $y$  for a given value of  $x$ . Regression, unlike correlation, requires that we have an explanatory variable and a response variable.

### Example

## 2.18 Fidgeting and fat gain.

Does fidgeting keep you slim? Some people don't gain weight even when they overeat. Perhaps fidgeting and other "nonexercise activity" (NEA) explains why—the body might spontaneously increase nonexercise activity when fed more. Researchers deliberately overfed 16 healthy young adults for 8 weeks. They measured fat gain (in kilograms) and, as an explanatory variable, increase in energy use (in calories) from activity other than deliberate exercise—fidgeting, daily living, and the like. Here are the data:<sup>18</sup>



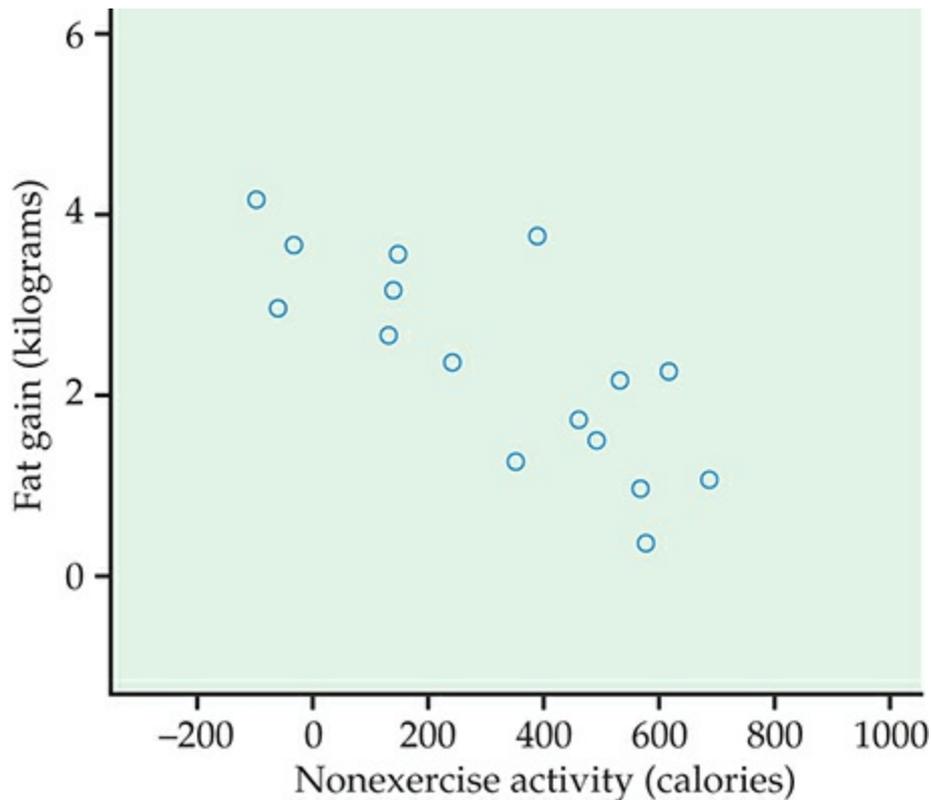
NEA increase (cal)	-94	-57	-29	135	143	151	245	355
Fat gain (kg)	4.2	3.0	3.7	2.7	3.2	3.6	2.4	1.3
NEA increase (cal)	392	473	486	535	571	580	620	690
Fat gain (kg)	3.8	1.7	1.6	2.2	1.0	0.4	2.3	1.1

Figure 2.16 is a scatterplot of these data. The plot shows a moderately strong negative linear association with no outliers. The correlation is  $r = -0.7786$ . People with larger increases in nonexercise activity do indeed gain less fat. A line drawn through the points will describe the overall pattern well.

### Fitting a line to data

When a scatterplot displays a linear pattern, we can describe the overall pattern by drawing a straight line through the points. Of course, no straight line passes exactly through all the points. **Fitting a line** to data means drawing a line that comes as close as possible to the points. The equation of a line fitted to the data gives a concise description of the relationship between the response variable  $y$  and the explanatory variable  $x$ . It is the numerical summary that supports the scatterplot, our graphical summary.

#### fitting a line



**FIGURE 2.16**

Fat gain after 8 weeks of overeating plotted against the increase in nonexercise activity over the same period, for Example 2.18.

## STRAIGHT LINES

Suppose that  $y$  is a response variable (plotted on the vertical axis) and  $x$  is an explanatory variable (plotted on the horizontal axis). A straight line relating  $y$  to  $x$  has an equation of the form

$$y = b_0 + b_1x$$

In this equation,  $b_1$  is the **slope**, the amount by which  $y$  changes when  $x$  increases by one unit. The number  $b_0$  is the **intercept**, the value of  $y$  when  $x = 0$ .

In practice, we will use software to obtain values of  $b_0$  and  $b_1$  for a given set of data.

## Example

### 2.19 Regression line for fat gain.

Any straight line describing the nonexercise activity data has the form

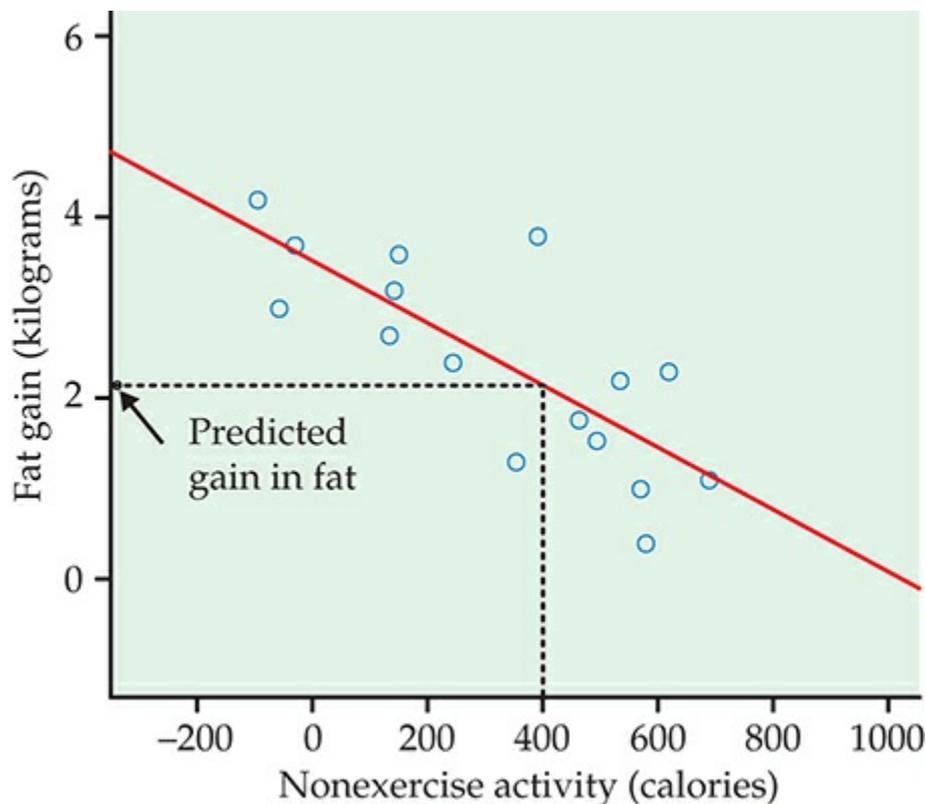
$$\text{fat gain} = b_0 + (b_1 \times \text{NEA increase})$$

In Figure 2.17 we have drawn the regression line with the equation

$$\text{fat gain} = 3.505 - (0.00344 \times \text{NEA increase})$$

The figure shows that this line fits the data well. The slope  $b_1 = -0.00344$  tells us that fat gained goes down by 0.00344 kilogram for each added calorie of NEA increase.

The slope  $b_1$  of a line  $y = b_0 + b_1x$  is the *rate of change* in the response  $y$  as the explanatory variable  $x$  changes. The slope of a regression line is an important numerical description of the relationship between the two variables. For Example 2.19, the intercept,  $b_0 = 3.505$  kilograms. This value is the estimated fat gain if NEA does not change. When we substitute the value zero for the NEA increase, the regression equation gives 3.505 (the intercept) as the predicted value of the fat gain.



## FIGURE 2.17

A regression line fitted to the nonexercise activity data and used to predict fat gain for an NEA increase of 400 calories, for Examples 2.19 and 2.20.

### USE YOUR KNOWLEDGE

#### 2.62 Plot the line.

Make a sketch of the data in Example 2.18 and plot the line

$$\text{fat gain} = 2.505 - (0.00344 \times \text{NEA increase})$$

on your sketch. Explain why this line does not give a good fit to the data.

### Prediction

We can use a regression line to *predict* the response  $y$  for a specific value of the explanatory variable  $x$ .

#### prediction

### Example

#### 2.20 Prediction for fat gain.

Based on the linear pattern, we want to predict the fat gain for an individual whose NEA increases by 400 calories when she overeats. To use the fitted line to predict fat gain, go “up and over” on the graph in Figure 2.17. From 400 calories on the  $x$  axis, go up to the fitted line and over to the  $y$  axis. The graph shows that the predicted gain in fat is a bit more than 2 kilograms.

If we have the equation of the line, it is faster and more accurate to substitute  $x = 400$  in the equation. The predicted fat gain is

$$\text{fat gain} = 3.505 - (0.00344 \times 400) = 2.13 \text{ kilograms}$$

The accuracy of predictions from a regression line depends on how much scatter about the line the data show. In Figure 2.17, fat gains for similar increases in NEA show a spread of 1 or 2 kilograms. The regression line summarizes the pattern but gives only roughly accurate predictions.

## USE YOUR KNOWLEDGE

### 2.63 Predict the fat gain.

Use the regression equation in Example 2.19 to predict the fat gain for a person whose NEA increases by 500 calories.

## Example

### 2.21 Is this prediction reasonable?

Can we predict the fat gain for someone whose nonexercise activity increases by 1500 calories when she overeats? We can certainly substitute 1500 calories into the equation of the line. The prediction is

$$\text{fat gain} = 3.505 - (0.00344 \times 1500) = -1.66 \text{ kilograms}$$

That is, we predict that this individual loses fat when she overeats. This prediction is not trustworthy. Look again at Figure 2.17. An NEA increase of 1500 calories is far outside the range of our data. We can't say whether increases this large ever occur, or whether the relationship remains linear at such extreme values. Predicting fat gain when NEA increases by 1500 calories *extrapolates* the relationship beyond what the data show.

## EXTRAPOLATION

**Extrapolation** is the use of a regression line for prediction far outside the range of values of the explanatory variable  $x$  used to obtain the line. Such predictions are often not accurate and should be avoided.

## USE YOUR KNOWLEDGE

### 2.64 Would you use the regression equation to predict?

Consider the following values for NEA increase:  $-400$ ,  $200$ ,  $500$ ,  $1000$ . For each, decide whether you would use the regression equation in Example 2.19 to predict fat gain or whether you would be concerned that the prediction would not be trustworthy because of extrapolation. Give reasons for your answers.

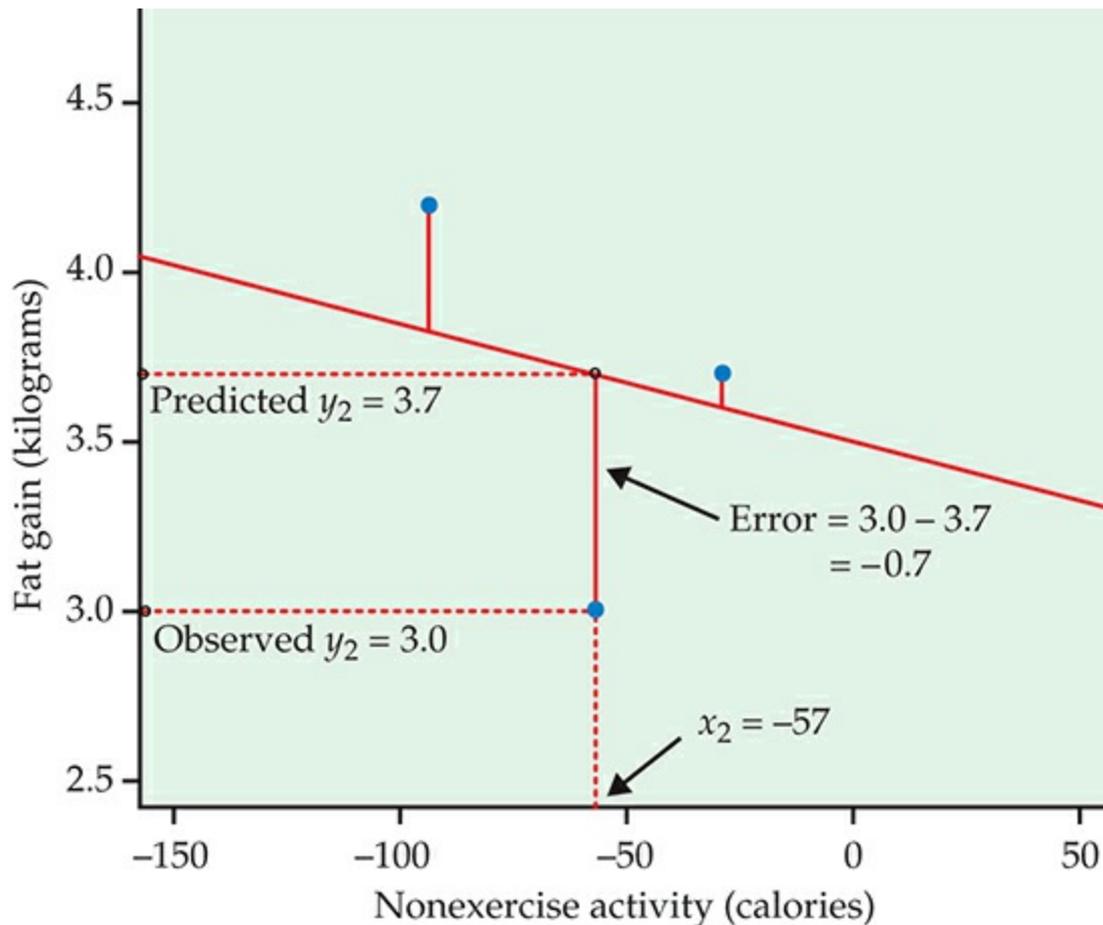
### Least-squares regression

Different people might draw different lines by eye on a scatterplot. This is especially true when the points are widely scattered. We need a way to draw a regression line that doesn't depend on our guess as to where the line should go. No line will pass exactly through all the points, but we want one that is as close as possible. We will use the line to predict  $y$  from  $x$ , so we want a line that is as close as possible to the points in the *vertical* direction. That's because the prediction errors we make are errors in  $y$ , which is the vertical direction in the scatterplot.

The line in Figure 2.17 predicts 2.13 kilograms of fat gain for an increase in nonexercise activity of 400 calories. If the actual fat gain turns out to be 2.3 kilograms, the error is

$$\begin{aligned}\text{error} &= \text{observed gain} - \text{predicted gain} \\ &= 2.3 - 2.13 = 0.17 \text{ kilograms}\end{aligned}$$

Errors are positive if the observed response lies above the line, and negative if the response lies below the line. We want a regression line that makes these prediction errors as small as possible. Figure 2.18 illustrates the idea. For clarity, the plot shows only three of the points from Figure 2.17, along with the line, on an expanded scale. The line passes below two of the points and above one of them. The vertical distances of the data points from the line appear as vertical line segments. A "good" regression line makes these distances as small as possible. There are many ways to make "as small as possible" precise. The most common is the *least-squares* idea. The line in Figures 2.17 and 2.18 is in fact the least-squares regression line.



**FIGURE 2.18**

The least-squares idea: make the errors in predicting  $y$  as small as possible by minimizing the sum of their squares.

## LEAST-SQUARES REGRESSION LINE

The **least-squares regression line of  $y$  on  $x$**  is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

Here is the least-squares idea expressed as a mathematical problem. We represent  $n$  observations on two variables  $x$  and  $y$  as

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

If we draw a line  $y = b_0 + b_1x$  through the scatterplot of these observations, the line predicts the value of  $y$  corresponding to  $x_i$  as  $\hat{y}_i = b_0 + b_1x_i$ . We write  $\hat{y}$  (read “y-hat”) in the equation of a regression line to emphasize that the line gives a *predicted* response  $\hat{y}$  for any  $x$ . The predicted response will usually not be exactly the same as the actually *observed* response  $y$ . The method of least squares chooses the line that makes the sum of the squares of these errors as small as possible. To

find this line, we must find the values of the intercept  $b_0$  and the slope  $b_1$  that minimize

$$\Sigma (\text{error})^2 = \Sigma (y_i - b_0 - b_1x_i)^2$$

for the given observations  $x_i$  and  $y_i$ . For the NEA data, for example, we must find the  $b_0$  and  $b_1$  that minimize

$$(4.2 - b_0 + 94b_1)^2 + (3.0 - b_0 + 57b_1)^2 + \dots + (1.1 - b_0 - 690b_1)^2$$

These values are the intercept and slope of the least-squares line.

You will use software or a calculator with a regression function to find the equation of the least-squares regression line from data on  $x$  and  $y$ . We will therefore give the equation of the least-squares line in a form that helps our understanding but is not efficient for calculation.

## EQUATION OF THE LEAST-SQUARES REGRESSION LINE

We have data on an explanatory variable  $x$  and a response variable  $y$  for  $n$  individuals. The means and standard deviations of the sample data are  $\bar{x}$  and  $s_x$  for  $x$  and  $\hat{y}$  and  $s_y$  for  $y$ , and the correlation between  $x$  and  $y$  is  $r$ . The **equation of the least-squares regression line** of  $y$  on  $x$  is

$$\hat{y} = b_0 + b_1x$$

with **slope**

$$b_1 = r \frac{s_y}{s_x}$$

and **intercept**

$$b_0 = \bar{y} - b_1\bar{x}$$

### Example

#### 2.22 Check the calculations.

Verify from the data in Example 2.18 that the mean and standard deviation of the 16 increases in NEA are

$$\bar{x}=324.8 \text{ calories} \quad \text{and} \quad s_x=257.66 \text{ calories}$$

The mean and standard deviation of the 16 fat gains are

$$\bar{y}=2.388 \text{ kg} \quad \text{and} \quad s_y=1.1389 \text{ kg}$$

The correlation between fat gain and NEA increase is  $r = -0.7786$ . The least-squares regression line of fat gain  $y$  on NEA increase  $x$  therefore has slope

$$\begin{aligned} b_1 &= r s_y / s_x = -0.7786(1.1389 / 257.66) \\ &= -0.00344 \text{ kg per calorie} \end{aligned}$$

and intercept

$$\begin{aligned} b_0 &= \bar{y} - b_1 \bar{x} \\ &= 2.388 - (-0.00344)(324.8) = 3.505 \text{ kg} \end{aligned}$$

The equation of the least-squares line is

$$\hat{y} = 3.505 - 0.00344x$$



*When doing calculations like this by hand, you may need to carry extra decimal places in the preliminary calculations to get accurate values of the slope and intercept. Using software or a calculator with a regression function eliminates this worry.*

## Interpreting the regression line

The slope  $b_1 = -0.00344$  kilograms per calorie in Example 2.22 is the change in fat gain as NEA increases. The units “kilograms of fat gained per calorie of NEA” come from the units of  $y$  (kilograms) and  $x$  (calories). Although the correlation does not change when we change the units of measurement, the equation of the least-squares line does change. The slope in grams per calorie would be 1000 times as large as the slope in kilograms per calorie, because there are 1000 grams in a kilogram. The small value of the slope,  $b_1 = -0.00344$ , does not mean that the effect of increased NEA on fat gain is small—it just reflects the choice of kilograms as the unit for fat gain. *The slope and intercept of the least-squares line depend on the units of measurement—you can’t conclude anything from their size.*

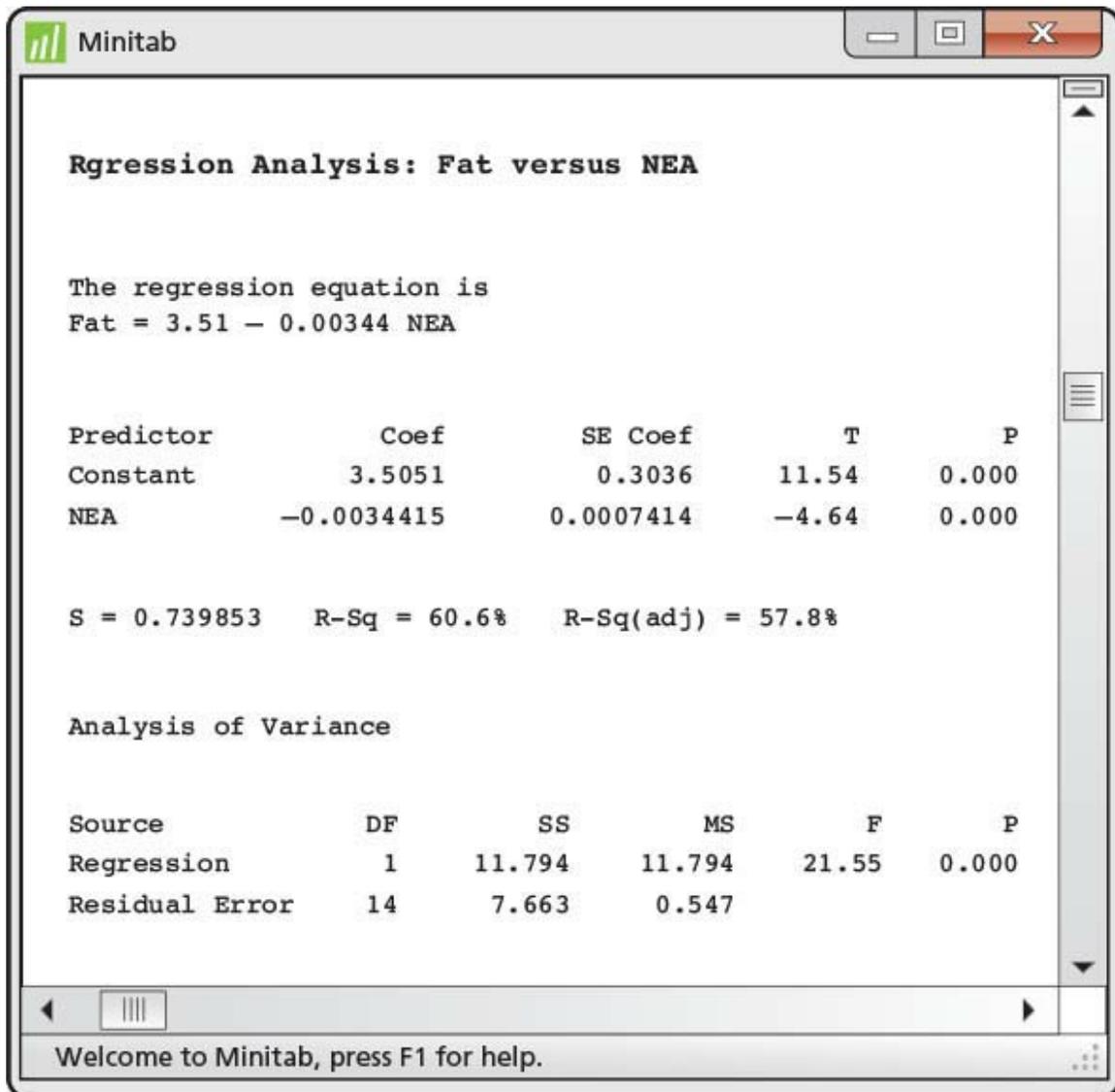


## Example

### 2.23 Regression using software.

Figure 2.19 displays the basic regression output for the nonexercise activity data from three statistical software packages. Other software produces very similar output. You can find the slope and intercept of the least-squares line, calculated to more decimal places than we need, in each output. The software also provides information that we do not yet need, including some that we trimmed from Figure 2.19.

Part of the art of using software is to ignore the extra information that is almost always present. Look for the results that you need. Once you understand a statistical method, you can read output from almost any software.



(a) Minitab

\*Output1 - IBM SPSS Statistics Viewer

→ Regression

[DataSet1]

Variables Entered/Removed<sup>a</sup>

Model	Variables Entered	Variables Removed	Method
1	NEA <sup>b</sup>		Enter

a. Dependent Variable: Fat  
b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.779 <sup>a</sup>	.606	.578	.7399

a. Predictors: (Constant), NEA

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11.794	1	11.794	21.546	.000 <sup>b</sup>
	Residual	7.663	14	.547		
	Total	19.458	15			

a. Dependent Variable: Fat  
b. Predictors: (Constant), NEA

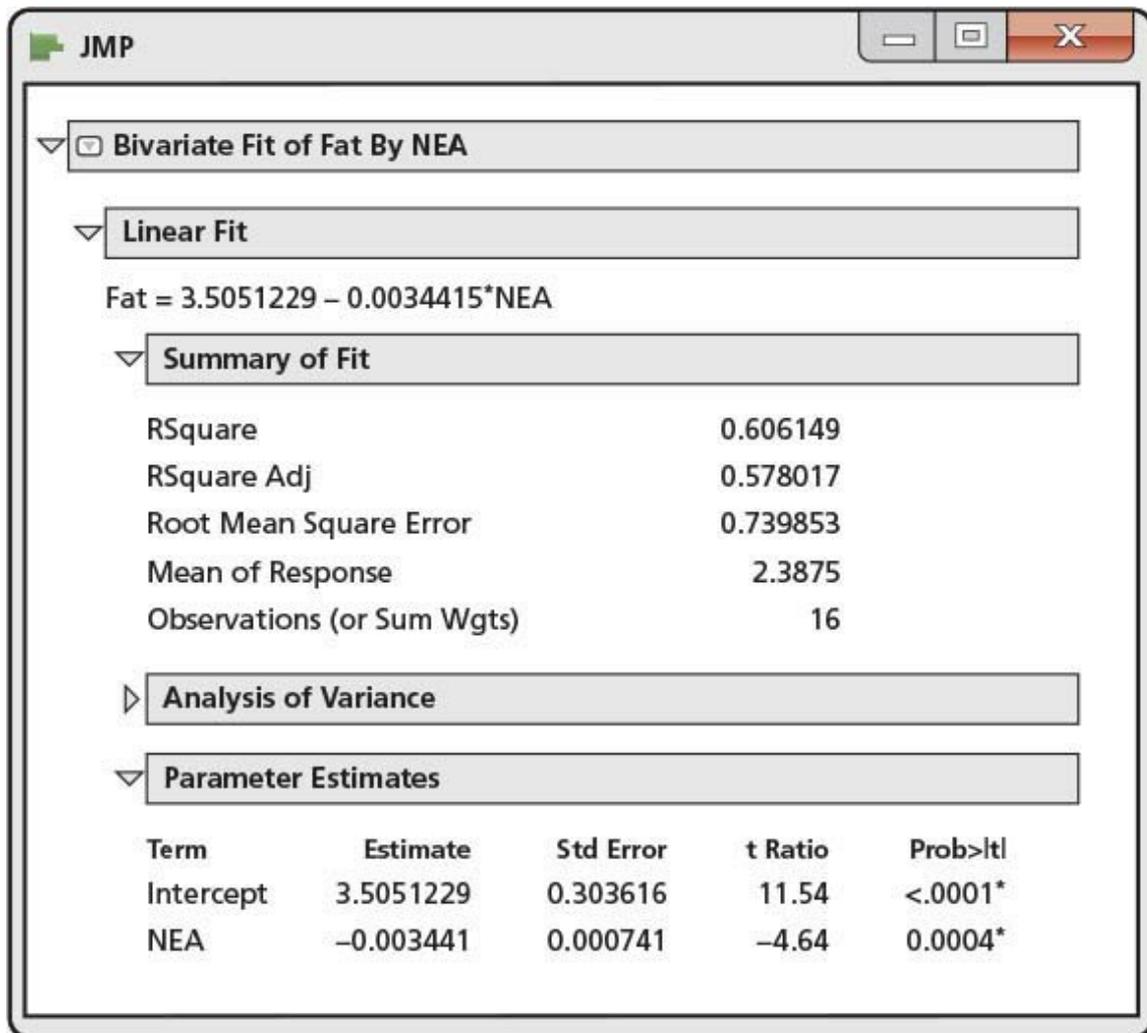
Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.505	.304		11.545	.000
	NEA	-.003	.001	-.779	-4.642	.000

a. Dependent Variable: Fat

IBM SPSS Statistics Processor is ready

(b) SPSS



(c) JMP

**FIGURE 2.19**

Regression results for the nonexercise activity data from three statistical software packages: (a) Minitab, (b) SPSS, and (c) JMP. Other software produces similar output.

## Facts about least-squares regression

Regression is one of the most common statistical settings, and least squares is the most common method for fitting a regression line to data. Here are some facts about least-squares regression lines.

**Fact 1.** There is a close connection between correlation and the slope of the least-squares line. The slope is

$$b_1 = r s_{y/x}$$

This equation says that along the regression line, **a change of one standard deviation in  $x$  corresponds to a change of  $r$  standard deviations in  $y$ .** When the variables are perfectly correlated ( $r = 1$  or  $r = -1$ ), the change in the predicted response  $\hat{y}$  is the same (in standard deviation units) as the change in  $x$ . Otherwise,

because  $-1 \leq r \leq 1$ , the change in  $\hat{y}$  is less than the change in  $x$ . As the correlation grows less strong, the prediction  $\hat{y}$  moves less in response to changes in  $x$ . Note that if the correlation is zero, then the slope of the least-squares regression line will be zero.

**Fact 2. The least-squares regression line always passes through the point  $(\bar{x}, \bar{y})$**  on the graph of  $y$  against  $x$ . So the least-squares regression line of  $y$  on  $x$  is the line with slope  $rs_y/s_x$  that passes through the point  $(\bar{x}, \bar{y})$ . We can describe regression entirely in terms of the basic descriptive measures  $\bar{x}$ ,  $s_x$ ,  $\bar{y}$ ,  $s_y$ , and  $r$ .

**Fact 3. The distinction between explanatory and response variables is essential in regression.** Least-squares regression looks at the distances of the data points from the line only in the  $y$  direction. If we reverse the roles of the two variables, we get a different least-squares regression line.

## Correlation and regression

Least-squares regression looks at the distances of the data points from the line only in the  $y$  direction. So the two variables  $x$  and  $y$  play different roles in regression.

### Example

#### 2.24 Laundry detergents.



#### LAUNDRY

Figure 2.20 is a scatterplot of the laundry detergent data described in Example 2.8 (page 87). There is a positive linear relationship. The two lines on the plot are the two least-squares regression lines. The regression line using price to predict rating is red. The regression line using rating to predict price is blue. *Regression of rating on price and regression of price on rating give different lines.* In the regression setting, you must decide which variable is explanatory.

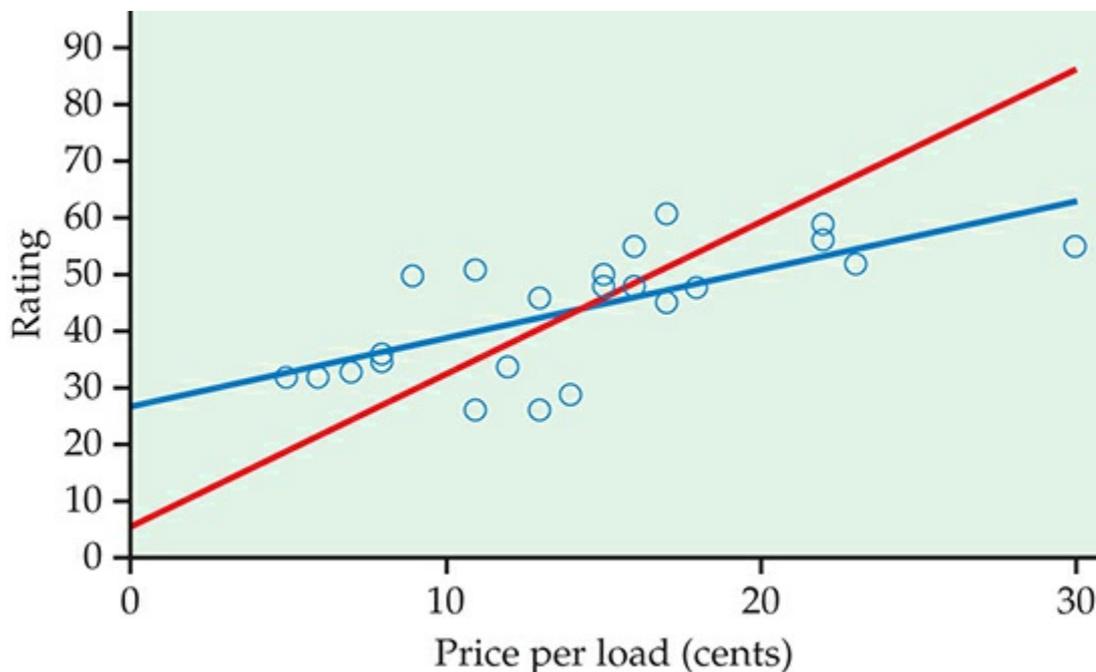
Even though the correlation  $r$  ignores the distinction between explanatory and response variables, there is a close connection between correlation and regression. We saw that the slope of the least-squares line involves  $r$ . Another connection between correlation and regression is even more important. In fact, the numerical value of  $r$  as a measure of the strength of a linear relationship is best interpreted by

thinking about regression. Here is the fact we need.

## $r^2$ IN REGRESSION

The **square of the correlation**,  $r^2$ , is the fraction of the variation in the values of  $y$  that is explained by the least-squares regression of  $y$  on  $x$ .

The correlation between NEA increase and fat gain for the 16 subjects in Example 2.18 (page 110) is  $r = -0.7786$ . Because  $r^2 = 0.606$ , the straight-line relationship between NEA and fat gain explains about 61% of the vertical scatter in fat gains in Figure 2.17 (page 112).



**FIGURE 2.20**

Scatterplot of price per load versus rating for 24 laundry detergents, from Example 2.8. The two lines are the two least-squares regression lines: using price per load to predict rating (red) and using rating to predict price per load (blue), for Example 2.24.

When you report a regression, give  $r^2$  as a measure of how successfully the regression explains the response. All three software outputs in Figure 2.19 include  $r^2$ , either in decimal form or as a percent.

When you see a correlation, square it to get a better feel for the strength of the association. Perfect correlation ( $r = -1$  or  $r = 1$ ) means that the points lie exactly on a line. Then  $r^2 = 1$  and all the variation in one variable is accounted for by the linear relationship with the other variable. If  $r = -0.7$  or  $r^2 = 0.7$ ,  $r^2 = 0.49$  and about half the variation is accounted for by the linear relationship. In the  $r^2$  scale, correlation  $\pm 0.7$  is about halfway between 0 and  $\pm 1$ .

## USE YOUR KNOWLEDGE

### 2.65 What fraction of the variation is explained?

Consider the following correlations:  $-0.8$ ,  $-0.4$ ,  $-0.2$ ,  $0$ ,  $0.3$ ,  $0.5$ , and  $0.9$ . For each give the fraction of the variation in  $y$  that is explained by the least-squares regression of  $y$  on  $x$ . Summarize what you have found from performing these calculations.

The use of  $r^2$  to describe the success of regression in explaining the response  $y$  is very common. It rests on the fact that there are two sources of variation in the responses  $y$  in a regression setting. Figure 2.17 (page 112) gives a rough visual picture of the two sources. The first reason for the variation in fat gains is that there is a relationship between fat gain  $y$  and increase in NEA  $x$ . As  $x$  increases from  $-94$  to  $690$  calories among the 16 subjects, it pulls fat gain  $y$  with it along the regression line in the figure. The linear relationship explains this part of the variation in fat gains.

The fat gains do not lie exactly on the line, however, but are scattered above and below it. This is the second source of variation in  $y$ , and the regression line tells us nothing about how large it is. The dashed lines in Figure 2.17 show a rough average for  $y$  when we fix a value of  $x$ . We use  $r^2$  to measure variation along the line as a fraction of the total variation in the fat gains. In Figure 2.17, about 61% of the variation in fat gains among the 16 subjects is due to the straight-line relationship between  $y$  and  $x$ . The remaining 39% is vertical scatter in the observed responses remaining after the line has fixed the predicted responses.

### Another view of $r^2$

Here is a more specific interpretation of  $r^2$ . The fat gains  $y$  in Figure 2.17 range from  $0.4$  to  $4.2$  kilograms. The variance of these responses, a measure of how variable they are, is

$$\text{variance of observed values } y = 1.297$$

Much of this variability is due to the fact that as  $x$  increases from  $-94$  to  $690$  calories it pulls  $y$  along with it. If the only variability in the observed responses were due to the straight-line dependence of fat gain on NEA, the observed gains would lie exactly on the regression line. That is, they would be the same as the predicted gains  $\hat{y}$ . We can compute the predicted gains by substituting the NEA values for each subject into the equation of the least-squares line. Their variance describes the variability in the predicted responses. The result is

variance of predicted values  $\hat{y} = 0.786$

This is what the variance would be if the responses fell exactly on the line, that is, if the linear relationship explained 100% of the observed variation in  $y$ . Because the responses don't fall exactly on the line, the variance of the predicted values is smaller than the variance of the observed values. Here is the fact we need:

$$r^2 = \frac{\text{variance of predicted values } \hat{y}}{\text{variance of observed values } y} \\ = \frac{0.786}{1.297} = 0.606$$

This fact is always true. The squared correlation gives the variance that the responses would have if there were no scatter about the least-squares line as a fraction of the variance of the actual responses. This is the exact meaning of "fraction of variation explained" as an interpretation of  $r^2$ .

These connections with correlation are special properties of least-squares regression. They are not true for other methods of fitting a line to data. One reason that least squares is the most common method for fitting a regression line to data is that it has many convenient special properties.

## SECTION 2.4 Summary

A **regression line** is a straight line that describes how a response variable  $y$  changes as an explanatory variable  $x$  changes.

The most common method of fitting a line to a scatterplot is least squares. The **least-squares regression line** is the straight line  $\hat{y} = b_0 + b_1x$  that minimizes the sum of the squares of the vertical distances of the observed  $y$ -values from the line.

You can use a regression line to **predict** the value of  $y$  for any value of  $x$  by substituting this  $x$  into the equation of the line. **Extrapolation** beyond the range of  $x$ -values spanned by the data is risky.

The **slope**  $b_1$  of a regression line  $\hat{y} = b_0 + b_1x$  is the rate at which the predicted response  $\hat{y}$  changes along the line as the explanatory variable  $x$  changes. Specifically,  $b_1$  is the change in  $\hat{y}$  when  $x$  increases by 1. The numerical value of the slope depends on the units used to measure  $x$  and  $y$ .

The **intercept**  $b_0$  of a regression line  $\hat{y} = b_0 + b_1x$  is the predicted response  $\hat{y}$  when the explanatory variable  $x = 0$ . This prediction is not particularly useful unless  $x$  can actually take values near 0.

The least-squares regression line of  $y$  on  $x$  is the line with slope  $b_1 = r s_y / s_x$  and intercept  $b_0 = \bar{y} - b_1 \bar{x}$ . This line always passes through the point  $(\bar{x}, \bar{y})$ .

**Correlation and regression** are closely connected. The correlation  $r$  is the slope of the least-squares regression line when we measure both  $x$  and  $y$  in standardized units. The square of the correlation  $r^2$  is the fraction of the variance of one variable that is explained by least-squares regression on the other variable.

## SECTION 2.4 Exercises

For Exercise 2.62, see page 112; for Exercise 2.63, see page 113; for Exercise 2.64, see page 113; and for Exercise 2.65, see page 120.

## 2.66 Bone strength.

Exercise 2.18 (page 98) gives the bone strengths of the dominant and the nondominant arms for 15 men who were controls in a study.  ARMSTR

(a) Plot the data. Use the bone strength in the nondominant arm as the explanatory variable and bone strength in the dominant arm as the response variable.

(b) The least-squares regression line for these data is

$$\text{dominant} = 2.74 + (0.936 \times \text{nondominant})$$

Add this line to your plot.

(c) Use the scatterplot (a graphical summary) with the least-squares line (a graphical display of a numerical summary) to write a short paragraph describing this relationship.

## 2.67 Bone strength for baseball players.

Refer to the previous exercise. Similar data for baseball players is given in Exercise 2.19 (page 98). Here is the equation of the least-squares line for the baseball players:

$$\text{dominant} = 0.886 + (1.373 \times \text{nondominant})$$

Answer parts (a) and (c) of the previous exercise for these data.  ARMSTR

## 2.68 Predict the bone strength.

Refer to Exercise 2.66. A young male who is not a baseball player has a bone strength of  $16.0 \text{ cm}^4/1000$  in his nondominant arm. Predict the bone strength in the dominant arm for this person.  ARMSTR

## 2.69 Predict the bone strength for a baseball player.

Refer to Exercise 2.67. A young male who is a baseball player has a bone strength of  $16.0 \text{ cm}^4/1000$  in his nondominant arm. Predict the bone strength in the dominant arm for this person.  ARMSTR

## 2.70 Compare the predictions.

Refer to the two previous exercises. You have predicted two dominant-arm bone strengths, one for a baseball player and one for a person who is not a baseball player. The nondominant bone strengths are both  $16.0 \text{ cm}^4/1000$ .  ARMSTR

(a) Compare the two predictions, baseball player minus control.

(b) Explain how the difference in the two predictions is an estimate of the effect of baseball-throwing exercise on the strength of arm bones.

- (c) For nondominant-arm strengths of 12 and 20 cm<sup>4</sup>/1000, repeat your predictions and take the differences. Make a table of the results of all three calculations (for 12, 16, and 20 cm<sup>4</sup>/1000).
- (d) Write a short summary of the results of your calculations for the three different nondominant-arm strengths. Be sure to include an explanation of why the differences are not the same for the three nondominant-arm strengths.

## 2.71 Least-squares regression for radioactive decay.

Refer to Exercise 2.22 (page 99) for the data on radioactive decay of barium-137m. Here are the data:  **DECAY**

Time	1	3	5	7
Count	578	317	203	118

- (a) Using the least-squares regression equation

$$\text{count} = 602.8 - (74.7 \times \text{time})$$

find the predicted values for the counts.

- (b) Compute the differences, observed count minus predicted count. How many of these are positive; how many are negative?
- (c) Square and sum the differences that you found in part (b).
- (d) Repeat the calculations that you performed in parts (a) to (c) using the equation

$$\text{count} = 500 - (100 \times \text{time})$$

- (e) In a short paragraph, explain the least-squares idea using the calculations that you performed in this exercise.

## 2.72 Least-squares regression for the log counts.

Refer to Exercise 2.23 (page 99), where you analyzed the radioactive decay of barium-137m data using log counts. Here are the data:  **DECAY**

Time	1	3	5	7
Log count	6.35957	5.75890	5.31321	4.77068

- (a) Using the least-squares regression equation

$$\log \text{count} = 6.593 - (0.2606 \times \text{time})$$

find the predicted values for the log counts.

- (b) Compute the differences, observed count minus predicted count. How many of these are positive; how many are negative?
- (c) Square and sum the differences that you found in part (b).
- (d) Repeat the calculations that you performed in parts (a) to (c) using the equation

$$\log \text{ count} = 7 - (0.2 \times \text{time})$$

(e) In a short paragraph, explain the least-squares idea using the calculations that you performed in this exercise.

### 2.73 College students by state.

Refer to Exercise 2.21 (page 99) and Figure 2.11 (page 100), where you examined the relationship between the number of undergraduate college students and the populations for the 50 states. In Exercise 2.45 (page 107) you calculated the correlation between these two variables. Here are some numerical summaries for these variables: for number of undergraduate college students, the mean is 302,136 and the standard deviation is 358,460; for population, the mean is 5,955,551 and the standard deviation is 6,620,733. The correlation between the number of undergraduate college students and the population is 0.98367. Use this information to find the least-squares regression line. Show your work.

### 2.74 College students by state without the four largest states.

Refer to the previous exercise. Let's eliminate the four largest states, which have populations greater than 15 million. Here are the numerical summaries: for number of undergraduate college students, the mean is 220,134 and the standard deviation is 165,270; for population, the mean is 4,367,448 and the standard deviation is 3,310,957. The correlation between the number of undergraduate college students and the population is 0.97081. Use this information to find the least-squares regression line. Show your work.

### 2.75 Make predictions and compare.

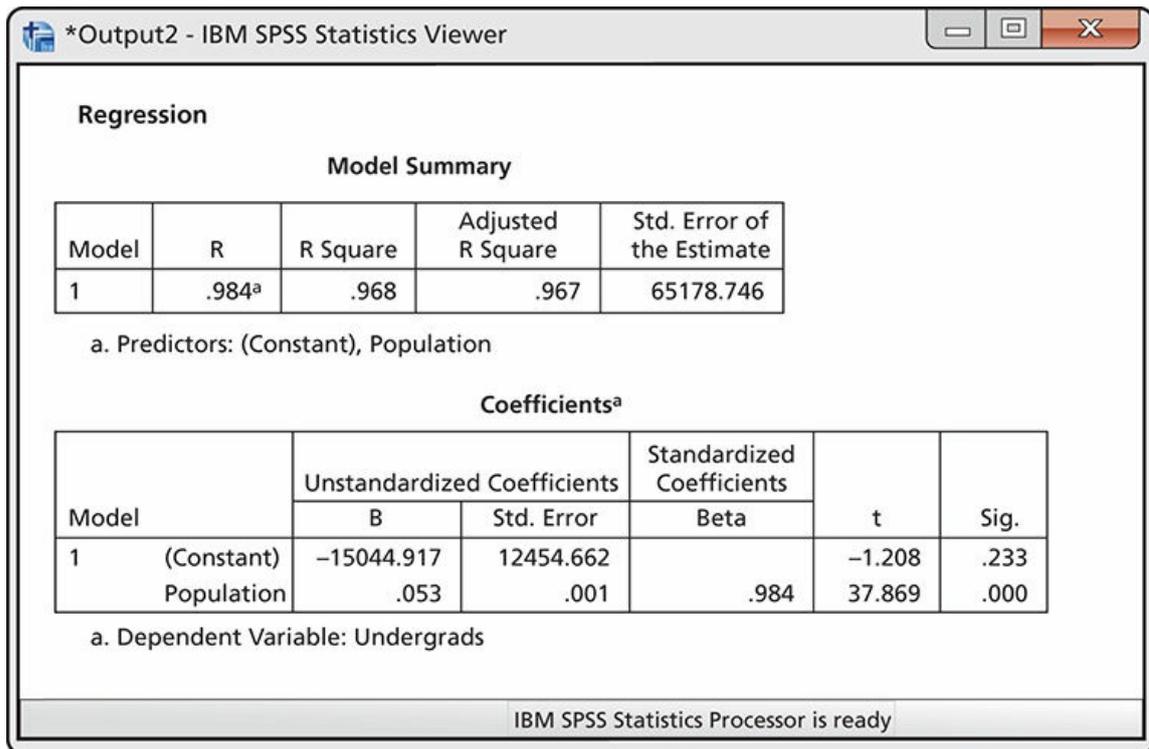
Refer to the two previous exercises. Consider a state with a population of 6 million (this value is approximately the median population for the 50 states).

- Using the least-squares regression equation for all 50 states, find the predicted number of undergraduate college students.
- Do the same using the least-squares regression equation for the 46 states with population less than 15 million.
- Compare the predictions that you made in parts (a) and (b). Write a short summary of your results and conclusions. Pay particular attention to the effect of including the four states with the largest populations in the prediction equation for a median-sized state.

### 2.76 College students by state.

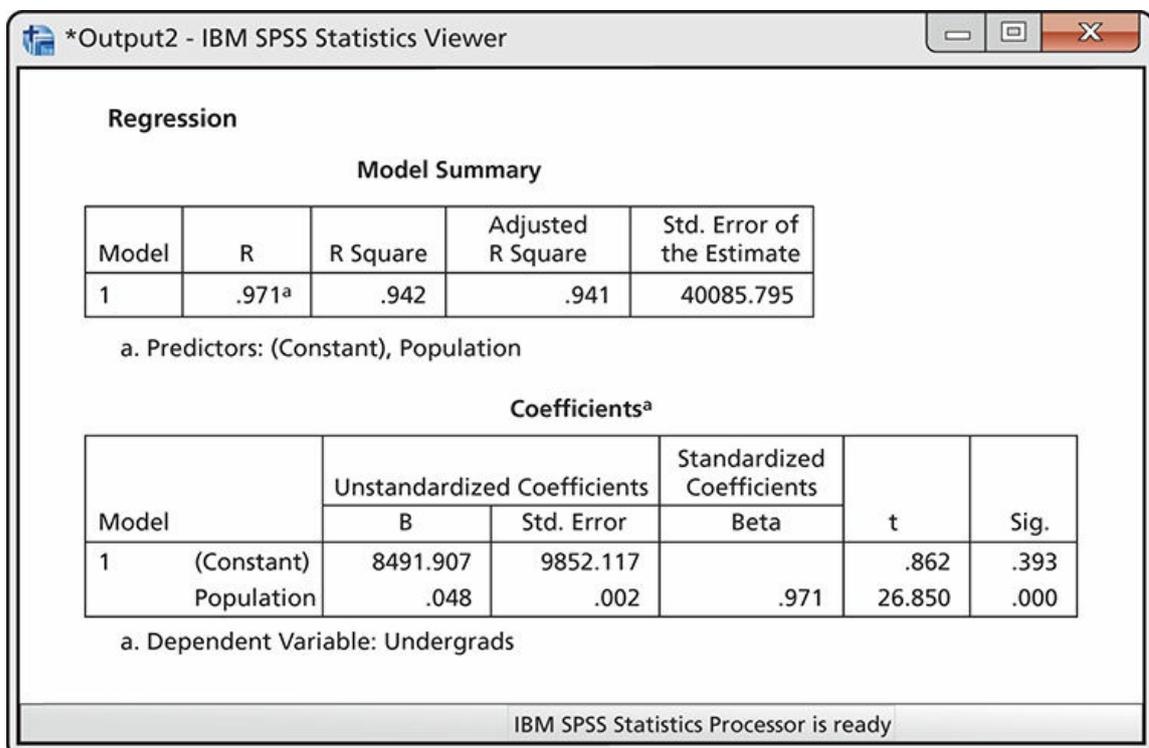
Refer to Exercise 2.21 (page 99), where you examined the relationship between the number of undergraduate college students and the populations for the 50 states. Figure 2.22 gives the output from a software package for the regression. Use this output to answer the following questions:  COLLEGE

- What is the equation of the least-squares regression line?
- What is the value of  $r^2$ ?
- Interpret the value of  $r^2$ .
- Does the software output tell you that the relationship is linear and not, for example, curved? Explain your answer.



**FIGURE 2.21**

SPSS output for predicting number of undergraduate college students using population for the 50 states, for Exercise 2.73.



**FIGURE 2.22**

SPSS output for predicting number of undergraduate college students using population, with the four largest states deleted, for Exercise 2.76.

## 2.77 College students by state without the four largest states.

Refer to Exercise 2.74 where you eliminated the four largest states that have populations greater than 15 million. Answer the questions in the previous exercise for the data set with the 46 states.  COL46

## 2.78 Data generated by software.

The following 20 observations on  $Y$  and  $X$  were generated by a computer program.  GENDATA

$Y$	$X$	$Y$	$X$
34.38	22.06	27.07	17.75
30.38	19.88	31.17	19.96
26.13	18.83	27.74	17.87
31.85	22.09	30.01	20.20
26.77	17.19	29.61	20.65
29.00	20.72	31.78	20.32
28.92	18.10	32.93	21.37
26.30	18.01	30.29	17.31
29.49	18.69	28.57	23.50
31.36	18.05	29.80	22.02

- Make a scatterplot and describe the relationship between  $Y$  and  $X$ .
- Find the equation of the least-squares regression line and add the line to your plot.
- What percent of the variability in  $Y$  is explained by  $X$ ?
- Summarize your analysis of these data in a short paragraph.

## 2.79 Alcohol and carbohydrates in beer.

Figure 2.10 (page 100) gives a scatterplot of carbohydrates versus percent alcohol in 153 brands of beer.  BEER

- Find the equation of the least-squares regression line for these data.
- Find the value of  $r^2$  and interpret it in the regression context.
- Write a short report on the relationship between carbohydrates and percent alcohol in beer. Include graphical and numerical summaries for each variable separately as well as graphical and numerical summaries for the relationship in your report.

## 2.80 Alcohol and carbohydrates in beer revisited.

Refer to the previous exercise. The data that you used includes an outlier.  BEER

- Remove the outlier and answer parts (a) through (c) for the new set of data.
- Write a short paragraph about the possible effects of outliers on a least-squares regression line and the

value of  $r^2$ , using this example to illustrate your ideas.

## 2.81 Always plot your data!

Table 2.2 presents four sets of data prepared by the statistician Frank Anscombe to illustrate the dangers of calculating without first plotting the data.<sup>19</sup>  ANSC

TABLE 2.2		Four Data Sets for Exploring Correlation and Regression										
<b>Data</b>	<b>Set A</b>											
$x$	10	8	13	9	11	14	6	4	12	7	5	
$y$	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68	
<b>Data</b>	<b>Set B</b>											
$x$	10	8	13	9	11	14	6	4	12	7	5	
$y$	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74	
<b>Data</b>	<b>Set C</b>											
$x$	10	8	13	9	11	14	6	4	12	7	5	
$y$	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73	
<b>Data</b>	<b>Set D</b>											
$x$	8	8	8	8	8	8	8	8	8	8	19	
$y$	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50	

- Without making scatterplots, find the correlation and the least-squares regression line for all four data sets. What do you notice? Use the regression line to predict  $y$  for  $x = 10$ .
- Make a scatterplot for each of the data sets and add the regression line to each plot.
- In which of the four cases would you be willing to use the regression line to describe the dependence of  $y$  on  $x$ ? Explain your answer in each case.

## 2.82 Add an outlier.

Refer to Exercise 2.78. Add an additional observation with  $Y = 44$  and  $X = 40$  to the data set. Repeat the analysis that you performed in Exercise 2.78 and summarize your results paying particular attention to the effect of this outlier.  GEN21A

## 2.83 Add a different outlier.

Refer to Exercise 2.78 and the previous exercise. Add an additional observation with  $Y = 30$  and  $X = 40$  to the original data set.  GEN21B

- Repeat the analysis that you performed in Exercise 2.78 and summarize your results paying particular attention to the effect of this outlier.

(b) In this exercise and in the previous one, you added an outlier to the original data set and reanalyzed the data. Write a short summary of the changes in correlations that can result from different kinds of outliers.

## 2.84 Progress in math scores.

Every few years, the National Assessment of Educational Progress asks a national sample of eighth-graders to perform the same math tasks. The goal is to get an honest picture of progress in math. Here are the last few national mean scores, on a scale of 0 to 500:<sup>20</sup>  NAEP

Year	1990	1992	1996	2000	2003	2005	2008	2011
Score	263	268	272	273	278	279	281	283

(a) Make a time plot of the mean scores, by hand. This is just a scatterplot of score against year. There is a slow linear increasing trend.

(b) Find the regression line of mean score on time step-by-step. First calculate the mean and standard deviation of each variable and their correlation (use a calculator with these functions). Then find the equation of the least-squares line from these. Draw the line on your scatterplot. What percent of the year-to-year variation in scores is explained by the linear trend?

(c) Now use software or the regression function on your calculator to verify your regression line.

## 2.85 The regression equation.

The equation of a least-squares regression line is  $y = 12 + 8x$ .

(a) What is the value of  $y$  for  $x = 3$ ?

(b) If  $x$  increases by one unit, what is the corresponding increase in  $y$ ?

(c) What is the intercept for this equation?

## 2.86 Metabolic rate and lean body mass.

Compute the mean and the standard deviation of the metabolic rates and lean body masses in Exercise 2.35 (page 101) and the correlation between these two variables. Use these values to find the slope of the regression line of metabolic rate on lean body mass. Also find the slope of the regression line of lean body mass on metabolic rate. What are the units for each of the two slopes?  BMASS

## 2.87 IQ and self-concept.

Table 1.3 (page 29) reports data on 78 seventh-grade students. We want to know how well each of IQ score and self-concept score predicts GPA using least-squares regression. We also want to know which of these explanatory variables predicts GPA better. Give numerical measures that answer these questions, and explain your answers.  SEVENGR

## 2.88 Use an applet for progress in math scores.

Go to the *Two-Variable Statistical Calculator* applet. Enter the data for the progress in math scores from Exercise 2.84 using the “User-entered data” option in the “Data” tab. Explore the data by clicking the other

tabs in the applet. Using only the results provided by the applet, write a short report summarizing the analysis of these data.

### **2.89 A property of the least-squares regression line.**

Use the equation for the least-squares regression line to show that this line always passes through the point  $(\bar{x}, \bar{y})$ .

### **2.90 Class attendance and grades.**

A study of class attendance and grades among first-year students at a state university showed that in general students who attended a higher percent of their classes earned higher grades. Class attendance explained 16% of the variation in grade index among the students. What is the numerical value of the correlation between percent of classes attended and grade index?

### **2.91 Revenue and value of NFL teams.**

In Exercises 2.36 and 2.54, you used scatterplots and correlations to examine the prediction of team value for 32 NFL teams using three different predictors. Now, find the least-squares regression equations. Write a short report summarizing your findings. Include plots, correlations, and the least-squares regression lines and a summary of your conclusions.  NFL

## 2.5 Cautions about Correlation and Regression

When you complete this section, you will be able to

- Calculate the residuals for a set of data using the equation of the least-squares regression line and the observed values of the explanatory variable.
- Use a plot of the residuals versus the explanatory variable to assess the fit of a regression line.
- Identify outliers and influential observations by examining scatterplots and residual plots.
- Identify lurking variables that can influence the interpretation of relationships between two variables.
- Explain the difference between association and causality when interpreting the relationship between two variables.

Correlation and regression are among the most common statistical tools. They are used in more elaborate form to study relationships among many variables, a situation in which we cannot see the essentials by studying a single scatterplot. We need a firm grasp of the use and limitations of these tools, both now and as a foundation for more advanced statistics.

### Residuals

A regression line describes the overall pattern of a linear relationship between an explanatory variable and a response variable. Deviations from the overall pattern are also important. In the regression setting, we see deviations by looking at the scatter of the data points about the regression line. The vertical distances from the points to the least-squares regression line are as small as possible in the sense that they have the smallest possible sum of squares. Because they represent “leftover” variation in the response after fitting the regression line, these distances are called *residuals*.

#### RESIDUALS

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. That is,

$$\begin{aligned}\text{residual} &= \text{observed } y - \text{predicted } y \\ &= y - \hat{y}\end{aligned}$$

## Example

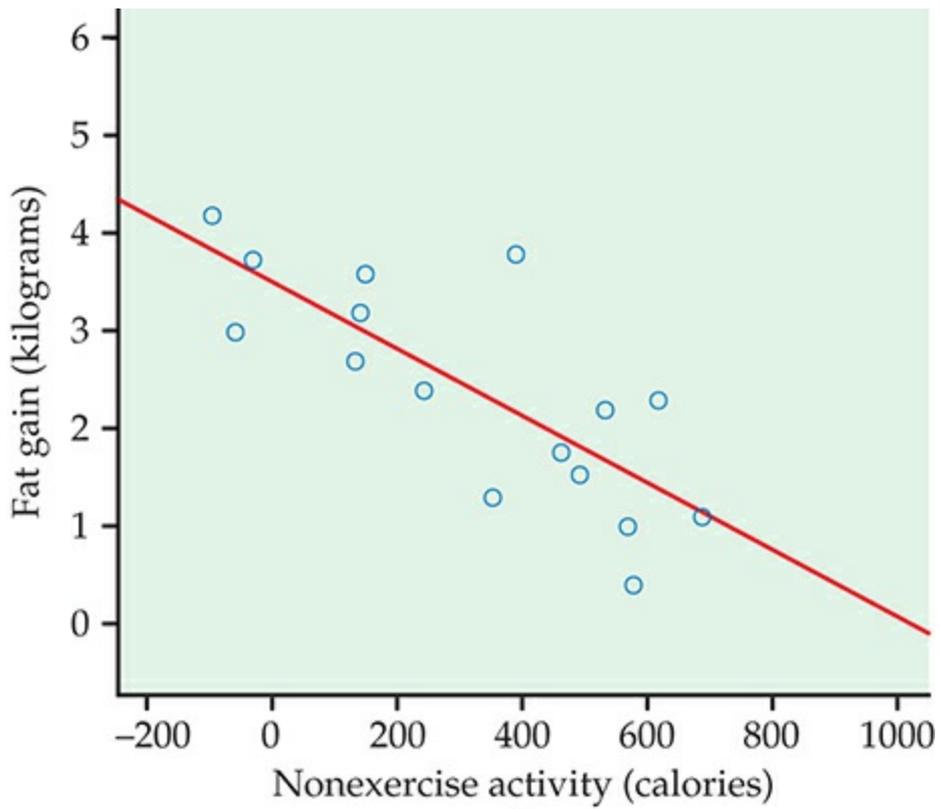
### 2.25 Residuals for fat gain.

Example 2.18 (page 110) describes measurements on 16 young people who volunteered to overeat for 8 weeks. Those whose nonexercise activity (NEA) spontaneously rose substantially gained less fat than others. Figure 2.23(a) is a scatterplot of these data. The pattern is linear. The least-squares line is

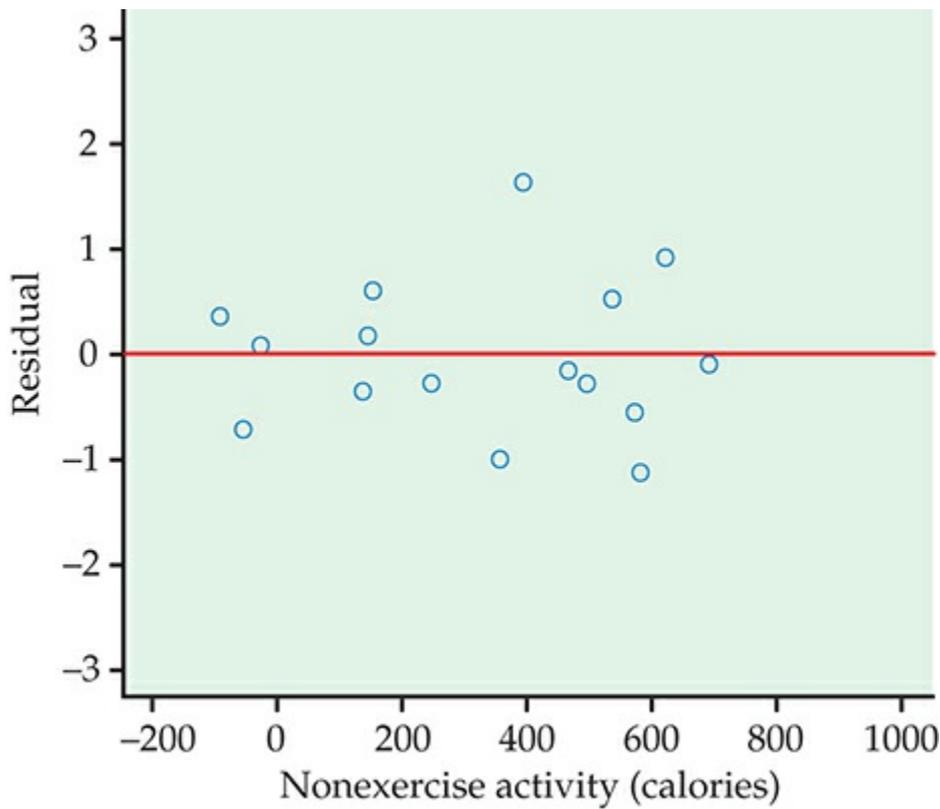
$$\text{fat gain} = 3.505 - (0.00344 \times \text{NEA increase})$$

One subject's NEA rose by 135 calories. That subject gained 2.7 kilograms of fat. The predicted gain for 135 calories is

$$\hat{y} = 3.505 - (0.00344 \times 135) = 3.04 \text{ kg}$$



(a)



(b)

**FIGURE 2.23**

(a) Scatterplot of fat gain versus increase in nonexercise activity, with the least-squares regression line, for Example 2.25. (b) Residual plot for the regression displayed in (a). The

line at  $y = 0$  marks the mean of the residuals.

The residual for this subject is therefore

$$\begin{aligned}\text{residual} &= \text{observed } y - \text{predicted } y \\ &= y - \hat{y} \\ &= 2.7 - 3.04 = -0.34 \text{ kg}\end{aligned}$$

Most regression software will calculate and store residuals for you.

## USE YOUR KNOWLEDGE

### 2.92 Find the predicted value and the residual.

Let's say that we have an individual in the NEA data set who has NEA increase equal to 144 calories and fat gain equal to 3.1 kg. Find the predicted value of fat gain for this individual and then calculate the residual. Explain why this residual is positive.

Because the residuals show how far the data fall from our regression line, examining the residuals helps us assess how well the line describes the data. Although residuals can be calculated from any model fitted to the data, the residuals from the least-squares line have a special property: **the mean of the least-squares residuals is always zero.**

## USE YOUR KNOWLEDGE

### 2.93 Find the sum of the residuals.

Here are the 16 residuals for the NEA data rounded to two decimal places:



0.37	-0.70	0.10	-0.34	0.19	0.61	-0.26	-0.98
1.64	-0.18	-0.23	0.54	-0.54	-1.11	0.93	-0.03

Find the sum of these residuals. Note that the sum is not exactly zero because of roundoff error.

You can see the residuals in the scatterplot of Figure 2.23(a) by looking at the vertical deviations of the points from the line. The *residual plot* in Figure 2.23(b) makes it easier to study the residuals by plotting them against the explanatory variable, increase in NEA.

## RESIDUAL PLOTS

A **residual plot** is a scatterplot of the regression residuals against the explanatory variable. Residual plots help us assess the fit of a regression line.

Because the mean of the residuals is always zero, the horizontal line at zero in Figure 2.23(b) helps orient us. This line (residual = 0) corresponds to the fitted line in Figure 2.23(a). The residual plot magnifies the deviations from the line to make patterns easier to see. If the regression line catches the overall pattern of the data, there should be *no pattern* in the residuals. That is, the residual plot should show an unstructured horizontal band centered at zero. The residuals in Figure 2.23(b) do have this irregular scatter.

You can see the same thing in the scatterplot of Figure 2.23(a) and the residual plot of Figure 2.23(b). It's just a bit easier in the residual plot. Deviations from an irregular horizontal pattern point out ways in which the regression line fails to catch the overall pattern. Here is an example.

### Example

#### 2.26 Patterns in birthrate and Internet user residuals.

In Exercise 2.28 (page 100) we used a scatterplot to study the relationship between birthrate and Internet users for 106 countries. In this scatterplot, Figure 2.11, we see that there are many countries with low numbers of Internet users. In addition, the relationship between births and Internet users appears to be curved. For low values of Internet users, there is a clear relationship, while

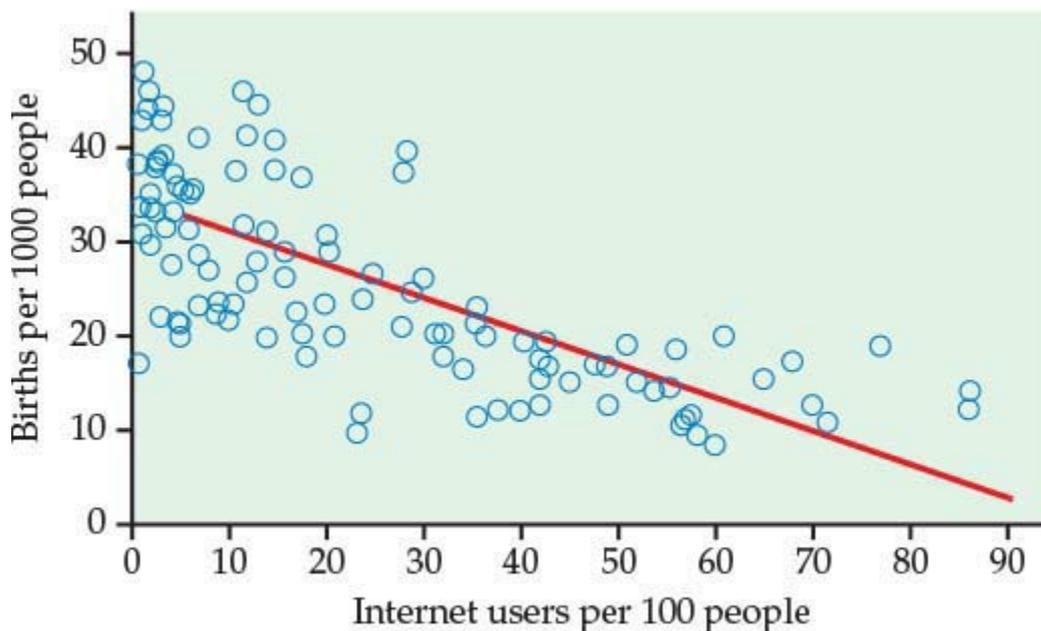
for higher values, the curve becomes relatively flat.

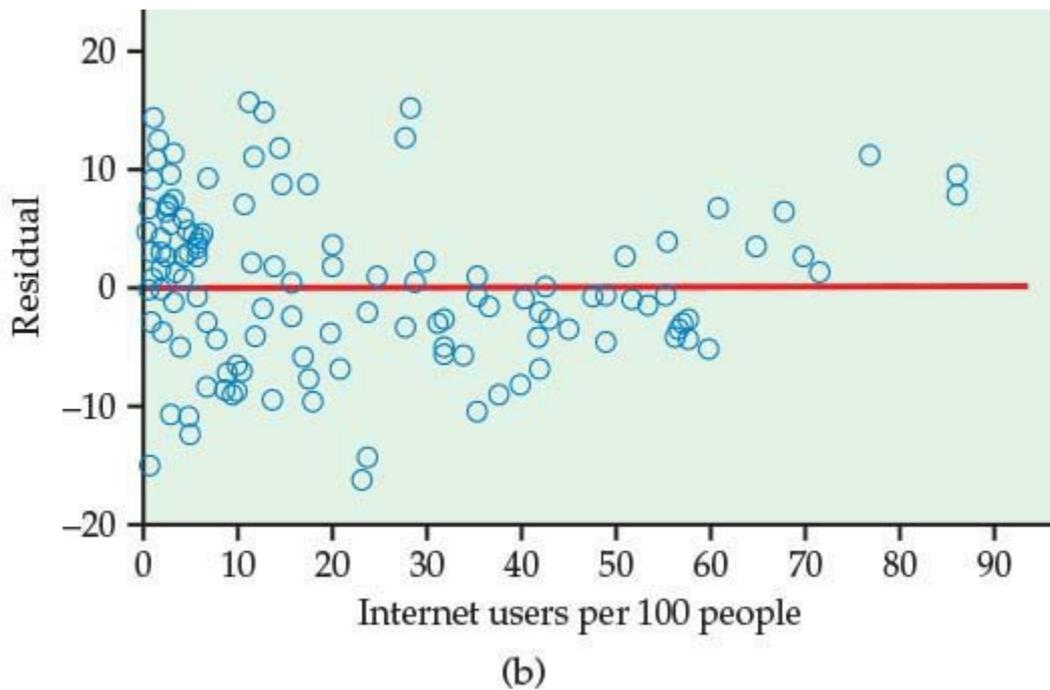


### INBIRTH

Figure 2.24(a) gives the data with the least-squares regression line, and Figure 2.24(b) plots the residuals. Look at the right part of Figure 2.24(b), where the values of Internet users are high. Here we see that the residuals tend to be positive.

The residual pattern in Figure 2.24(b) is characteristic of a simple curved relationship. *There are many ways in which a relationship can deviate from a linear pattern.* We now have an important tool for examining these deviations. Use it frequently and carefully when you study relationships.





**FIGURE 2.24**

(a) Scatterplot of birthrate versus Internet users, with the least-squares regression line, for Example 2.26. (b) Residual plot for the regression displayed in (a). The line at  $y = 0$  marks the mean of the residuals.

## Outliers and influential observations

When you look at scatterplots and residual plots, look for striking individual points as well as for an overall pattern. Here is an example of data that contain some unusual cases.

### Example

#### 2.27 Diabetes and blood sugar.

People with diabetes must manage their blood sugar levels carefully. They measure their fasting plasma glucose (FPG) several times a day with a glucose meter. Another measurement, made at regular medical checkups, is called HbA1c. This is roughly the percent of red blood cells that have a glucose molecule attached. It measures average exposure to glucose over a period of several months.

This diagnostic test is becoming widely used and is sometimes called A1c by health care professionals. Table 2.3 gives data on both HbA1c and FPG for

18 diabetics five months after they completed a diabetes education class.<sup>21</sup>

Because both FPG and HbA1c measure blood glucose, we expect a positive association. The scatterplot in Figure 2.25(a) shows a surprisingly weak relationship, with correlation  $r = 0.4819$ . The line on the plot is the least-squares regression line for predicting FPG from HbA1c. Its equation is

$$\hat{y} = 66.4 + 10.41x$$

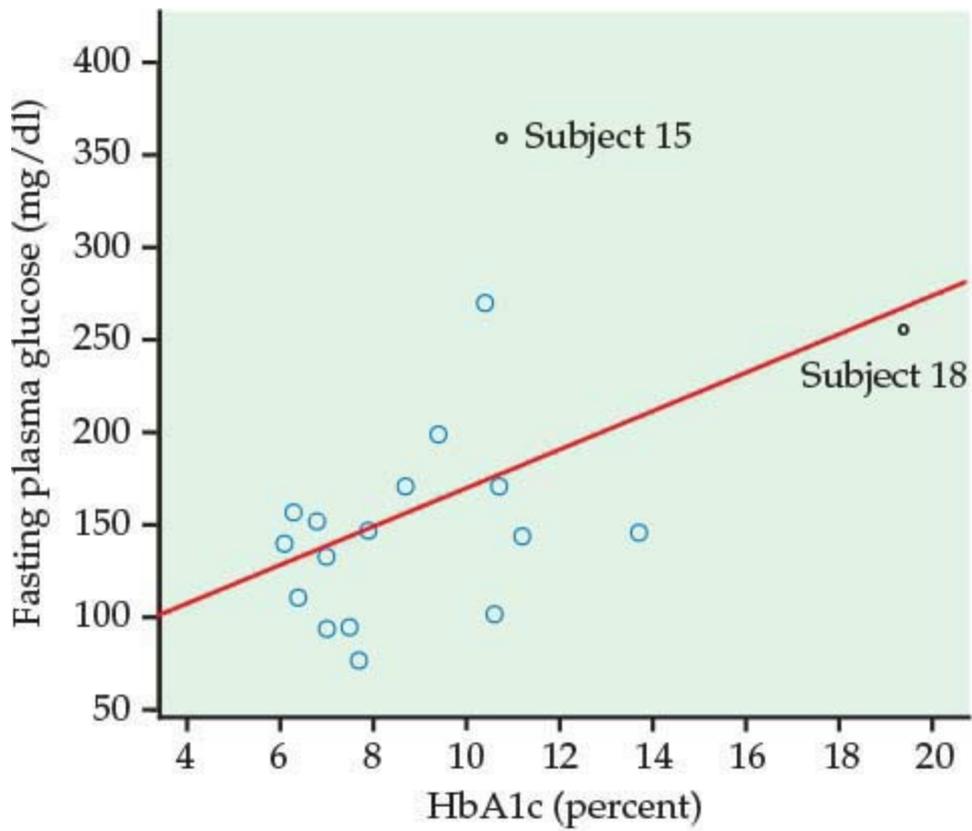
It appears that one-time measurements of FPG can vary quite a bit among people with similar long-term levels, as measured by HbA1c. This is why A1c is an important new diagnostic test.

Two unusual cases are marked in Figure 2.25(a). Subjects 15 and 18 are unusual in different ways. Subject 15 has dangerously high FPG and lies far from the regression line in the  $y$  direction. Subject 18 is close to the line but far out in the  $x$  direction. The residual plot in Figure 2.25(b) confirms that Subject 15 has a large residual and that Subject 18 does not.

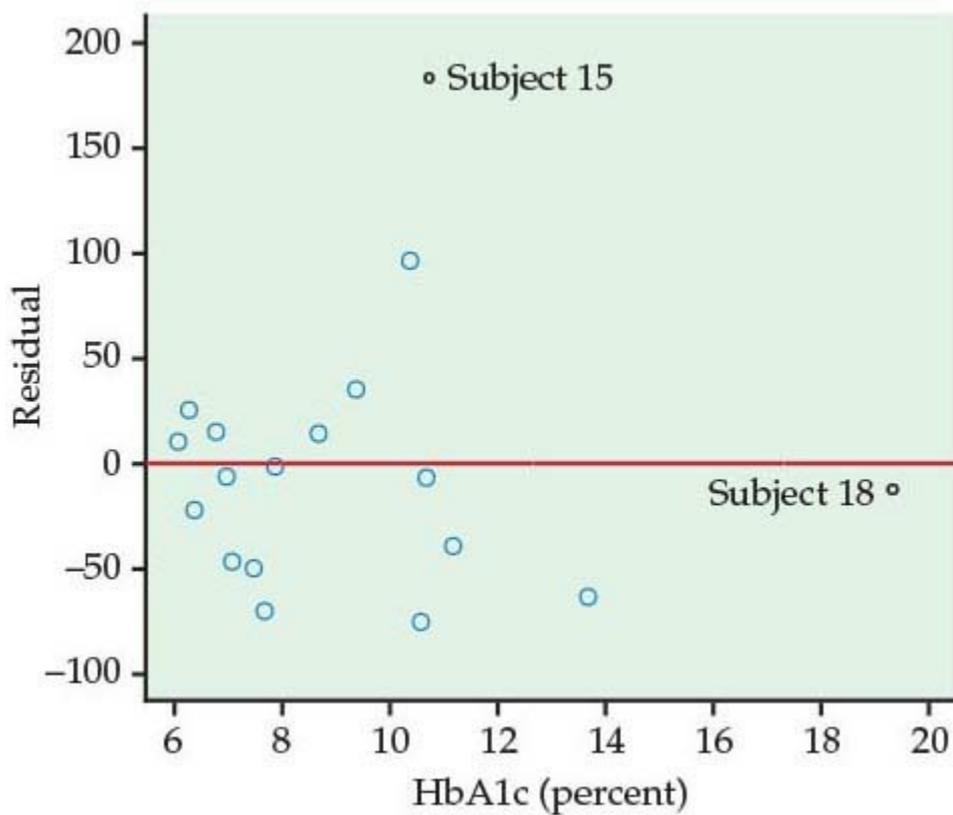
Points that are outliers in the  $x$  direction, like Subject 18, can have a strong influence on the position of the regression line. Least-squares lines make the sum of squares of the vertical distances to the points as small as possible. A point that is extreme in the  $x$  direction with no other points near it pulls the line toward itself.

**TABLE 2.3** Two Measures of Glucose Level in Diabetics

Subject	HbA1c (%)	FPG (mg/ml)	Subject	HbA1c (%)	FPG (mg/ml)	Subject	HbA1c (%)	FPG (mg/ml)
1	6.1	141	7	7.5	96	13	10.6	103
2	6.3	158	8	7.7	78	14	10.7	172
3	6.4	112	9	7.9	148	15	10.7	359
4	6.8	153	10	8.7	172	16	11.2	145
5	7.0	134	11	9.4	200	17	13.7	147
6	7.1	95	12	10.4	271	18	19.3	255



(a)



(b)

**FIGURE 2.25**

(a) Scatterplot of fasting plasma glucose against HbA1c (which measures long-term blood

glucose), with the least-squares regression line, for Example 2.27. (b) Residual plot for the regression of fasting plasma glucose on HbA1c. Subject 15 is an outlier in fasting plasma glucose. Subject 18 is an outlier in HbA1c that may be influential but does not have a large residual.

## OUTLIERS AND INFLUENTIAL OBSERVATIONS IN REGRESSION

An **outlier** is an observation that lies outside the overall pattern of the other observations. Points that are outliers in the  $y$  direction of a scatterplot have large regression residuals, but other outliers need not have large residuals.

An observation is **influential** for a statistical calculation if removing it would markedly change the result of the calculation. Points that are outliers in the  $x$  direction of a scatterplot are often influential for the least-squares regression line.

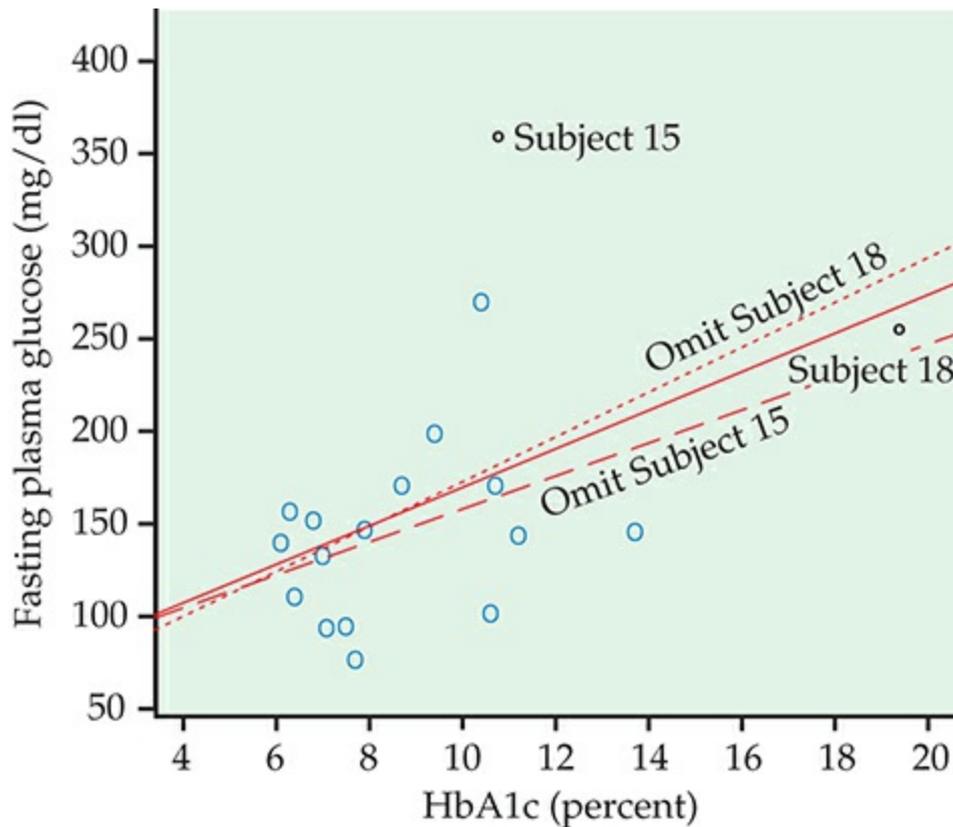
Influence is a matter of degree—how much does a calculation change when we remove an observation? It is difficult to assess influence on a regression line without actually doing the regression both with and without the suspicious observation. A point that is an outlier in  $x$  is often influential. But if the point happens to lie close to the regression line calculated from the other observations, then its presence will move the line only a little and the point will not be influential.

The influence of a point that is an outlier in  $y$  depends on whether there are many other points with similar values of  $x$  that hold the line in place. Figures 2.25(a) and (b) identify two unusual observations. How influential are they?

### Example

#### 2.28 Influential observations.

Subjects 15 and 18 both influence the correlation between FPG and HbA1c, in opposite directions. Subject 15 weakens the linear pattern; if we drop this point, the correlation increases from  $r = 0.4819$  to  $r = 0.5684$ . Subject 18 extends the linear pattern; if we omit this subject, the correlation drops from  $r = 0.4819$  to  $r = 0.3837$ .



**FIGURE 2.26**

Three regression lines for predicting fasting plasma glucose from HbA1c, for Example 2.28. The solid line uses all 18 subjects. The dotted line leaves out Subject 18. The dashed line leaves out Subject 15. “Leaving one out” calculations are the surest way to assess influence.

To assess influence on the least-squares line, we recalculate the line leaving out a suspicious point. Figure 2.26 shows three least-squares lines. The solid line is the regression line of FPG on HbA1c based on all 18 subjects. This is the same line that appears in Figure 2.25(a). The dotted line is calculated from all subjects except Subject 18. You see that point 18 does pull the line down toward itself. But the influence of Subject 18 is not very large—the dotted and solid lines are close together for HbA1c values between 6 and 14, the range of all except Subject 18.

The dashed line omits Subject 15, the outlier in  $y$ . Comparing the solid and dashed lines, we see that Subject 15 pulls the regression line up. The influence is again not large, but it exceeds the influence of Subject 18.

We did not need the distinction between outliers and influential observations in Chapter 1. A single large salary that pulls up the mean salary  $\bar{x}$  for a group of workers is an outlier because it lies far above the other salaries. It is also influential because the mean changes when it is removed. In the regression setting, however, not all outliers are influential. Because influential observations draw the regression line toward themselves, we may not be able to spot them by looking for large residuals.

## Beware of the lurking variable

Correlation and regression are powerful tools for measuring the association between two variables and for expressing the dependence of one variable on the other. These tools must be used with an awareness of their limitations. We have seen that

- Correlation measures *only linear association*, and fitting a straight line makes sense only when the overall pattern of the relationship is linear. Always plot your data before calculating.
- *Extrapolation* (using a fitted model far outside the range of the data that we used to fit it) often produces unreliable predictions.
- Correlation and least-squares regression are *not resistant*. Always plot your data and look for potentially influential points.

Another caution is even more important: the relationship between two variables can often be understood only by taking other variables into account. *Lurking variables* can make a correlation or regression misleading.

### LURKING VARIABLE

A **lurking variable** is a variable that is not among the explanatory or response variables in a study and yet may influence the interpretation of relationships among those variables.

### Example

#### 2.29 Discrimination in medical treatment?

Studies show that men who complain of chest pain are more likely to get detailed tests and aggressive treatment such as bypass surgery than are women with similar complaints. Is this association between gender and treatment due to discrimination?

Perhaps not. Men and women develop heart problems at different ages—women are, on the average, between 10 and 15 years older than men. Aggressive treatments are more risky for older patients, so doctors may hesitate to recommend them. Lurking variables—the patient's age and condition—may explain the relationship between gender and doctors'

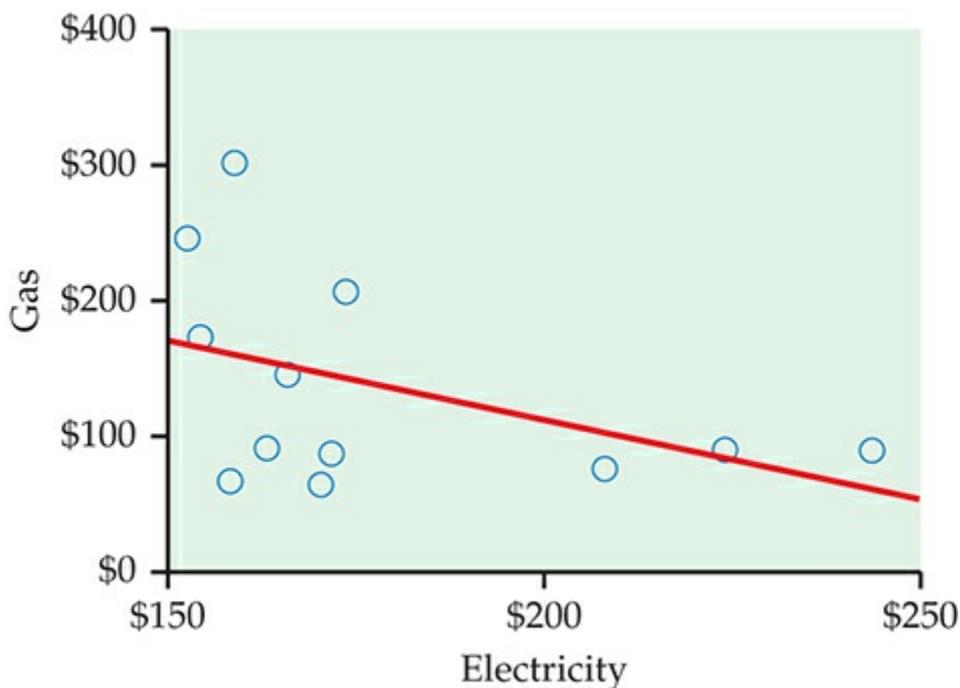
decisions.

Here is an example of a different type of lurking variable.

## Example

### 2.30 Gas and electricity bills.

A single-family household receives bills for gas and electricity each month. The 12 observations for a recent year are plotted with the least-squares regression line in Figure 2.27. We have arbitrarily chosen to put the electricity bill on the  $x$  axis and the gas bill on the  $y$  axis. There is a clear negative association. Does this mean that a high electricity bill causes the gas bill to be low and vice versa?



**FIGURE 2.27**

Scatterplot with least-squares regression line for predicting a household's monthly charges for gas using its monthly charges for electricity, for Example 2.30.

To understand the association in this example, we need to know a little more about the two variables. In this household, heating is done by gas and cooling is done by electricity. Therefore, in the winter months the gas bill will be relatively high and the electricity bill will be relatively low. The pattern is

reversed in the summer months. The association that we see in this example is due to a lurking variable: time of year.

Correlations that are due to lurking variables are sometimes called “nonsense correlations.” The correlation is real. What is nonsense is the suggestion that the variables are directly related so that changing one of the variables *causes* changes in the other. The question of causation is important enough to merit separate treatment in Section 2.7. For now, just *remember that an association between two variables  $x$  and  $y$  can reflect many types of relationship among  $x$ ,  $y$ , and one or more lurking variables.*



## ASSOCIATION DOES NOT IMPLY CAUSATION

An association between an explanatory variable  $x$  and a response variable  $y$ , even if it is very strong, is not by itself good evidence that changes in  $x$  actually cause changes in  $y$ .

Lurking variables sometimes create a correlation between  $x$  and  $y$ , as in Examples 2.29 and 2.30. *When you observe an association between two variables, always ask yourself if the relationship that you see might be due to a lurking variable.* As in Example 2.30, time is often a likely candidate.



## Beware of correlations based on averaged data

Regression or correlation studies sometimes work with averages or other measures that combine information from many individuals. For example, if we plot the average height of young children against their age in months, we will see a very strong positive association with correlation near 1. But individual children of the same age vary a great deal in height. A plot of height against age for individual children will show much more scatter and lower correlation than the plot of average height against age.



*A correlation based on averages over many individuals is usually higher than the correlation between the same variables based on data for individuals.* This fact reminds us again of the importance of noting exactly what variables a statistical study involves.

## **Beware of restricted ranges**

The range of values for the explanatory variable in a regression can have a large impact on the strength of the relationship. For example, if we use age as a predictor of reading ability for a sample of students in the third grade, we will probably see little or no relationship. However, if our sample includes students from grades 1 through 8, we would expect to see a relatively strong relationship. We call this phenomenon the *restricted-range problem*.

### **restricted-range problem**

#### **Example**

##### **2.31 A test for job applicants.**

Your company gives a test of cognitive ability to job applicants before deciding whom to hire. Your boss has asked you to use company records to see if this test really helps predict the performance ratings of employees. The restricted-range problem may make it difficult to see a strong relationship between test scores and performance ratings. The current employees were selected by a mechanism that is likely to result in scores that tend to be higher than those of the entire pool of applicants.

## **BEYOND THE BASICS**

### **Data mining**

Chapters 1 and 2 of this text are devoted to the important aspect of statistics called *exploratory data analysis* (EDA). We use graphs and numerical summaries to examine data, searching for patterns and paying attention to

striking deviations from the patterns we find. In discussing regression, we advanced to using the pattern we find (in this case, a linear pattern) for prediction.

Suppose now that we have a truly enormous database, such as all purchases recorded by the cash register scanners of a national retail chain during the past week. Surely this treasure chest of data contains patterns that might guide business decisions. If we could see clearly the types of activewear preferred in large California cities and compare the preferences of small Midwest cities—right now, not at the end of the season—we might improve profits in both parts of the country by matching stock with demand. This sounds much like EDA, and indeed it is. Exploring really large databases in the hope of finding useful patterns is called *data mining*. Here are some distinctive features of data mining:

#### data mining

- When you have terabytes of data, even straightforward calculations and graphics become very time-consuming. So efficient algorithms are very important.
- The structure of the database and the process of storing the data (the fashionable term is *data warehousing*), perhaps by unifying data scattered across many departments of a large corporation, require careful consideration.
- Data mining requires automated tools that work based on only vague queries by the user. The process is too complex to do step-by-step as we have done in EDA.

All these features point to the need for sophisticated computer science as a basis for data mining. Indeed, data mining is often viewed as a part of computer science. Yet many statistical ideas and tools—mostly tools for dealing with multidimensional data, not the sort of thing that appears in a first statistics course—are very helpful. Like many other modern developments, data mining crosses the boundaries of traditional fields of study.

Do remember that the perils we associate with blind use of correlation and regression are yet more perilous in data mining, where the fog of an immense database can prevent clear vision. Extrapolation, ignoring lurking variables, and confusing association with causation are traps for the unwary data miner.

## SECTION 2.5 Summary

You can examine the fit of a regression line by plotting the **residuals**, which are the differences between the observed and predicted values of  $y$ . Be on the lookout for points with unusually large residuals and also for nonlinear patterns and uneven variation about the line.

Also look for **influential observations**, individual points that substantially change the regression line. Influential observations are often outliers in the  $x$  direction, but they need not have large residuals.

Correlation and regression must be **interpreted with caution**. Plot the data to be sure that the relationship is roughly linear and to detect outliers and influential observations.

**Lurking variables** may explain the relationship between the explanatory and response variables. Correlation and regression can be misleading if you ignore important lurking variables.

We cannot conclude that there is a cause-and-effect relationship between two variables just because they are strongly associated. **High correlation does not imply causation.**

A **correlation based on averages** is usually higher than if we used data for individuals.

## SECTION 2.5 Exercises

*For Exercise 2.92, see page 128; and for Exercise 2.93, see page 128.*

### 2.94 Bone strength.

Exercise 2.18 (page 98) gives the bone strengths of the dominant and the nondominant arms for 15 men who were controls in a study. The least-squares regression line for these data is

$$\text{dominant} = 2.74 + (0.936 \times \text{nondominant})$$

Here are the data for the first four cases:

ID	Nondominant	Dominant	ID	Nondominant	Dominant
1	16.3	15.7	3	18.7	17.9
2	26.9	25.2	4	22.0	19.1

Calculate the residuals for these four cases.  **ARMSTR**

### 2.95 Bone strength for baseball players.

Refer to the previous exercise. Similar data for baseball players is given in Exercise 2.19 (page 98). The equation of the least-squares line for the baseball players is

$$\text{dominant} = 0.886 + (1.373 \times \text{nondominant})$$

Here are the data for the first four cases:

ID	Nondominant	Dominant	ID	Nondominant	Dominant
16	19.3	17.0	18	25.2	17.7
17	19.0	16.9	19	40.3	21.2

Calculate the residuals for these four cases.  **ARMSTR**

## 2.96 Least-squares regression for radioactive decay.

Refer to Exercise 2.22 (page 99) for the data on radioactive decay of barium-137m. Here are the data: 

**DECAY**

Time	1	3	5	7
Count	578	317	203	118

(a) Using the least-squares regression equation

$$\text{count} = 602.8 - (74.7 \times \text{time})$$

and the observed data, find the residuals for the counts.

(b) Plot the residuals versus time.

(c) Write a short paragraph assessing the fit of the least-squares regression line to these data based on your interpretation of the residual plot.

## 2.97 Least-squares regression for the log counts.

Refer to Exercise 2.23 (page 99), where you analyzed the radioactive decay of barium-137m data using log counts. Here are the data:  **DECAY**

Time	1	3	5	7
Log count	6.35957	5.75890	5.31321	4.77068

(a) Using the least-squares regression equation

$$\log \text{count} = 6.593 - (0.2606 \times \text{time})$$

and the observed data, find the residuals for the counts.

(b) Plot the residuals versus time.

(c) Write a short paragraph assessing the fit of the least-squares regression line to these data based on your interpretation of the residual plot.

## 2.98 College students by state.

Refer to Exercise 2.21 (page 99), where you examined the relationship between the number of undergraduate college students and the populations for the 50 states.  **COLLEGE**

(a) Make a scatterplot of the data with the least-squares regression line.

(b) Plot the residuals versus population.

(c) Focus on California, the state with the largest population. Is this state an outlier when you consider only the distribution of population? Explain your answer and describe what graphical and numerical summaries you used as the basis for your conclusion.

(d) Is California an outlier in the distribution of undergraduate college students? Explain your answer and describe what graphical and numerical summaries you used as the basis for your conclusion.

(e) Is California an outlier when viewed in terms of the relationship between number of undergraduate college students and population? Explain your answer and describe what graphical and numerical summaries you used as the basis for your conclusion.

(f) Is California influential in terms of the relationship between number of undergraduate college students and population? Explain your answer and describe what graphical and numerical summaries you used as the basis for your conclusion.

## 2.99 College students by state using logs.

Refer to the previous exercise. Answer parts (a) through (f) for that exercise using the logs of both variables. Write a short paragraph summarizing your findings and comparing them with those from the previous exercise.  COLLEGE

## 2.100 Compare numbers of college students over time.

The data file COLYEAR gives the numbers of college undergraduate students for the years 1970, 1980, 1990, 2000, 2006, 2007, 2008, 2009, and 2011 for each of the 50 states. For this exercise, we will focus on the years 2011 and 2006.<sup>22</sup>  COLYEAR

(a) Make a scatterplot of the data with number of undergraduate students in year 2006 as the explanatory variable and number of undergraduate students in year 2011 as the response variable. Include the least-squares regression line on your plot.

(b) Plot the residuals versus the number of undergraduate students in 2006.

(c) Give a simple explanation of what it means for a state to have a positive residual.

(d) Are there any outliers or influential observations? Give reasons for your answers.

(e) Compare the scatterplot with the residual plot as a graphical tool for detecting outliers.

## 2.101 College students over time using logs.

Refer to the previous exercise. Let's examine the effect of using a log transformation on the numbers of undergraduate college students.  COLYEAR

(a) Make a scatterplot of the data with the least-squares regression line.

(b) Plot the residuals versus the log of the number of undergraduate students in 2006.

(c) Are there any outliers or influential observations? Give reasons for your answers.

(d) Compare your results for this exercise with those for the previous exercise.

(e) Discuss some advantages and disadvantages of using logs for these data.

## 2.102 Make some scatterplots.

For each of the following scenarios, make a scatterplot with 10 observations that show a moderate positive association, plus one that illustrates the unusual case. Explain each of your answers.

(a) An outlier in  $x$  that is not influential for the regression.

- (b) An outlier in  $x$  that is influential for the regression.
- (c) An influential observation that is not an outlier in  $x$ .
- (d) An observation that is influential for the intercept but not for the slope.

### 2.103 What's wrong?

Each of the following statements contains an error. Describe each error and explain why the statement is wrong.

- (a) An influential observation will always have a large residual.
- (b) High correlation is never present when there is causation.
- (c) If we have data at values of  $x$  equal to 1, 2, 3, 4, and 5, and we try to predict the value of  $y$  for  $x = 2.5$  using a least-squares regression equation, we are doing an extrapolation.

### 2.104 What's wrong?

Each of the following statements contains an error. Describe each error and explain why the statement is wrong.

- (a) If the residuals are all negative, this implies that there is a negative relationship between the response variable and the explanatory variable.
- (b) A strong negative relationship does not imply that there is an association between the explanatory variable and the response variable.
- (c) A lurking variable is always something that can be measured.

### 2.105 Internet use and babies.

Exercise 2.28 (page 100) explores the relationship between Internet use and birthrate for 106 countries. Figure 2.11 (page 100) is a scatterplot of the data. It shows a negative association between these two variables. Do you think that this plot indicates that Internet use causes people to have fewer babies? Give another possible explanation for why these two variables are negatively associated. 

### 2.106 A lurking variable.

The effect of a lurking variable can be surprising when individuals are divided into groups. In recent years, the mean SAT score of all high school seniors has increased. But the mean SAT score has decreased for students at each level of high school grades (A, B, C, and so on). Explain how grade inflation in high school (the lurking variable) can account for this pattern. *A relationship that holds for each group within a population need not hold for the population as a whole. In fact, the relationship can even change direction.*

### 2.107 How's your self-esteem?

People who do well tend to feel good about themselves. Perhaps helping people feel good about themselves will help them do better in their jobs and in life. For a time, raising self-esteem became a goal in many schools and companies. Can you think of explanations for the association between high self-esteem and good performance other than “Self-esteem causes better work”?

## 2.108 Are big hospitals bad for you?

A study shows that there is a positive correlation between the size of a hospital (measured by its number of beds  $x$ ) and the median number of days  $y$  that patients remain in the hospital. Does this mean that you can shorten a hospital stay by choosing a small hospital? Why?

## 2.109 Does herbal tea help nursing-home residents?

A group of college students believes that herbal tea has remarkable powers. To test this belief, they make weekly visits to a local nursing home, where they visit with the residents and serve them herbal tea. The nursing-home staff reports that after several months many of the residents are healthier and more cheerful. We should commend the students for their good deeds but doubt that herbal tea helped the residents. Identify the explanatory and response variables in this informal study. Then explain what lurking variables account for the observed association.

## 2.110 Price and ounces.

In Example 2.2 (page 82) and Exercise 2.3 (page 84) we examined the relationship between the price and the size of a Mocha Frappuccino. The 12-ounce Tall drink costs \$3.75, the 16-ounce Grande is \$4.35, and the 24-ounce Venti is \$4.85.

- Plot the data and describe the relationship. (Explain why you should plot size in ounces on the  $x$  axis.)
- Find the least-squares regression line for predicting the price using size. Add the line to your plot.
- Draw a vertical line from the least-squares line to each data point. This gives a graphical picture of the residuals.
- Find the residuals and verify that they sum to zero.
- Plot the residuals versus size. Interpret this plot.



## 2.111 Use the applet.

It isn't easy to guess the position of the least-squares line by eye. Use the *Correlation and Regression* applet to compare a line you draw with the least-squares line. Click on the scatterplot to create a group of 15 to 20 points from lower left to upper right with a clear positive straight-line pattern (correlation around 0.7). Click the "Draw line" button and use the mouse to draw a line through the middle of the cloud of points from lower left to upper right. Note the "thermometer" that appears above the plot. The red portion is the sum of the squared vertical distances from the points in the plot to the least-squares line. The green portion is the "extra" sum of squares for your line—it shows by how much your line misses the smallest possible sum of squares.

- You drew a line by eye through the middle of the pattern. Yet the right-hand part of the bar is probably almost entirely green. What does that tell you?
- Now click the "Show least-squares line" box. Is the slope of the least-squares line smaller (the new line is less steep) or larger (line is steeper) than that of your line? If you repeat this exercise several times, you will consistently get the same result. *The least-squares line minimizes the vertical distances of the points from the line. It is not the line through the "middle" of the cloud of points.* This is one reason why it is hard to draw a good regression line by eye.



## 2.112 Use the applet.

Go to the *Correlation and Regression* applet. Click on the scatterplot to create a group of 10 points in the lower-right corner of the scatterplot with a strong straight-line pattern (correlation about  $-0.9$ ). Now click the “Show least-squares line” box to display the regression line.

- (a) Add one point at the upper left that is far from the other 10 points but exactly on the regression line. Why does this outlier have no effect on the line even though it changes the correlation?
- (b) Now drag this last point down until it is opposite the group of 10 points. You see that one end of the least-squares line chases this single point, while the other end remains near the middle of the original group of 10. What makes the last point so influential?

### 2.113 Education and income.

There is a strong positive correlation between years of education and income for economists employed by business firms. (In particular, economists with doctorates earn more than economists with only a bachelor’s degree.) There is also a strong positive correlation between years of education and income for economists employed by colleges and universities. But when all economists are considered, there is a *negative* correlation between education and income. The explanation for this is that business pays high salaries and employs mostly economists with bachelor’s degrees, while colleges pay lower salaries and employ mostly economists with doctorates. Sketch a scatterplot with two groups of cases (business and academic) that illustrates how a strong positive correlation within each group and a negative overall correlation can occur together.

### 2.114 Dangers of not looking at a plot.

Table 2.2 (page 125) presents four sets of data prepared by the statistician Frank Anscombe to illustrate the dangers of calculating without first plotting the data.<sup>23</sup>  ANSC

- (a) Use  $x$  to predict  $y$  for each of the four data sets. Find the predicted values and residuals for each of the four regression equations.
- (b) Plot the residuals versus  $x$  for each of the four data sets.
- (c) Write a summary of what the residuals tell you for each data set, and explain how the residuals help you to understand these data.

## 2.6 Data Analysis for Two-Way Tables

When you complete this section, you will be able to

- Identify the row variable, the column variable, and the cells in a two-way table.
- Find and interpret the joint distribution in a two-way table.
- Find and interpret the marginal distributions in a two-way table.
- Use the conditional distributions to describe the relationship displayed in a two-way table.
- Determine the joint distribution, the marginal distributions, and the conditional distributions in a two-way table from software output.
- Interpret examples of Simpson's paradox.

When we study relationships between two variables, one of the first questions we ask is whether each variable is quantitative or categorical. For two quantitative variables, we use a scatterplot to examine the relationship, and we fit a line to the data if the relationship is approximately linear. If one of the variables is quantitative and the other is categorical, we can use the methods in Chapter 1 to describe the distribution of the quantitative variable for each value of the categorical variable. This leaves us with the situation where both variables are categorical. In this section we discuss methods for studying these relationships.

Some variables—such as gender, race, and occupation—are inherently categorical. Other categorical variables are created by grouping values of a quantitative variable into classes. Published data are often reported in grouped form to save space. To describe categorical data, we use the *counts* (frequencies) or *percents* (relative frequencies) of individuals that fall into various categories.



quantitative and categorical variables, p. 3

### The two-way table

A key idea in studying relationships between two variables is that both variables must be measured on the same individuals or cases. When both variables are categorical, the raw data are summarized in a ***two-way table*** that gives counts of observations for each combination of values of the two categorical variables. Here is an example.

**Example****2.32 Is the calcium intake adequate?**

IOM

Young children need calcium in their diet to support the growth of their bones. The Institute of Medicine provides guidelines on how much calcium should be consumed for people of different ages.<sup>24</sup> One study examined whether or not a sample of children consumed an adequate amount of calcium, based on these guidelines. Since there are different requirements for children aged 5 to 10 years and for children aged 11 to 13 years, the children were classified into these two age groups. For each student, his or her calcium intake was classified as meeting or not meeting the requirement. There were 2029 children in the study. Here are the data:<sup>25</sup>

Two-way table for “met requirement” and age		
Met requirement	Age (years)	
	5 to 10	11 to 13
No	194	557
Yes	861	417

We see that 194 children aged 5 to 10 did not meet the calcium requirement, and 861 children aged 5 to 10 years met the calcium requirement.

**USE YOUR KNOWLEDGE****2.115 Read the table.**

## IOM

How many children aged 11 to 13 met the requirement? How many did not?

For the calcium requirement example, we could view age as an explanatory variable and “met requirement” as a response variable. This is why we put age in the columns (like the  $x$  axis in a scatterplot) and “met requirement” in the rows (like the  $y$  axis in a scatterplot). We call “met requirement” the *row variable* because each horizontal row in the table describes whether or not the requirement was met. Age is the *column variable* because each vertical column describes one age group. Each combination of values for these two variables is called a *cell*. For example, the cell corresponding to children who are 5 to 10 years old and who have not met the requirement contains the number 194. This table is called a  $2 \times 2$  table because there are 2 rows and 2 columns.

row variable

column variable

cell

To describe relationships between two categorical variables, we compute different types of percents. Our job is easier if we expand the basic two-way table by adding various totals. We illustrate the idea with our calcium requirement example.

### Example

#### 2.33 Add the margins to the table.



We expand the table in Example 2.32 by adding the totals for each row, for each column, and the total number of all the observations. Here is the result:

Two-way table for “met requirement” and age

Met requirement	Age (years)		Total
	5 to 10	11 to 13	
No	194	557	751
Yes	861	417	1278
Total	1055	974	2029

In this study there were 1055 children aged 5 to 10. The total number of children who did not meet the calcium requirement is 751, and the total number of children in the study is 2029.

## USE YOUR KNOWLEDGE

### 2.116 Read the margins of the table.



IOM

How many children aged 5 to 10 were subjects in the calcium requirement study? What is the total number of children who did not meet the calcium requirement?

In this example, be sure that you understand how the table is obtained from the raw data. Think about a data file with one line per subject. There would be 2029 lines or records in this data set. In the two-way table, each individual is counted once and only once. As a result, the sum of the counts in the table is the total number of individuals in the data set. *Most errors in the use of categorical-data methods come from a misunderstanding of how these tables are constructed.*



## Joint distribution

We are now ready to compute some proportions that help us understand the data in a two-way table. Suppose that we are interested in the children aged 5 to 10 years who do not meet the calcium requirement. The proportion of these is simply 194 divided by 2029, or 0.0956. We would estimate that 9.56% of children in the

population from which this sample was drawn are 5- to 10-year-olds who do not meet the calcium requirement. For each cell, we can compute a proportion by dividing the cell entry by the total sample size. The collection of these proportions is the *joint distribution* of the two categorical variables.

joint distribution

## Example

### 2.34 The joint distribution.

For the calcium requirement example, the joint distribution of “met requirement” and age is



IOM

Joint distribution of “met requirement” and age		
Met requirement	Age (years)	
	5 to 10	11 to 13
No	0.0956	0.2745
Yes	0.4243	0.2055

Because this is a distribution, the sum of the proportions should be 1. For this example the sum is 0.9999. The difference is due to roundoff error.

## USE YOUR KNOWLEDGE

### 2.117 Explain the computation.



IOM

Explain how the entry for the children aged 11 to 13 who met the calcium requirement in Example 2.34 is computed from the table in Example 2.33.

How might we use the information in the joint distribution for this example? Suppose that we were to develop an outreach unit to increase the consumption of calcium. The distribution suggests that the older students should be targeted if we have to make a choice because of limited funds. Of the children aged 11 to 13 years, 27.45% do not meet the calcium requirement; but only 9.56% of the children aged 5 to 10 years do not meet the requirement. For other uses of these data, we may need to calculate different numerical summaries. Let's look at the distribution of age.

## Marginal distributions

When we examine the distribution of a single variable in a two-way table, we are looking at a *marginal distribution*. There are two marginal distributions, one for each categorical variable in the two-way table. They are very easy to compute.

marginal distribution

### Example

#### 2.35 The marginal distribution of age.

Look at the table in Example 2.33. The total numbers of children aged 5 to 10 and children aged 11 to 13 are given in the bottom row, labeled "Total." Our sample has 1055 children aged 5 to 10 and 974 children aged 11 to 13. To find the marginal distribution of age we simply divide these numbers by the total sample size, 2029. The marginal distribution of age is



Marginal distribution of age		
	5 to 10	11 to 13
Proportion	0.52	0.48

Note that the proportions sum to 1; there is no roundoff error.

Often we prefer to use percents rather than proportions. Here is the marginal distribution of age described with percents:

Marginal distribution of age		
	5 to 10	11 to 13
Percent	52%	48%

Which form do you prefer?

The percent of children in each age group is approximately the same. This is interesting because the first category includes six ages (5, 6, 7, 8, 9, and 10); whereas the second includes only three ages (11, 12, and 13). Recall that the age categories were chosen in this way because the Institute of Medicine defined the calcium requirement differently for these age groups. In this study, the children were selected from grades 4, 5, and 6. The distribution of ages within these grades explains the marginal distribution of age for our sample.

The other marginal distribution for this example is the distribution of “met requirement.”

## Example

### 2.36 The marginal distribution of “met requirement.”

Here is the marginal distribution of “met requirement,” in percents:



Marginal distribution of “met requirement”		
	No	Yes
Percent	37.01%	62.99%

## USE YOUR KNOWLEDGE

## 2.118 Explain the marginal distribution.



IOM

Explain how the marginal distribution of “met requirement” given in Example 2.36 is computed from the entries in the table given in Example 2.33.

Each marginal distribution from a two-way table is a distribution for a single categorical variable. We can use a bar graph or a pie chart to display such a distribution. For our two-way table, we will be content with numerical summaries: for example, 52% of the children are aged 5 to 10, and 37% of the children are not meeting their calcium requirement. When we have more rows or columns, the graphical displays are particularly useful.



bar graphs and pie charts, p. 9

## Describing relations in two-way tables

The table in Example 2.33 contains much more information than the two marginal distributions of age alone and “met requirement” alone. We need to do a little more work to examine the relationship. *Relationships among categorical variables are described by calculating appropriate percents from the counts given.* What percents do you think we should use to describe the relationship between age and meeting the calcium requirement?

### Example

#### 2.37 Meeting the calcium requirement for children aged 5 to 10.



IOM

What percent of the children aged 5 to 10 in our sample met the calcium requirement? This is the count of the children who are 5 to 10 years old and who met the calcium requirement as a percent of the number children who are 5 to 10 years old:

$$8611055=0.8161=82\%$$

## USE YOUR KNOWLEDGE

### 2.119 Find the percent.



IOM

Show that the percent of children 11 to 13 years old who met the calcium requirement is 43%.

## Conditional distributions

In Example 2.37 we looked at the children aged 5 to 10 alone and examined the distribution of the other categorical variable, “met requirement.” Another way to say this is that we conditioned on the value of age, 5 to 10 years old. Similarly, we can condition on the value of age being 11 to 13 years old. When we condition on the value of one variable and calculate the distribution of the other variable, we obtain a *conditional distribution*. Note that in Example 2.37 we calculated only the percent for children aged 5 to 10 years. The complete conditional distribution gives the proportions or percents for all possible values of the conditioning variable.

conditional distribution

## Example

**2.38 Conditional distribution of “met requirement” for children aged 5 to 10.**

For children aged 5 to 10 years, the conditional distribution of the “met requirement” variable in terms of percents is



IOM

<b>Conditional distribution of “met requirement” for children aged 5 to 10</b>		
	<b>No</b>	<b>Yes</b>
Percent	18.39%	81.61%

Note that we have included the percents for both of the possible values, Yes and No, of the “met requirement” variable. These percents sum to 100%.

## USE YOUR KNOWLEDGE

### 2.120 A conditional distribution.

Perform the calculations to show that the conditional distribution of “met requirement” for children aged 11 to 13 years is



IOM

<b>Conditional distribution of “met requirement” for children aged 11 to 13</b>		
	<b>No</b>	<b>Yes</b>
Percent	57.19%	42.81%

Comparing the conditional distributions (Example 2.38 and Exercise 2.120) reveals the nature of the association between age and meeting the calcium requirement. In this set of data the older children are more likely to fail to meet the calcium requirement.

Bar graphs can help us to see relationships between two categorical variables. No single graph (such as a scatterplot) portrays the form of the relationship between categorical variables, and no single numerical measure (such as the correlation) summarizes the strength of an association. Bar graphs are flexible enough to be helpful, but you must think about what comparisons you want to display. For numerical measures, we must rely on well-chosen percents or on more

advanced statistical methods.<sup>26</sup>

*A two-way table contains a great deal of information in compact form. Making that information clear almost always requires finding percents. You must decide which percents you need. Of course, we prefer to use software to compute the joint, marginal, and conditional distributions.*



## Example

### 2.39 Software output.



Figure 2.28 gives computer output for the data in Example 2.32 using Minitab, SPSS, and JMP. There are minor variations among software packages, but these outputs are typical of what is usually produced. Each cell in the  $2 \times 2$  table has four entries. These are the count (the number of observations in the cell), the conditional distributions for rows and columns, and the joint distribution. Note that all of these are expressed as percents rather than proportions. Marginal totals and distributions are given in the rightmost column and the bottom row.

Most software packages order the row and column labels numerically or alphabetically. In general, it is better to use words rather than numbers for the column labels. This sometimes involves some additional work, but it avoids the kind of confusion that can result when you forget the real values associated with each numerical value. You should verify that the entries in Figure 2.28 correspond to the calculations that we performed in Examples 2.34 to 2.38. In addition, verify the calculations for the conditional distributions of age for each value of “met requirement.”

Rows: Age      Columns: Met

	No	Yes	All
A05to10	194	861	1055
	18.39	81.61	100.00
	25.83	67.37	52.00
	9.56	42.43	52.00
Allto13	557	417	974
	57.19	42.81	100.00
	74.17	32.63	48.00
	27.45	20.55	48.00
All	751	1278	2029
	37.01	62.99	100.00
	100.00	100.00	100.00
	37.01	62.69	100.00

Cell Contents:      Count  
                          % of Row  
                          % of Column  
                          % of Total

Show the ReportPad

(a) Minitab

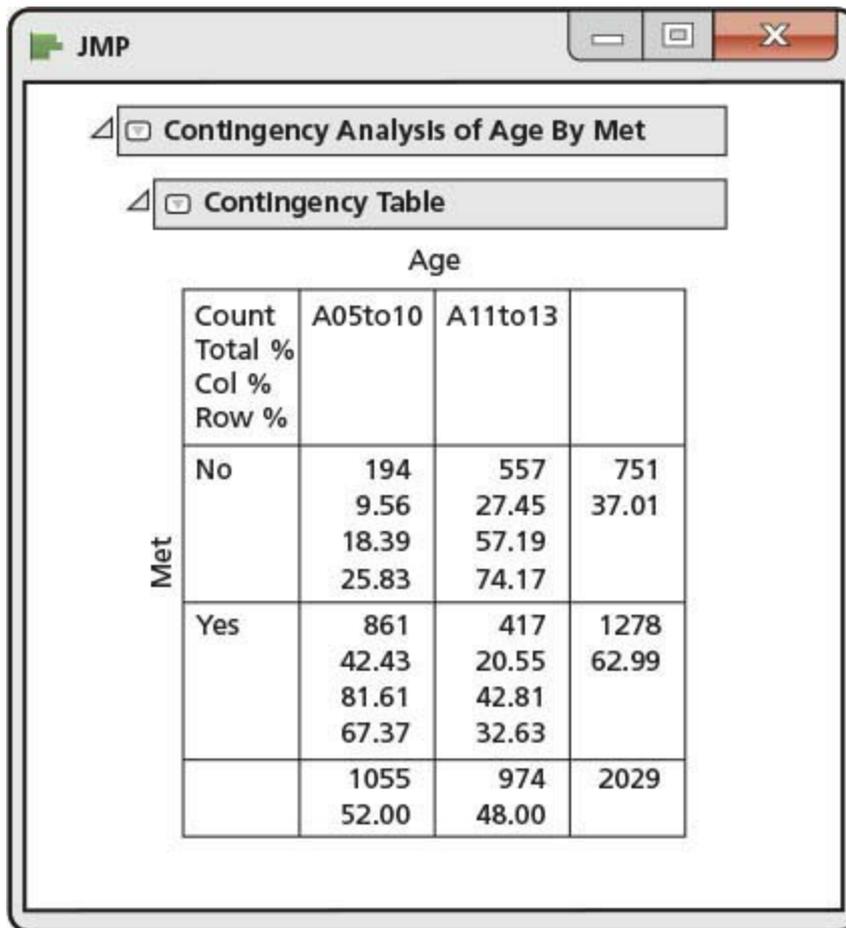
\*Output1 - IBM SPSS Statistics Viewer

**Met \* Age Crosstabulation**

			Age		Total
			A05to10	A11to13	
Met	No	Count	194	557	751
		% within Met	25.8%	74.2%	100.0%
		% within Age	18.4%	57.2%	37.0%
		% of Total	9.6%	27.5%	37.0%
Yes	Yes	Count	861	417	1278
		% within Met	67.4%	32.6%	100.0%
		% within Age	81.6%	42.8%	63.0%
		% of Total	42.4%	20.6%	63.0%
Total	Total	Count	1055	974	2029
		% within Met	52.0%	48.0%	100.0%
		% within Age	100.0%	100.0%	100.0%
		% of Total	52.0%	48.0%	100.0%

IBM SPSS Statistics Processor is ready H: 115, W: 476 pt.

(b) SPSS



(c) JMP

**FIGURE 2.28**

Computer output for the calcium requirement study, for Example 2.39.(a) Minitab, (b) SPSS, (c) JMP.

**Simpson’s paradox**

As is the case with quantitative variables, the effects of lurking variables can strongly influence relationships between two categorical variables. Here is an example that demonstrates the surprises that can await the unsuspecting consumer of data.

**Example**

**2.40 Which customer service representative is better?**



### CUSTSER

A customer service center has a goal of resolving customer questions in 10 minutes or less. Here are the records for two representatives:

Goal met	Representative	
	Alexis	Peyton
Yes	172	118
No	28	82
Total	200	200

Alexis has met the goal 172 times out of 200, a success rate of 86%. For Peyton, the success rate is 118 out of 200, or 59%. Alexis clearly has the better success rate.

Let's look at the data in a little more detail. The data summarized come from two different weeks in the year.

## Example

### 2.41 Look at the data more carefully.



### CUSTSER

Here are the counts broken down by week:

Goal met	Week 1		Week 2	
	Alexis	Peyton	Alexis	Peyton
Yes	162	19	10	99
No	18	1	10	81
Total	180	20	20	180

For Week 1, Alexis met the goal 90% of the time (162/180), while Peyton met the goal 95% of the time (19/20). Peyton had the better performance in Week 1. What about Week 2? Here Alexis met the goal 50% of the time (10/20), while the success rate for Peyton was 55% (99/180). Peyton again had the

better performance. How does this analysis compare with the analysis that combined the counts for the two weeks? That analysis clearly showed that Alexis had the better performance, 59% versus 86%.

These results can be explained by a lurking variable, Week. The first week was during a period when the product had been in use for several months. Most of the calls to the customer service center concerned problems that had been encountered before. The representatives were trained to answer these questions and usually had no trouble in meeting the goal of resolving the problems quickly. On the other hand, the second week occurred shortly after the release of a new version of the product. Most of the calls during this week concerned new problems that the representatives had not yet encountered. Many more of these questions took longer than the 10-minute goal to resolve.

Look at the totals in the bottom row of the detailed table. During the first week, when calls were easy to resolve, Alexis handled 180 calls and Peyton handled 20. The situation was exactly the opposite during the second week, when the calls were difficult to resolve. There were 20 calls for Alexis and 180 for Peyton.

The original two-way table, which did not take account of week, was misleading. This example illustrates *Simpson's paradox*.

## SIMPSON'S PARADOX

An association or comparison that holds for all of several groups can reverse direction when the data are combined to form a single group. This reversal is called **Simpson's paradox**.

The lurking variables in our Simpson's paradox example, week and problem difficulty, are categorical. That is, they break the observations into groups by, work week. *Simpson's paradox is an extreme form of the fact that observed associations can be misleading when there are lurking variables.*



The data in Example 2.41 are given in a **three-way table** that reports counts for each combination of three categorical variables: week, representative, and whether or not the goal was met. In our example, we constructed the three-way table by constructing two two-way tables for representative by goal, one for each week. The original table in Example 2.40 can be obtained by adding the corresponding counts for these two tables. This process is called **aggregating** the data. When we aggregated data in Example 2.40 we ignored the variable week, which then became a lurking variable. *Conclusions that seem obvious when we look only at aggregated*

*data can become quite different when the data are examined in more detail.*

### three-way table

### aggregation



## SECTION 2.6 Summary

A **two-way table** of counts organizes data about two categorical variables. Values of the **row variable** label the rows that run across the table, and values of the **column variable** label the columns that run down the table. Two-way tables are often used to summarize large amounts of data by grouping outcomes into categories.

The **joint distribution** of the row and column variables is found by dividing the count in each cell by the total number of observations.

The **row totals** and **column totals** in a two-way table give the **marginal distributions** of the two variables separately. It is clearer to present these distributions as percents of the table total. Marginal distributions do not give any information about the relationship between the variables.

To find the **conditional distribution** of the row variable for one specific value of the column variable, look only at that one column in the table. Find each entry in the column as a percent of the column total.

There is a conditional distribution of the row variable for each column in the table. Comparing these conditional distributions is one way to describe the association between the row and the column variables. It is particularly useful when the column variable is the explanatory variable. When the row variable is explanatory, find the conditional distribution of the column variable for each row and compare these distributions.

**Bar graphs** are a flexible means of presenting categorical data. There is no single best way to describe an association between two categorical variables.

We present data on three categorical variables in a **three-way table**, printed as separate two-way tables for each level of the third variable. A comparison between two variables that holds for each level of a third variable can be changed or even reversed when the data are **aggregated** by summing over all levels of the third variable. **Simpson's paradox** refers to the reversal of a comparison by aggregation. It is an example of the potential effect of lurking variables on an observed association.

## SECTION 2.6 Exercises

*For Exercise 2.115, see page 140; for 2.116, see page 141; for 2.117, see page 142; for 2.118, see page 143; for 2.119, see page 144; and for 2.120, see page 144.*

## 2.121 Does drivers ed help?

A study is planned to look at the effect of drivers education programs on accidents. The driving records of all drivers under 18 in a given year will classify each driver as having taken a drivers education course or not. The drivers will also be classified with respect to the number of accidents that they had in the year after they received their license. The categories are zero, one, and two or more accidents.

- (a) There are two variables in this study. Do you think that one is an explanatory variable and that the other is a response variable? Explain your answer.
- (b) Sketch a two-way table that could be used to organize the data. Which variable is the row variable? Which variable is the column variable?
- (c) How many cells are in the table? Describe in words what each of the cells will contain when the data are collected.

## 2.122 Music and video games.

You are planning a study of undergraduates in which you will examine the relationship between listening to music and playing video games. The study subjects will be asked how much time they spend in each of these activities during a typical day. The choices for both activities will be a half hour or less, more than a half hour but less than an hour, and more than an hour.

- (a) There are two variables in this study. Do you think that one is an explanatory variable and that the other is a response variable? Explain your answer.
- (b) Sketch a two-way table that could be used to organize the data. Which variable is the row variable? Which variable is the column variable?
- (c) How many cells are in the table? Describe in words what each of the cells will contain when the data are collected.

## 2.123 Eight is enough.

A healthy body needs good food, and healthy teeth are needed to chew our food so that it can nourish our bodies. The U.S. Army has recognized this fact and requires recruits to pass a dental examination. If you wanted to be a soldier in the Spanish American War, which took place in 1898, you needed to have at least eight teeth. Here is the statement of the requirement:  **TEETH**

*Unless an applicant has at least four sound double teeth, one above and one below on each side of the mouth, and so opposed as to serve the purpose of mastication, he should be rejected.*

A study reported the rejection data for enlistment candidates classified by age. Here are the data:<sup>27</sup>

Rejected	Age (years)					
	Under 20	20 to 25	25 to 30	30 to 35	35 to 40	Over 40
Yes	68	647	1,114	1,783	2,887	3,801
No	58,884	77,992	55,597	43,994	47,569	39,985

- (a) Which variable is the explanatory variable? Which variable is the response variable? Give reasons for

your answer.

(b) Find the joint distribution. Write a brief summary explaining the major features of this distribution.

(c) Find the two marginal distributions. Write a brief summary explaining the major features of these distributions.

(d) Which conditional distribution would you choose to explain the relationship between these two variables? Explain your answer.

(e) Find the conditional distribution that you chose in part (d), and write a summary that includes your interpretation of the relationship based on this conditional distribution.

## 2.124 Survival and class on the *Titanic*.

In Exercise 1.27 (page 25) you created a graphical summary of the number of passengers who survived classified by the accommodations that they had on the ship: first, second, or third class. Let's look at these data with a two-way table.  TITANIC

(a) Create a two-way table that you could use to explore the relationship between survival and class.

(b) Which variable is the explanatory variable and which is the response variable? Give reasons for your answers.

(c) Find the two marginal distributions. Write a brief summary explaining the major features of these distributions.

(d) Which conditional distribution would you choose to explain the relationship between these two variables? Explain your answer.

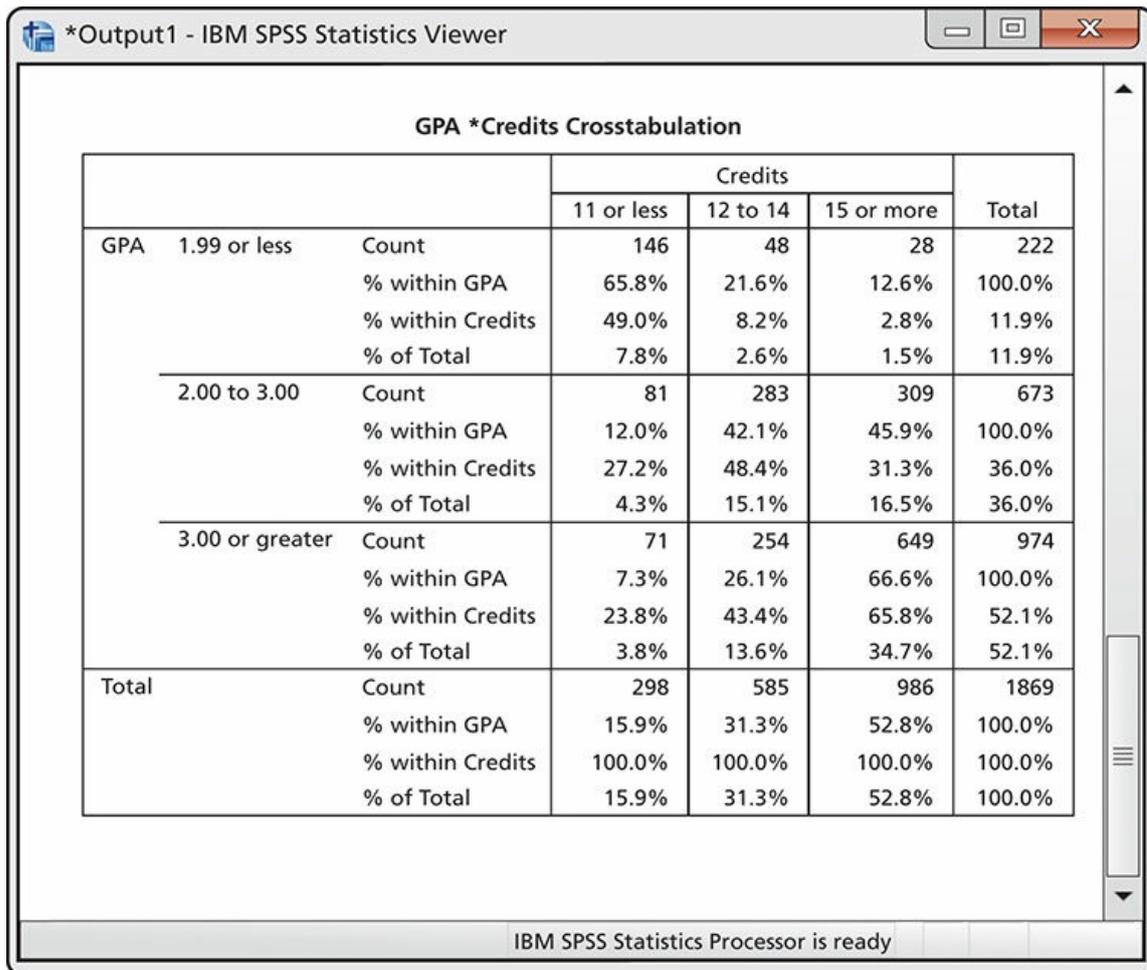
(e) Find the conditional distribution that you chose in part (d), and write a summary that includes your interpretation of the relationship based on this conditional distribution.

## 2.125 Number of credits and grade point average.

A study of undergraduate students examined the relationship between the number of credits taken in a semester and the grade point average. Credits were classified as less than 12, 12 to 14, and 15 or more. For grade point average, three categories were used: less than 2.0, 2.0 to 3.0, and 3.0 and higher.<sup>28</sup> Figure 2.29 gives software output for these data. Use this output to analyze these data, and write a report summarizing your work. Be sure to include a discussion of whether or not you consider this relationship to involve an explanatory variable and a response variable.

## 2.126 Punxsutawney Phil.

At Gobbler's Knob in Punxsutawney, Pennsylvania, there is a gathering every year on February 2. A groundhog, always named Phil, is the center of attraction. If Phil sees his shadow when he emerges from his burrow, tradition says that there will be six more weeks of winter. If he does not see his shadow, spring has arrived. How well has Phil done at predicting the arrival of spring for the past several years? The National Oceanic and Atmospheric Administration has collected data for the 24 years from 1988 to 2011. For each year, whether or not Phil sees his shadow is recorded. This is compared with the February temperature for that year, classified as above or below normal.<sup>29</sup> Figure 2.30 gives software output for these data. Use this output to analyze these data, and write a report summarizing your work. Be sure to include a discussion of whether or not you consider this relationship to involve an explanatory variable and a response variable.



**FIGURE 2.29**

Computer output for the credit and grade point data, for Exercise 2.125.

### 2.127 Exercise and adequate sleep.

A survey of 656 boys and girls who were 13 to 18 years old asked about adequate sleep and other health-related behaviors. The recommended amount of sleep is six to eight hours per night.<sup>30</sup> In the survey 59% of the respondents reported that they got less than this amount of sleep on school nights. An exercise scale was developed and was used to classify the students as above or below the median in this domain. Here is the  $2 \times 2$  table of counts with students classified as getting or not getting adequate sleep and by the exercise variable:  **SLEEP**

Enough sleep	Exercise	
	High	Low
Yes	151	115
No	148	242

- Find the distribution of adequate sleep for the high exercisers.
- Do the same for the low exercisers.
- Summarize the relationship between adequate sleep and exercise using the results of parts (a) and (b).

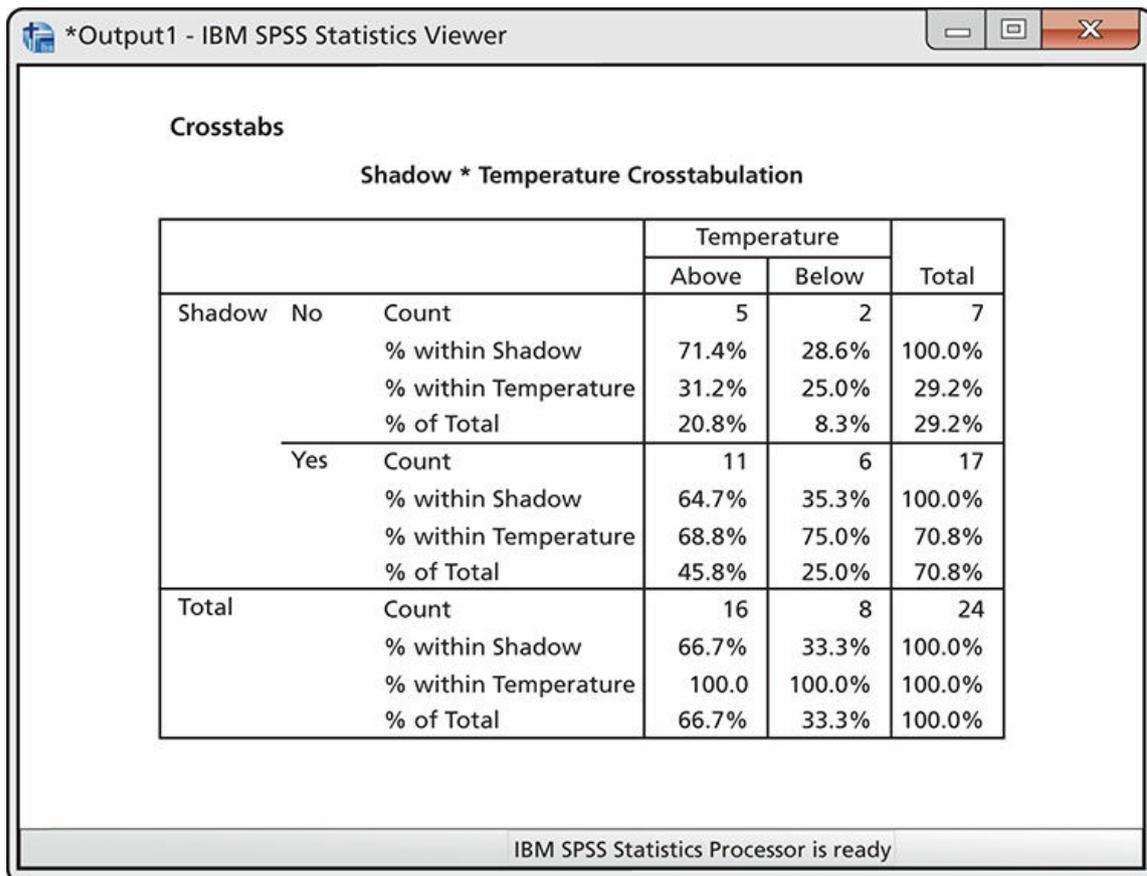
## 2.128 Adequate sleep and exercise.

Refer to the previous exercise.  **SLEEP**

- Find the distribution of exercise for those who get adequate sleep.
- Do the same for those who do not get adequate sleep.
- Write a short summary of the relationship between adequate sleep and exercise using the results of parts (a) and (b).
- Compare this summary with your summary from part (c) of the previous exercise. Which do you prefer? Give a reason for your answer.

## 2.129 Which hospital is safer?

Insurance companies and consumers are interested in the performance of hospitals. The government releases data about patient outcomes in hospitals that can be useful in making informed health care decisions. Here is a two-way table of data on the survival of patients after surgery in two hospitals. All patients undergoing surgery in a recent time period are included. “Survived” means that the patient lived at least 6 weeks following surgery.  **HOSP**



The screenshot shows the IBM SPSS Statistics Viewer window titled '\*Output1 - IBM SPSS Statistics Viewer'. The main content is a Crosstabs table titled 'Shadow \* Temperature Crosstabulation'. The table displays the relationship between 'Shadow' (No, Yes) and 'Temperature' (Above, Below), including counts and percentages for each cell and row/column totals.

			Temperature		Total
			Above	Below	
Shadow	No	Count	5	2	7
		% within Shadow	71.4%	28.6%	100.0%
		% within Temperature	31.2%	25.0%	29.2%
		% of Total	20.8%	8.3%	29.2%
Yes	Count	11	6	17	
	% within Shadow	64.7%	35.3%	100.0%	
	% within Temperature	68.8%	75.0%	70.8%	
	% of Total	45.8%	25.0%	70.8%	
Total	Count	16	8	24	
	% within Shadow	66.7%	33.3%	100.0%	
	% within Temperature	100.0%	100.0%	100.0%	
	% of Total	66.7%	33.3%	100.0%	

IBM SPSS Statistics Processor is ready

**FIGURE 2.30**

Computer output for the Punxsutawney Phil data, for Exercise 2.126.

Hospital A	Hospital B
------------	------------

Died	63	16
Survived	2037	784
Total	2100	800

What percent of Hospital A patients died? What percent of Hospital B patients died? These are the numbers one might see reported in the media.

### 2.130 Patients in “poor” or “good” condition.

Refer to the previous exercise. Not all surgery cases are equally serious, however. Patients are classified as being in either “poor” or “good” condition before surgery. Here are the data broken down by patient condition. The entries in the original two-way table are just the sums of the “poor” and “good” entries in this pair of tables. 

Good Condition		
	Hospital A	Hospital B
Died	6	8
Survived	594	592
Total	600	600

Poor Condition		
	Hospital A	Hospital B
Died	57	8
Survived	1443	192
Total	1500	200

- Find the death rate for Hospital A patients who were classified as “poor” before surgery. Do the same for Hospital B. In which hospital do “poor” patients fare better?
- Repeat (a) for patients classified as “good” before surgery.
- What is your recommendation to someone facing surgery and choosing between these two hospitals?
- How can Hospital A do better in both groups, yet do worse overall? Look at the data and carefully explain how this can happen.

### 2.131 Complete the table.

Here are the row and column totals for a two-way table with two rows and two columns:

$a$	$b$	200
$c$	$d$	200
200	200	400

Find *two different* sets of counts  $a$ ,  $b$ ,  $c$ , and  $d$  for the body of the table that give these same totals. This shows that the relationship between two variables cannot be obtained from the two individual distributions of the variables.

### 2.132 Construct a table with no association.

Construct a  $3 \times 3$  table of counts where there is no apparent association between the row and column variables.

## 2.7 The Question of Causation

When you complete this section, you will be able to

- Identify the differences among causation, common response, and confounding in explaining an association.
- Apply the five criteria for establishing causation.

In many studies of the relationship between two variables, the goal is to establish that changes in the explanatory variable *cause* changes in the response variable. Even when a strong association is present, however, the conclusion that this association is due to a causal link between the variables is often hard to justify. What ties between two variables (and others lurking in the background) can explain an observed association? What constitutes good evidence for causation? We begin our consideration of these questions with a set of observed associations. In each case, there is a clear association between variable  $x$  and variable  $y$ . Moreover, the association is positive whenever the direction makes sense.

### Explaining Association

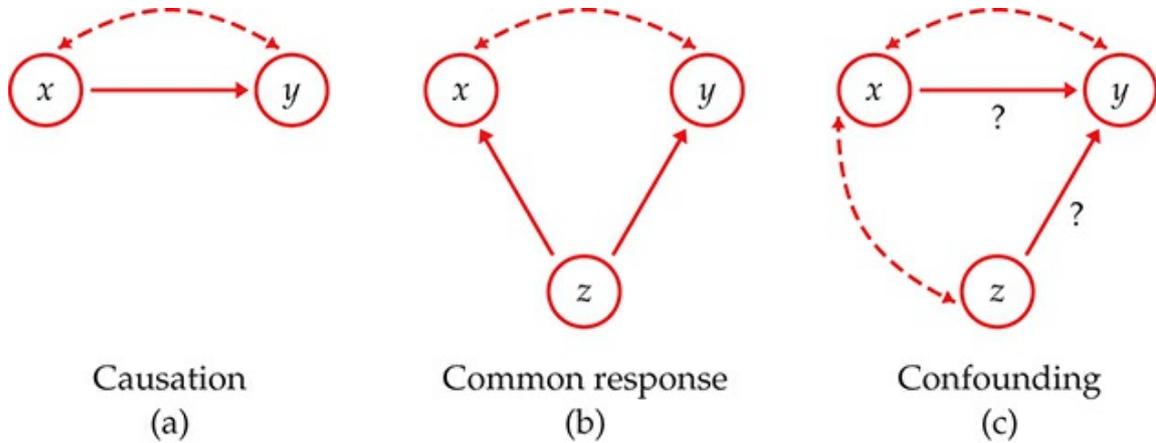
#### Example

#### 2.42 Observed associations.

Here are some examples of observed association between  $x$  and  $y$ :

1.  $x$  = mother's body mass index  
 $y$  = daughter's body mass index
2.  $x$  = amount of the artificial sweetener saccharin in a rat's diet  
 $y$  = count of tumors in the rat's bladder
3.  $x$  = a student's SAT score as a high school senior  
 $y$  = a student's first-year college grade point average
4.  $x$  = monthly flow of money into stock mutual funds  
 $y$  = monthly rate of return for the stock market

5.  $x$  = whether a person regularly attends religious services  
 $y$  = how long the person lives
6.  $x$  = the number of years of education a worker has  
 $y$  = the worker's income



**FIGURE 2.31**

Possible explanations for an observed association. The dashed double-arrow lines show an association. The solid arrows show a cause-and-effect link. The variable  $x$  is explanatory,  $y$  is a response variable, and  $z$  is a lurking variable.

### Explaining association: causation

Figure 2.31 shows in outline form how a variety of underlying links between variables can explain association. The dashed double-arrow line represents an observed association between the variables  $x$  and  $y$ . Some associations are explained by a direct cause-and-effect link between these variables. The first diagram in Figure 2.31 shows “ $x$  causes  $y$ ” by a solid arrow running from  $x$  to  $y$ .

Items 1 and 2 in Example 2.42 are examples of direct causation. *Even when direct causation is present, very often it is not a complete explanation of an association between two variables.* The best evidence for causation comes from experiments that actually change  $x$  while holding all other factors fixed. If  $y$  changes, we have good reason to think that  $x$  caused the change in  $y$ .



### Explaining association: common response

“Beware of the lurking variable” is good advice when thinking about an association between two variables. The second diagram in Figure 2.31 illustrates **common response**. The observed association between the variables  $x$  and  $y$  is explained by a lurking variable  $z$ . Both  $x$  and  $y$  change in response to changes in  $z$ .

This common response creates an association even though there may be no direct causal link between  $x$  and  $y$ .

### common response

The third and fourth items in Example 2.42 illustrate how common response can create an association.

### Explaining association: confounding

For the first item in Example 2.42 we expect that inheritance explains part of the association between the body mass indexes (BMIs) of daughters and their mothers. Can we use  $r$  or  $r^2$  to say how much inheritance contributes to the daughters' BMIs? No. It may well be that mothers who are overweight also set an example of little exercise, poor eating habits, and lots of television. Their daughters pick up these habits to some extent, so the influence of heredity is mixed up with influences from the girls' environment. We call this mixing of influences *confounding*.

## CONFOUNDING

Two variables are **confounded** when their effects on a response variable cannot be distinguished from each other. The confounded variables may be either explanatory variables or lurking variables or both.



*When many uncontrolled variables are related to a response variable, you should always ask whether or not confounding of several variables prevents you from drawing conclusions about causation.* The third diagram in Figure 2.31 illustrates confounding. Both the explanatory variable  $x$  and the lurking variable  $z$  may influence the response variable  $y$ . Because  $x$  is confounded with  $z$ , we cannot distinguish the influence of  $x$  from the influence of  $z$ . We cannot say how strong the direct effect of  $x$  on  $y$  is. In fact, it can be hard to say if  $x$  influences  $y$  at all.

The last two associations in Example 2.42 (Items 5 and 6) are explained in part by confounding.

Many observed associations are at least partly explained by lurking variables. Both common response and confounding involve the influence of a lurking variable (or variables)  $z$  on the response variable  $y$ . The distinction between these two types of relationship is less important than the common element, the influence

of lurking variables. The most important lesson of these examples is one we have already emphasized: **even a very strong association between two variables is not by itself good evidence that there is a cause-and-effect link between the variables.**

## **Establishing causation**

How can a direct causal link between  $x$  and  $y$  be established? The best method—indeed, the only fully compelling method—of establishing causation is to conduct a carefully designed experiment in which the effects of possible lurking variables are controlled. Chapter 3 explains how to design convincing experiments.

Many of the sharpest disputes in which statistics plays a role involve questions of causation that cannot be settled by experiment. Does gun control reduce violent crime? Does living near power lines cause cancer? Has “outsourcing” work to overseas locations reduced overall employment in the United States? All these questions have become public issues. All concern associations among variables. And all have this in common: they try to pinpoint cause and effect in a setting involving complex relations among many interacting variables. Common response and confounding, along with the number of potential lurking variables, make observed associations misleading. Experiments are not possible for ethical or practical reasons. We can’t assign some people to live near power lines or compare the same nation with and without strong gun controls.

### **Example**

#### **2.43 Power lines and leukemia.**

Electric currents generate magnetic fields. So living with electricity exposes people to magnetic fields. Living near power lines increases exposure to these fields. Really strong fields can disturb living cells in laboratory studies. Some people claim that the weaker fields we experience if we live near power lines cause leukemia in children.



It isn't ethical to do experiments that expose children to magnetic fields. It's hard to compare cancer rates among children who happen to live in more and less exposed locations because leukemia is rare and locations vary in many ways other than magnetic fields. We must rely on studies that compare children who have leukemia with children who don't.

A careful study of the effect of magnetic fields on children took five years and cost \$5 million. The researchers compared 638 children who had leukemia and 620 who did not. They went into the homes and actually measured the magnetic fields in the children's bedrooms, in other rooms, and at the front door. They recorded facts about nearby power lines for the family home and also for the mother's residence when she was pregnant. Result: no evidence of more than a chance connection between magnetic fields and childhood leukemia.<sup>31</sup>

"No evidence" that magnetic fields are connected with childhood leukemia doesn't prove that there is no risk. It says only that a careful study could not find any risk that stands out from the play of chance that distributes leukemia cases across the landscape. Critics continue to argue that the study failed to measure some lurking variables, or that the children studied don't fairly represent all children. Nonetheless, a carefully designed study comparing children with and without leukemia is a great advance over haphazard and sometimes emotional counting of cancer cases.

## **Example**

### **2.44 Smoking and lung cancer.**

Despite the difficulties, it is sometimes possible to build a strong case for causation in the absence of experiments. The evidence that smoking causes lung cancer is about as strong as nonexperimental evidence can be.

Doctors had long observed that most lung cancer patients were smokers. Comparison of smokers and similar nonsmokers showed a very strong association between smoking and death from lung cancer. Could the association be due to common response? Might there be, for example, a genetic factor that predisposes people both to nicotine addiction and to lung cancer? Smoking and lung cancer would then be positively associated even if smoking had no direct effect on the lungs. Or perhaps confounding is to blame. It might be that smokers live unhealthy lives in other ways (diet, alcohol, lack of exercise) and that some other habit confounded with smoking is a cause of lung cancer. How were these objections overcome?

Let's answer this question in general terms: what are the criteria for establishing causation when we cannot do an experiment?

- *The association is strong.* The association between smoking and lung cancer is very strong.
- *The association is consistent.* Many studies of different kinds of people in many countries link smoking to lung cancer. That reduces the chance that a lurking variable specific to one group or one study explains the association.
- *Higher doses are associated with stronger responses.* People who smoke more cigarettes per day or who smoke over a longer period get lung cancer more often. People who stop smoking reduce their risk.
- *The alleged cause precedes the effect in time.* Lung cancer develops after years of smoking.
- *The alleged cause is plausible.* Experiments show that tars from cigarette smoke cause cancer when applied to the backs of mice.

Medical authorities do not hesitate to say that smoking causes lung cancer. The U.S. Surgeon General states that cigarette smoking is “the largest avoidable cause of death and disability in the United States.”<sup>32</sup> The evidence for causation is strong—but it is not as strong as the evidence provided by well-designed experiments.

## SECTION 2.7 Summary

Some observed associations between two variables are due to a **cause-and-effect** relationship between these variables, but others are explained by **lurking variables**.

The effect of lurking variables can operate through **common response** if changes in both the explanatory and the response variables are caused by changes in lurking variables. **Confounding** of two variables (either explanatory or lurking variables or both) means that we cannot distinguish their effects on the response variable.

Establishing that an association is due to causation is best accomplished by conducting an **experiment** that changes the explanatory variable while controlling

other influences on the response.

In the absence of experimental evidence, be cautious in accepting claims of causation. Good evidence of causation requires (1) a strong association, (2) that appears consistently in many studies, (3) that has higher doses associated with stronger responses, (4) with the alleged cause preceding the effect in time, and (5) that is plausible.

## **SECTION 2.7 Exercises**

### **2.133 Examples of association.**

Give three examples of association: one due to causation, one due to common response, and one due to confounding. Use your examples to write a short paragraph explaining the differences among these three explanations for an observed association.

### **2.134 The five criteria for establishing causation.**

Consider the five criteria for establishing causation. Explain how each of these, if not established, seriously weakens the case that an association is due to causation.

### **2.135 Iron and anemia.**

A lack of adequate iron in the diet is associated with anemia, a condition in which the body does not have enough red blood cells. However, anemia is also associated with malaria and infections with worms called helminths. Discuss these observed associations using the framework of Figure 2.31.

### **2.136 Stress and lack of sleep in college students.**

Studies of college students have shown that stress and lack of sleep are associated. Do you think that lack of sleep causes stress or that stress causes lack of sleep? Write a short paragraph summarizing your opinions.

### **2.137 Online courses.**

Many colleges offer online versions of some courses that are also taught in the classroom. It often happens that the students who enroll in the online version do better than the classroom students on the course exams. This does not show that online instruction is more effective than classroom teaching, because the people who sign up for online courses are often quite different from the classroom students. Suggest some student characteristics that you think could be confounded with online versus classroom. Use a diagram like Figure 2.31(c) to illustrate your ideas.

### **2.138 Marriage and income.**

Data show that men who are married, and also divorced or widowed men, earn quite a bit more than men who have never been married. This does not mean that a man can raise his income by getting married. Suggest several lurking variables that you think are confounded with marital status and that help explain the association between marital status and income. Use a diagram like Figure 2.31(c) to illustrate your

ideas.

### **2.139 Exercise and self-confidence.**

A college fitness center offers an exercise program for staff members who choose to participate. The program assesses each participant's fitness, using a treadmill test, and also administers a personality questionnaire. There is a moderately strong positive correlation between fitness score and score for self-confidence. Is this good evidence that improving fitness increases self-confidence? Explain why or why not.

### **2.140 Computer chip manufacturing and miscarriages.**

A study showed that women who work in the production of computer chips have abnormally high numbers of miscarriages. The union claimed that exposure to chemicals used in production caused the miscarriages. Another possible explanation is that these workers spend most of their work time standing up. Illustrate these relationships in a diagram like one of those in Figure 2.31.

### **2.141 Hospital size and length of stay.**

A study shows that there is a positive correlation between the size of a hospital (measured by its number of beds  $x$ ) and the median number of days  $y$  that patients remain in the hospital. Does this mean that you can shorten a hospital stay by choosing a small hospital? Use a diagram like one of those in Figure 2.31 to explain the association.

### **2.142 Watching TV and low grades.**

Children who watch many hours of television get lower grades in school, on the average, than those who watch less TV. Explain clearly why this fact does not show that watching TV *causes* poor grades. In particular, suggest some other variables that may be confounded with heavy TV viewing and may contribute to poor grades.

### **2.143 Artificial sweeteners.**

People who use artificial sweeteners in place of sugar tend to be heavier than people who use sugar. Does this mean that artificial sweeteners cause weight gain? Give a more plausible explanation for this association.

### **2.144 Exercise and mortality.**

A sign in a fitness center says, "Mortality is halved for men over 65 who walk at least 2 miles a day."

- (a) Mortality is eventually 100% for everyone. What do you think "mortality is halved" means?
- (b) Assuming that the claim is true, explain why this fact does not show that exercise *causes* lower mortality.

### **2.145 Effect of a math skills refresher initiative.**

Students enrolling in an elementary statistics course take a pretest that assesses their math skills. Those

who receive low scores are given the opportunity to take three one-hour refresher sessions designed to review the basic math skills needed for the statistics course. Those who took the refresher sessions performed worse than those who did not on the final exam in the statistics course. Can you conclude that the refresher course has a negative impact on performance in the statistics course? Explain your answer.

## CHAPTER 2 Exercises

### 2.146 Survival and gender on the *Titanic*.

In Exercise 2.124 (page 149) you examined the relationship between survival and class on the *Titanic*. The data file TITANIC contains data on the gender of the *Titanic* passengers. Examine the relationship between survival and gender and write a short summary of your findings. 

TITANIC

### 2.147 Survival, class, and gender on the *Titanic*.

Refer to the previous exercise and Exercise 2.124 (page 149). When we looked at survival and class, we ignored gender. When we looked at survival and gender, we ignored class. Are we missing something interesting about these data when we choose this approach to the analysis? Here is one way to answer this question.  TITANIC

- Create two separate two-way tables. One for survival and class for the women and another for survival and class for the men.
- Perform an analysis of the relationship between survival and class for the women. Summarize your findings.
- Perform an analysis of the relationship between survival and class for the men. Summarize your findings.
- Compare the analyses that you performed in parts (b) and (c). Write a short report on the relationship between survival and the two explanatory variables, class and gender.

### 2.148 Fan loyalty.

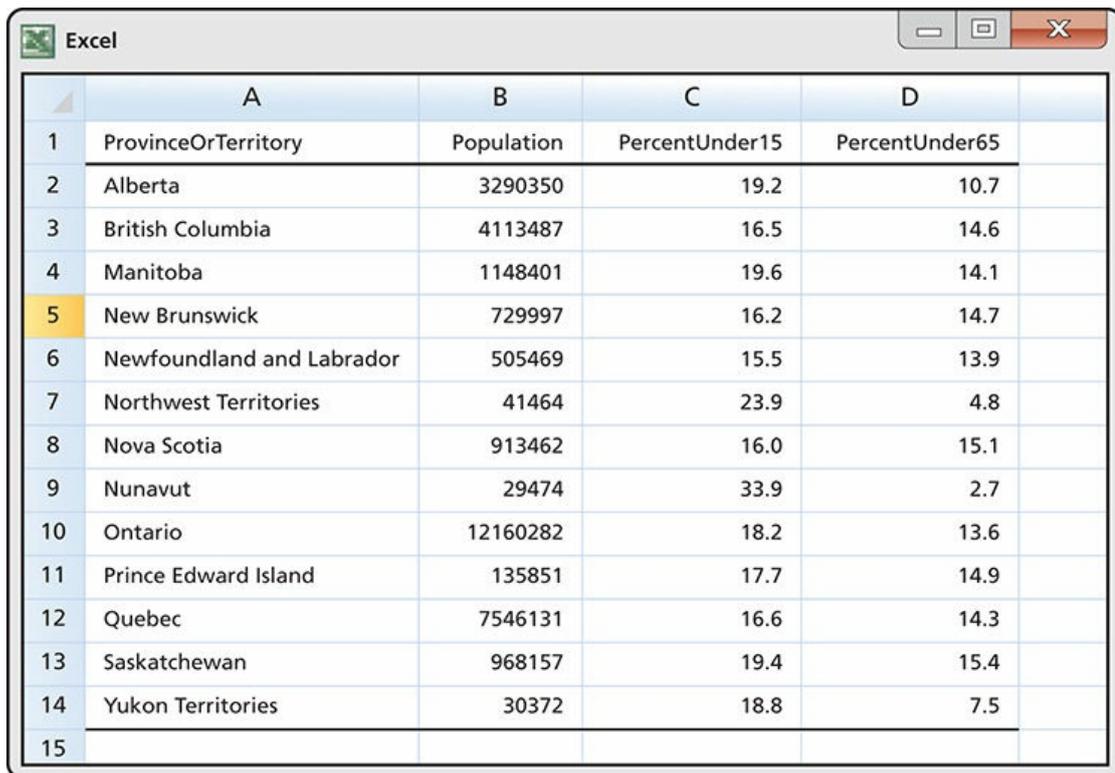
A study of fan loyalty compared Chicago Cubs fans with Arizona Diamondbacks fans. Fans of each team were classified as diehard, more loyal than most, or less loyal than most. A report of the study included the following results. For the Chicago Cubs, 43.3% of the fans were diehards, 34.1% of the fans were more loyal than most, and 22.6% were less loyal than most. For the Arizona Diamondbacks, 26.9% of the fans were diehards, 61.2% of the fans were more loyal than most, and 11.9% were less loyal than most. The report said that there were 115 fans who provided data for the study.<sup>33</sup>

- Write a short summary of what the data presented tell you about the Cubs fans and the Diamondbacks fans. Use graphical and numerical summaries.
- Can you construct a two-way table that could be used to compare fans of these two teams? If yes, construct the table. If no, explain what additional information you would need to construct the table.

### 2.149 Marketing in Canada.

Many consumer items are marketed to particular age groups in a population. To plan such marketing strategies, it is helpful to know the demographic profile for different areas. Statistics Canada provides a great deal of demographic data organized in different ways.<sup>34</sup> Figure 2.32 gives the percent of the population over 65 years and the percent under 15 years for each of the 13 Canadian provinces and territories. Figure 2.33 is a scatterplot of the percent of the population over 65 versus the percent under 15. 

(a) Write a short paragraph explaining what the plot tells you about these two demographic groups in the 13 Canadian provinces and territories.



	A	B	C	D
1	ProvinceOrTerritory	Population	PercentUnder15	PercentUnder65
2	Alberta	3290350	19.2	10.7
3	British Columbia	4113487	16.5	14.6
4	Manitoba	1148401	19.6	14.1
5	New Brunswick	729997	16.2	14.7
6	Newfoundland and Labrador	505469	15.5	13.9
7	Northwest Territories	41464	23.9	4.8
8	Nova Scotia	913462	16.0	15.1
9	Nunavut	29474	33.9	2.7
10	Ontario	12160282	18.2	13.6
11	Prince Edward Island	135851	17.7	14.9
12	Quebec	7546131	16.6	14.3
13	Saskatchewan	968157	19.4	15.4
14	Yukon Territories	30372	18.8	7.5
15				

**FIGURE 2.32**

Percent of the population over 65 years and percent of the population under 15 years in the 13 Canadian provinces and territories, for Exercise 2.149.

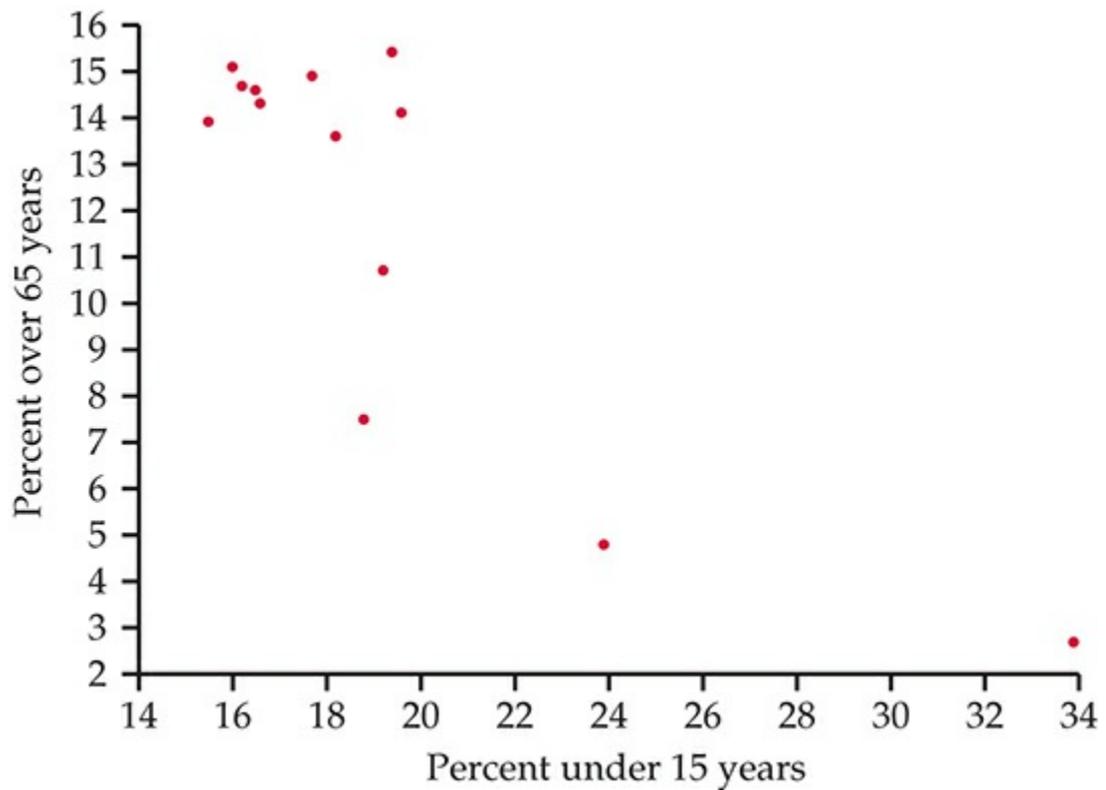
(b) Find the correlation between the percent of the population over 65 and the percent under 15. Does the correlation give a good numerical summary of the strength of this relationship? Explain your answer.

### 2.150 Nunavut.

Refer to the previous exercise and Figures 2.32 and 2.33. 

- (a) Do you think that Nunavut is an outlier?
- (b) Make a residual plot for these data. Comment on the size of the residual for Nunavut. Use this information to expand on your answer to part (a).
- (c) Find the value of the correlation without Nunavut. How does this compare with the value you computed in part (b) of the previous exercise?
- (d) Write a short paragraph about Nunavut based on what you have found in this exercise and the

previous one.



**FIGURE 2.33**

Scatterplot of percent of the population over 65 years versus percent of the population under 15 years for the 13 Canadian provinces and territories, for Exercise 2.149.

**TABLE 2.4 Dwelling Permits, Sales, and Production for 21 European Countries**

Country	Dwelling permits	Sales	Production
Australia	116	137	109
Belgium	125	105	112
Canada	224	122	101
Czech Republic	178	134	162
Denmark	121	126	109
Finland	105	136	125
France	145	121	104
Germany	54	100	119
Greece	117	136	102
Hungary	109	140	155
Ireland	92	123	144
Japan	86	99	109
Korea (South)	158	110	156
Luxembourg	145	161	118
Netherlands	160	107	109
New Zealand	127	139	112
Norway	125	136	94

Poland	163	139	159
Portugal	53	112	105
Spain	122	123	108
Sweden	180	142	116

### 2.151 Compare the provinces with the territories.

Refer to the previous exercise. The three Canadian territories are the Northwest Territories, Nunavut, and the Yukon Territories. All the other entries in Figure 2.32 are provinces.  CANADAP

- Generate a scatterplot of the Canadian demographic data similar to Figure 2.33 but with the points labeled “P” for provinces and “T” for territories.
- Use your new scatterplot to write a new summary of the demographics for the 13 Canadian provinces and territories.

### 2.152 Dwelling permits and sales for 21 European countries.

The Organisation for Economic Co-operation and Development collects data on Main Economic Indicators (MEIs) for many countries. Each variable is recorded as an index with the year 2000 serving as a base year. This means that the variable for each year is reported as a ratio of the value for the year divided by the value for 2000. Use of indices in this way makes it easier to compare values for different countries. Table 2.4 gives the values of three MEIs for 21 countries.<sup>35</sup>  MEIS

- Make a scatterplot with sales as the response variable and permits issued for new dwellings as the explanatory variable. Describe the relationship. Are there any outliers or influential observations?
- Find the least-squares regression line and add it to your plot.
- What is the predicted value of sales for a country that has an index of 160 for dwelling permits?
- The Netherlands has an index of 160 for dwelling permits. Find the residual for this country.
- What percent of the variation in sales is explained by dwelling permits?

### 2.153 Dwelling permits and production.

Refer to the previous exercise.  MEIS

- Make a scatterplot with production as the response variable and permits issued for new dwellings as the explanatory variable. Describe the relationship. Are there any outliers or influential observations?
- Find the least-squares regression line and add it to your plot.
- What is the predicted value of production for a country that has an index of 160 for dwelling permits?
- The Netherlands has an index of 160 for dwelling permits. Find the residual for this country.
- What percent of the variation in production is explained by dwelling permits? How does this value compare with the value that you found in the previous exercise for the percent of variation in

sales that is explained by building permits?

## 2.154 Sales and production.

Refer to the previous two exercises.  **MEIS**

- Make a scatterplot with sales as the response variable and production as the explanatory variable. Describe the relationship. Are there any outliers or influential observations?
- Find the least-squares regression line and add it to your plot.
- What is the predicted value of sales for a country that has an index of 125 for production?
- Finland has an index of 125 for production. Find the residual for this country.
- What percent of the variation in sales is explained by production? How does this value compare with the percents of variation that you calculated in the two previous exercises?

## 2.155 Remote deposit capture.

The Federal Reserve has called remote deposit capture (RDC) “the most important development the [U.S.] banking industry has seen in years.” This service allows users to scan checks and to transmit the scanned images to a bank for posting.<sup>36</sup> In its annual survey of community banks, the American Bankers Association asked banks whether or not they offered this service.<sup>37</sup> Here are the results classified by the asset size (in millions of dollars) of the bank:  **RDCSIZE**

Asset size	Offer RDC	
	Yes	No
Under \$100	63	309
\$101 to \$200	59	132
\$201 or more	112	85

Summarize the results of this survey question numerically and graphically. Write a short paragraph explaining the relationship between the size of a bank, measured by assets, and whether or not RDC is offered.

## 2.156 How does RDC vary across the country?

The survey described in the previous exercise also classified community banks by region. Here is the  $6 \times 2$  table of counts:<sup>38</sup>  **RDCREG**

Region size	Offer RDC	
	Yes	No
Northeast	28	38
Southeast	57	61
Central	53	84
Midwest	63	181
Southwest	27	51
West	61	76

Summarize the results of this survey question numerically and graphically. Write a short paragraph explaining the relationship between the location of a bank and whether or not RDC is offered.

## 2.157 Fields of study for college students.

The following table gives the number of students (in thousands) graduating from college with degrees in several fields of study for seven countries:<sup>39</sup>  FOS

Field of study	Canada	France	Germany	Italy	Japan	U.K.	U.S.
Social sciences, business, law	64	153	66	125	250	152	878
Science, mathematics, engineering	35	111	66	80	136	128	355
Arts and humanities	27	74	33	42	123	105	397
Education	20	45	18	16	39	14	167
Other	30	289	35	58	97	76	272

- Calculate the marginal totals and add them to the table.
- Find the marginal distribution of country and give a graphical display of the distribution.
- Do the same for the marginal distribution of field of study.

## 2.158 Fields of study by country for college students.

In the previous exercise you examined data on fields of study for graduating college students from seven countries.  FOS

- Find the seven conditional distributions giving the distribution of graduates in the different fields of study for each country.
- Display the conditional distributions graphically.
- Write a paragraph summarizing the relationship between field of study and country.

## 2.159 Graduation rates.

One of the factors used to evaluate undergraduate programs is the proportion of incoming students who graduate. This quantity, called the graduation rate, can be predicted by other variables such as the SAT or ACT scores and the high school records of the incoming students. One of the components that *U.S. News & World Report* uses when evaluating colleges is the difference between the actual graduation rate and the rate predicted by a regression equation.<sup>40</sup> In this chapter, we call this quantity the residual. Explain why the residual is a better measure to evaluate college graduation rates than the raw graduation rate.

## 2.160 Popularity of a first name.

The Social Security Administration maintains lists of the top 1000 names for boys and girls born each year since 1879.<sup>41</sup> The name “Atticus” made the list in five recent years. Here are the ranks for those years:  ATTICUS

Year	2004	2005	2006	2007	2008	2009	2010	2011

Rank	937	792	768	685	686	608	558	462
------	-----	-----	-----	-----	-----	-----	-----	-----

- (a) Plot rank versus year.
- (b) Find the equation of the least-squares regression line and add it to your plot.
- (c) Do these data suggest that the name “Atticus” has become more popular, less popular, or stayed the same in popularity over this period of time? Give reasons for your answer.

### 2.161 You select the name.

Refer to the previous exercise. Choose a first name and find the rank of this name for the past several years from the Social Security website, [ssa.gov/OACT/babynames](http://ssa.gov/OACT/babynames). Answer the questions from the previous exercise for this name.

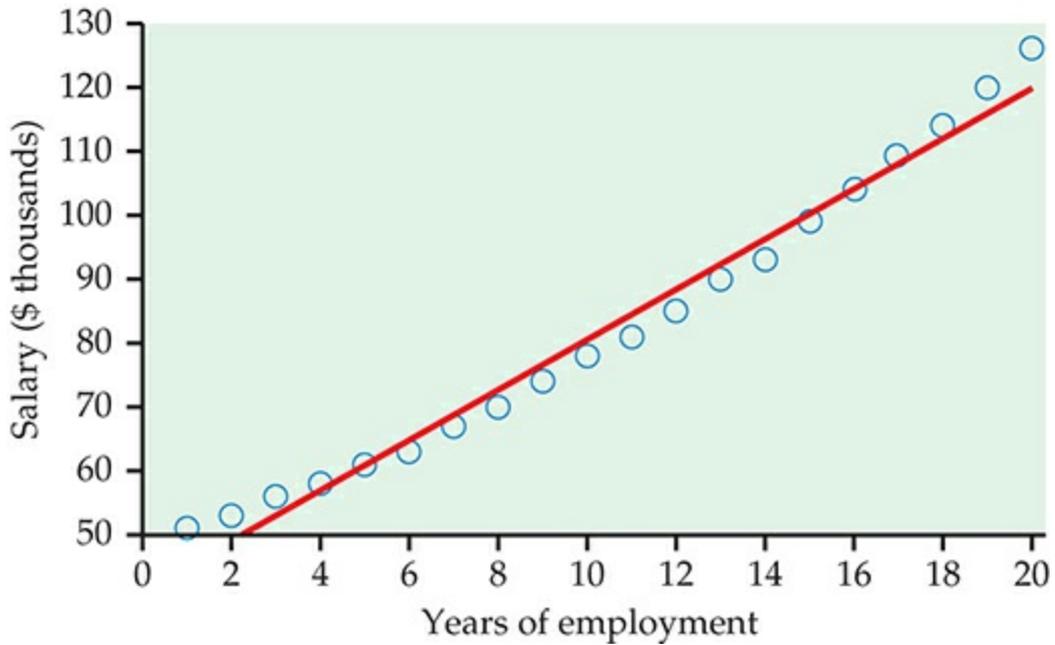
### 2.162 Salaries and raises.

For this exercise we consider a hypothetical employee who starts working in Year 1 with a salary of \$50,000. Each year her salary increases by approximately 5%. By Year 20, she is earning \$126,000.

The following table gives her salary for each year (in thousands of dollars):  RAISES

Year	Salary	Year	Salary	Year	Salary	Year	Salary
1	50	6	63	11	81	16	104
2	53	7	67	12	85	17	109
3	56	8	70	13	90	18	114
4	58	9	74	14	93	19	120
5	61	10	78	15	99	20	126

- (a) Figure 2.34 is a scatterplot of salary versus year, with the least-squares regression line. Describe the relationship between salary and year for this person.
- (b) The value of  $r^2$  for these data is 0.9832. What percent of the variation in salary is explained by year? Would you say that this is an indication of a strong linear relationship? Explain your answer.



**FIGURE 2.34**

Plot of salary versus year for an individual who receives approximately a 5% raise each year for 20 years, with the least-squares regression line, for Exercise 2.162.

### 2.163 Look at the residuals.

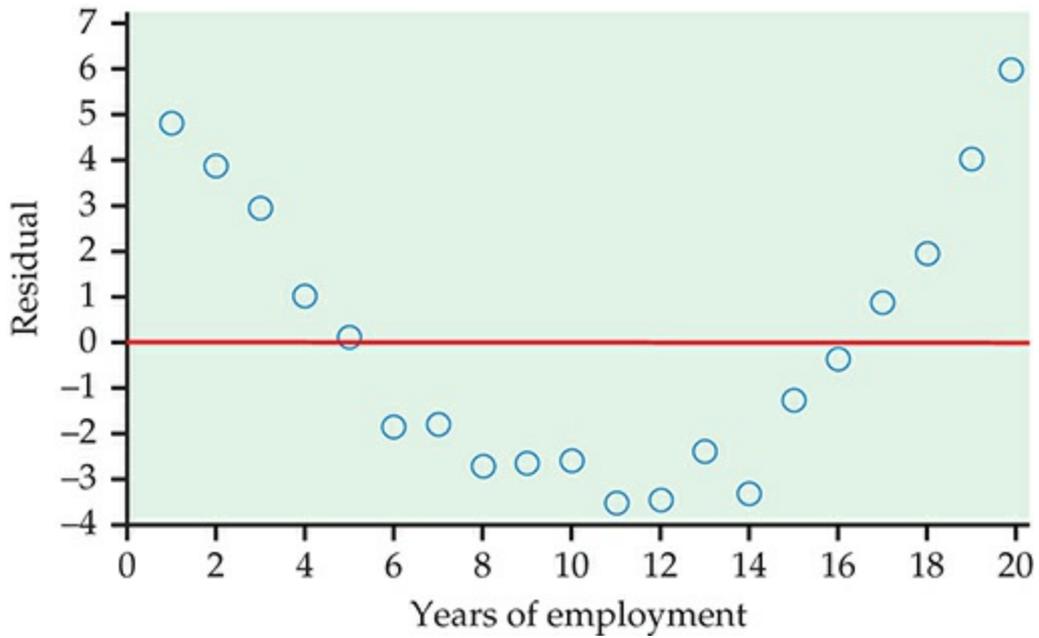
Refer to the previous exercise. Figure 2.35 is a plot of the residuals versus year.  RAISES

(a) Interpret the residual plot.

(b) Explain how this plot highlights the deviations from the least-squares regression line that you can see in Figure 2.34.

### 2.164 Try logs.

Refer to the previous two exercises. Figure 2.36 is a scatterplot with the least-squares regression line for log salary versus year. For this model,  $r^2 = 0.9995$ .  RAISES

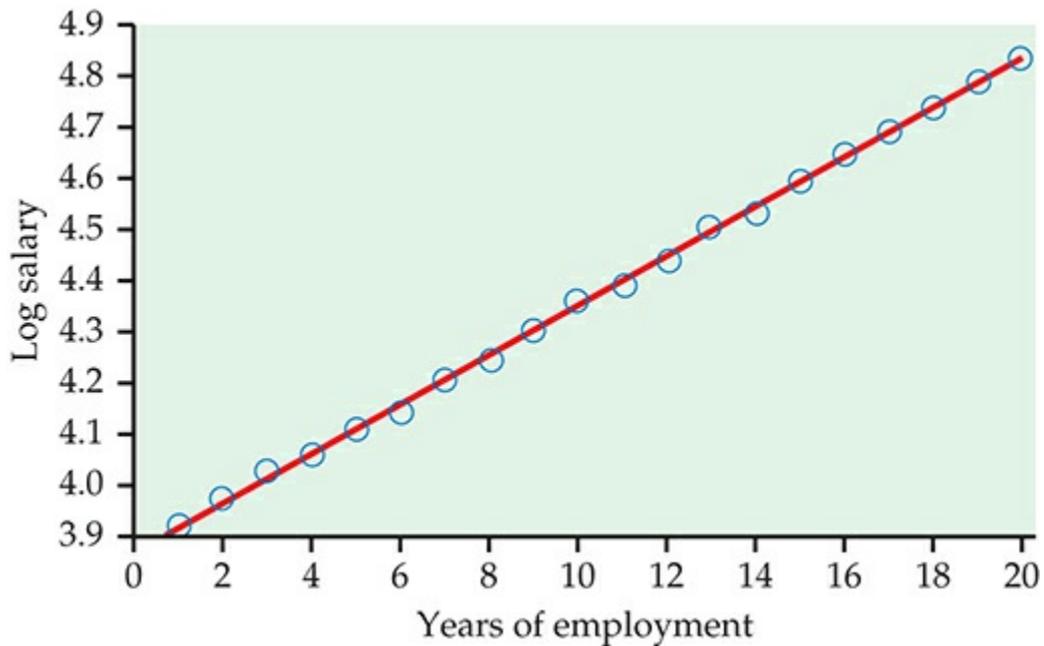


**FIGURE 2.35**

Plot of residuals versus year for an individual who receives approximately a 5% raise each year for 20 years, for Exercise 2.163.

(a) Compare this plot with Figure 2.34. Write a short summary of the similarities and the differences.

(b) Figure 2.37 is a plot of the residuals for the model using year to predict log salary. Compare this plot with Figure 2.35 and summarize your findings.



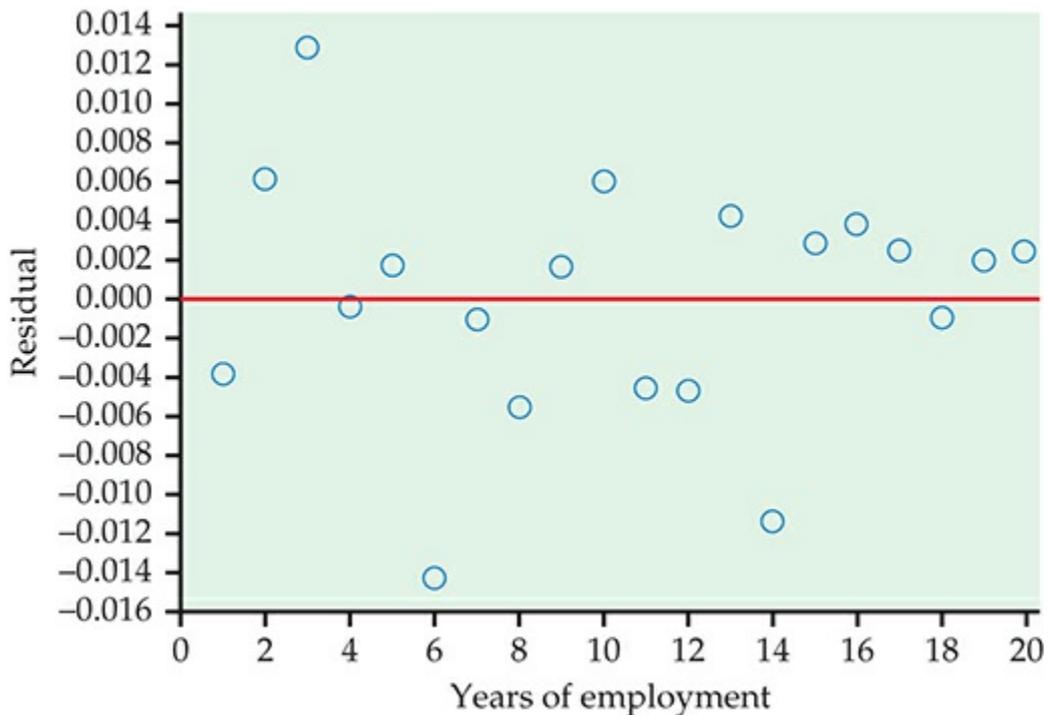
**FIGURE 2.36**

Plot of log salary versus year for an individual who receives approximately a 5% raise each year for 20 years, with the least-squares regression line, for Exercise 2.164.

**2.165 Make some predictions.**

The individual whose salary we have been studying wants to do some financial planning. Specifically, she would like to predict her salary 5 years into the future, that is, for Year 25. She is willing to assume that her employment situation will be stable for the next 5 years and that it will be similar to the last 20 years.  **RAISES**

(a) Predict her salary for Year 25 using the least-squares regression equation constructed to predict salary from year.



**FIGURE 2.37**

Plot of residuals, based on log salary, versus year for an individual who receives approximately a 5% raise each year for 20 years, for Exercise 2.164.

(b) Predict her salary for Year 25 using the least-squares regression equation constructed to predict log salary from year. Note that you will need to take the predicted log salary and convert this value back to the predicted salary. Many calculators have a function that will perform this operation.

(c) Which prediction do you prefer? Explain your answer.

(d) Someone looking at the numerical summaries and not the plots for these analyses says that because both models have very high values of  $r^2$ , they should perform equally well in doing this prediction. Write a response to this comment.

(e) Discuss the value of graphical summaries and the problems of extrapolation using what you have learned in studying these salary data.

### 2.166 Faculty salaries.

Here are the salaries for a sample of professors in a mathematics department at a large midwestern university for the academic years 2012–2013 and 2013–2014.  **FACULTY**

2012–2013 salary (\$)	2013–2014 salary (\$)	2012–2013 salary (\$)	2013–2014 salary (\$)
146,600	147,700	139,650	142,350

115,800	118,600	135,160	138,485
112,000	115,500	77,792	82,072
101,700	105,800	76,000	82,000
115,000	117,180	85,500	88,700
114,790	117,240	144,850	147,830
106,500	111,100	125,506	128,906
152,000	156,080	118,100	121,200

- (a) Construct a scatterplot with the 2013–2014 salaries on the vertical axis and the 2012–2013 salaries on the horizontal axis.
- (b) Comment on the form, direction, and strength of the relationship in your scatterplot.
- (c) What proportion of the variation in 2013–2014 salaries is explained by 2012–2013 salaries?

### 2.167 Find the line and examine the residuals.

Refer to the previous exercise.  FACULTY

- (a) Find the least-squares regression line for predicting 2013–2014 salaries from 2012–2013 salaries.
- (b) Analyze the residuals, paying attention to any outliers or influential observations. Write a summary of your findings.

### 2.168 Bigger raises for those earning less.

Refer to the previous two exercises. The 2012–2013 salaries do an excellent job of predicting the 2013–2014 salaries. Is there anything more that we can learn from these data? In this department there is a tradition of giving higher-than-average percent raises to those whose salaries are lower.

Let's see if we can find evidence to support this idea in the data.  FACULTY

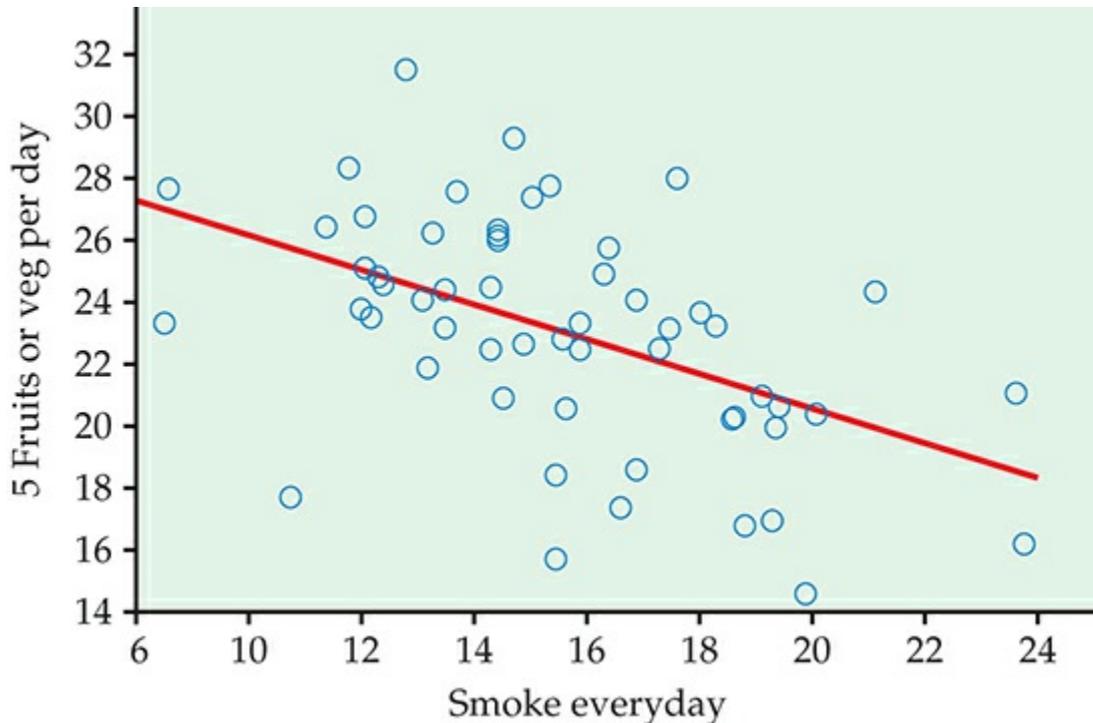
- (a) Compute the percent raise for each faculty member. Take the difference between the 2013–2014 salary and the 2012–2013 salary, divide by the 2012–2013 salary, and then multiply by 100. Make a scatterplot with raise as the response variable and the 2012–2013 salary as the explanatory variable. Describe the relationship that you see in your plot.
- (b) Find the least-squares regression line and add it to your plot.
- (c) Analyze the residuals. Are there any outliers or influential cases? Make a graphical display and include this in a short summary of your conclusions.
- (d) Is there evidence in the data to support the idea that greater percent raises are given to those with lower salaries? Include numerical and graphical summaries to support your conclusion.

### 2.169 Firefighters and fire damage.

Someone says, “There is a strong positive correlation between the number of firefighters at a fire and the amount of damage the fire does. So sending lots of firefighters just causes more damage.” Explain why this reasoning is wrong.

### 2.170 Eating fruits and vegetables and smoking.

The Centers for Disease Prevention and Control Behavior Risk Factor Surveillance System (BRFSS) collects data related to health conditions and risk behaviors.<sup>42</sup> Aggregated data by state are in the BRFSS data file. Figure 2.38 is a plot of two of the BRFSS variables. “5 Fruits or veg per day” is the percent of adults in the state who report eating at least five servings of fruits or vegetables per day; “Smoke everyday” is the percent who smoke every day.  BRFSS



**FIGURE 2.38**

Fruit and vegetable consumption versus smoking, with least-squares regression line, for Example 2.170.

- Describe the relationship between “5 Fruits or veg per day” and “Smoke everyday.” Explain why you might expect this type of association.
- Find the correlation between the two variables.
- For Utah, 23.3% eat at least five servings of fruits or vegetables per day and 8.5% smoke every day. Find Utah on the plot and describe its position relative to the other states.
- For California, the percents are 27.7% for fruits or vegetables and 8.6% for smoking. Find California on the plot and describe its position relative to the other states.
- Pick your favorite state and write a short summary of its position relative to states that you would consider to be similar. Then use Table 2.5 to determine if your idea is supported by the data. Summarize your results.



### 2.171 Education and eating fruits and vegetables.

Refer to the previous exercise. The BRFSS data file contains a variable called EdCollege, the proportion of adults who have completed college.  BRFSS

- Plot the data with 5 Fruits and vegetables per day on the  $x$  axis and EdCollege on the  $y$  axis. Describe the overall pattern of the data.

(b) Add the least-squares regression line to your plot. Does the line give a summary of the overall pattern? Explain your answer.

State	Fruits & vegetables (%)	Smoking (%)	State	Fruits & vegetables (%)	Smoking (%)
Alabama	20.3	18.6	Montana	25.7	16.4
Alaska	23.4	15.9	Nebraska	20.9	14.5
Arizona	24.1	13.1	Nevada	23.7	18.0
Arkansas	20.4	20.1	New Hampshire	27.9	15.4
California	27.7	8.6	New Jersey	26.4	11.4
Colorado	24.8	12.3	New Mexico	23.2	13.5
Connecticut	28.3	11.8	New York	26.8	12.1
Delaware	25.0	16.3	North Carolina	20.6	15.6
District of Columbia	31.5	12.8	North Dakota	22.5	15.9
Florida	24.4	13.5	Ohio	21.0	19.1
Georgia	24.5	14.3	Oklahoma	14.6	19.9
Guam	24.3	21.1	Oregon	26.3	14.4
Hawaii	23.5	12.2	Pennsylvania	24.1	16.9
Idaho	24.6	12.4	Puerto Rico	17.7	10.8
Illinois	22.5	14.3	Rhode Island	26.1	14.4
Indiana	20.6	19.4	South Carolina	17.4	16.6
Iowa	18.5	15.5	South Dakota	15.7	15.5
Kansas	18.6	16.9	Tennessee	23.3	18.3
Kentucky	21.1	23.6	Texas	23.8	12.0
Louisiana	16.9	19.3	Utah	23.3	8.5
Maine	28.0	17.6	Vermont	29.3	14.7
Maryland	27.6	13.7	Virginia	27.3	15.0
Massachusetts	26.2	13.3	Washington	25.1	12.1
Michigan	22.6	17.3	West Virginia	16.2	23.8
Minnesota	21.9	13.2	Wisconsin	22.7	14.9
Mississippi	16.8	18.8	Wyoming	23.3	17.5
Missouri	19.9	19.4			

(c) Pick out a few states and use their position in the graph to write a short summary of how they compare with other states.

(d) Can you conclude that earning a college degree will cause you to eat five servings of fruits and vegetables per day? Explain your answer.

## 2.172 Predicting text pages.

The editor of a statistics text would like to plan for the next edition. A key variable is the number of pages that will be in the final version. Text files are prepared by the authors using a word processor called LaTeX, and separate files contain figures and tables. For the previous edition of the text, the number of pages in the LaTeX files can easily be determined, as well as the number of pages in the final version of the text. Here are the data:  **TEXTP**

Chapter	1	2	3	4	5	6	7	8	9	10	11	12	13
LaTeX pages	77	73	59	80	45	66	81	45	47	43	31	46	26
Text pages	99	89	61	82	47	68	87	45	53	50	36	52	19

- Plot the data and describe the overall pattern.
- Find the equation of the least-squares regression line and add the line to your plot.
- Find the predicted number of pages for the next edition if the number of LaTeX pages is 62.
- Write a short report for the editor explaining to her how you constructed the regression equation and how she could use it to estimate the number of pages in the next edition of the text.

## 2.173 Plywood strength.

How strong is a building material such as plywood? To be specific, support a 24-inch by 2-inch strip of plywood at both ends and apply force in the middle until the strip breaks. The modulus of rupture (MOR) is the force needed to break the strip. We would like to be able to predict MOR without actually breaking the wood. The modulus of elasticity (MOE) is found by bending the wood without breaking it. Both MOE and MOR are measured in pounds per square inch. Here are data for 32 specimens of the same type of plywood:<sup>43</sup>  **MOEMOR**

MOE	MOR	MOE	MOR	MOE	MOR	MOE	MOR
2,005,400	11,591	1,774,850	10,541	2,181,910	12,702	1,747,010	11,794
1,166,360	8,542	1,457,020	10,314	1,559,700	11,209	1,791,150	11,413
1,842,180	12,750	1,959,590	11,983	2,372,660	12,799	2,535,170	13,920
2,088,370	14,512	1,720,930	10,232	1,580,930	12,062	1,355,720	9,286
1,615,070	9,244	1,355,960	8,395	1,879,900	11,357	1,646,010	8,814
1,938,440	11,904	1,411,210	10,654	1,594,750	8,889	1,472,310	6,326
2,047,700	11,208	1,842,630	10,223	1,558,770	11,565	1,488,440	9,214
2,037,520	12,004	1,984,690	13,499	2,212,310	15,317	2,349,090	13,645

Can we use MOE to predict MOR accurately? Use the data to write a discussion of this question.

## 2.174 Distribution of the residuals.

Some statistical methods require that the residuals from a regression line have a Normal distribution. The residuals for the nonexercise activity example are given in Exercise 2.93 (page 128). Is their distribution close to Normal? Make a Normal quantile plot to find out.

## 2.175 An example of Simpson's paradox.

Mountain View University has professional schools in business and law. Here is a three-way table of applicants to these professional schools, categorized by gender, school, and admission decision:<sup>44</sup>

 **ADMITS**

<b>Business</b>			<b>Law</b>		
	<b>Admit</b>			<b>Admit</b>	
<b>Gender</b>	<b>Yes</b>	<b>No</b>	<b>Gender</b>	<b>Yes</b>	<b>No</b>
Male	400	200	Male	90	110
Female	200	100	Female	200	200

(a) Make a two-way table of gender by admission decision for the combined professional schools by summing entries in the three-way table.

(b) From your two-way table, compute separately the percents of male and female applicants admitted. Male applicants are admitted to Mountain View's professional schools at a higher rate than female applicants.

(c) Now compute separately the percents of male and female applicants admitted by the business school and by the law school.

(d) Explain carefully, as if speaking to a skeptical reporter, how it can happen that Mountain View appears to favor males when this is not true within each of the professional schools.

## 2.176 Construct an example with four schools.

Refer to the previous exercise. Make a similar table that illustrates the same point for a hypothetical university having four different schools. Carefully summarize your table with the appropriate percents.

## 2.177 Class size and class level.

A university classifies its classes as either "small" (fewer than 40 students) or "large." A dean sees that 62% of Department A's classes are small, while Department B has only 40% small classes. She wonders if she should cut Department A's budget and insist on larger classes. Department A responds to the dean by pointing out that classes for third- and fourth-year students tend to be smaller than classes for first- and second-year students. The following three-way table gives the counts of classes by department, size, and student audience. Write a short report for the dean that summarizes these data. Start by computing the percents of small classes in the two departments and include other numerical and graphical comparisons as needed. Here are the numbers of classes to be analyzed:  **CSIZE**

<b>Year</b>	<b>Department A</b>			<b>Department B</b>		
	<b>Large</b>	<b>Small</b>	<b>Total</b>	<b>Large</b>	<b>Small</b>	<b>Total</b>
First	2	0	2	18	2	20
Second	9	1	10	40	10	50
Third	5	15	20	4	16	20
Fourth	4	16	20	2	14	16

## 2.178 Health conditions and risk behaviors.

The data file BRFSS gives several variables related to health conditions and risk behaviors as well as demographic information for the 50 states, the District of Columbia, Guam, and Puerto Rico. Pick at least three pairs of variables to analyze. Write a short report on your findings.  **BRFSS**

# 3 Producing Data

CHAPTER



- 3.1 Sources of Data
- 3.2 Design of Experiments
- 3.3 Sampling Design
- 3.4 Toward Statistical Inference
- 3.5 Ethics

# Introduction



In Chapters 1 and 2 we learned some basic tools of *data analysis*. We used graphs and numbers to describe data. When we do **exploratory data analysis**, we rely heavily on plotting the data. We look for patterns that suggest interesting conclusions or questions for further study. However, *exploratory analysis alone can rarely provide convincing evidence for its conclusions, because striking patterns that we find in data can arise from many sources.*

## exploratory data analysis

The validity of the conclusions that we draw from an analysis of data depends not only on the use of the best methods to perform the analysis but also on the quality of the data. Therefore, Section 3.1 begins this chapter with a short overview on sources of data.

The two main sources for quality data are designed experiments and sample surveys. We study these two sources in Sections 3.2 and 3.3, respectively.

Statistical techniques for producing data are the foundation for **statistical inference**, which answers specific questions with a known degree of confidence. In Section 3.4, we discuss some basic ideas related to inference.

## statistical inference

Should an experiment or sample survey that could possibly provide interesting and important information always be performed? How can we safeguard the privacy of subjects in a sample survey? What constitutes the mistreatment of people or animals who are studied in an experiment? These are questions of **ethics**. In Section 3.5, we address ethical issues related to the design of studies and the analysis of data.

## ethics

## 3.1 Sources of Data

When you complete this section, you will be able to

- Identify anecdotal data and, using specific examples, explain why they have limited value.
- Identify available data and explain how they can be used in specific examples.
- Identify data collected from sample surveys and explain how they can be used in specific examples.
- Identify data collected from experiments and explain how they can be used in specific examples.
- Distinguish data that are from experiments, from observational studies that are sample surveys, and from observational studies that are not sample surveys.
- Identify the treatment in an experiment.

There are many sources of data. Some data are very easy to collect but they may not be very useful. Other data require careful planning and need professional staff to gather. These can be much more useful. Whatever the source, a good statistical analysis will start with a careful study of the source of the data. Here is one type of source.

### Anecdotal data

It is tempting to simply draw conclusions from our own experience, making no use of more broadly representative data. A magazine article about Pilates says that men need this form of exercise even more than women do. The article describes the benefits that two men received from taking Pilates classes. A newspaper ad states that a particular brand of windows is “considered to be the best” and says that “now is the best time to replace your windows and doors.” These types of stories, or *anecdotes*, sometimes provide quantitative data. However, this type of data does not give us a sound basis for drawing conclusions.

#### ANECDOTAL DATA

**Anecdotal data** represent individual cases, which often come to our attention

because they are striking in some way. These cases are not necessarily representative of any larger group of cases.

## USE YOUR KNOWLEDGE

### 3.1 Do flu shots work?

A friend tells you that she received a flu shot and then got the flu. Can you conclude that flu shots don't work? Explain your answer.

### 3.2 Describe an anecdote.

Find an example from some recent experience where anecdotal evidence was used to draw a conclusion that is not justified. Describe the example and explain why it should not be used in this way.

### 3.3 I didn't do it.

A professional athlete is accused of using performance-enhancing drugs. He calls a news conference and denies the charges. Is this sufficient information to conclude that he did not use performance-enhancing drugs? Explain your answer.

### 3.4 Are all vehicles this good?

A friend has driven a Toyota Camry for more than 200,000 miles and with only the usual service maintenance expenses. Explain why not all Camry owners can expect this kind of performance.

Not all anecdotal data are bad. The experiences of an individual or a small group of individuals might suggest an interesting study that could be performed using more carefully collected data.

## Available data

Occasionally, data are collected for a particular purpose but can also serve as the basis for drawing sound conclusions about other research questions. We use the term *available data* for this type of data.

## AVAILABLE DATA

**Available data** are data that were produced for some other purpose but that may help answer a question of interest.

The library and the Internet can be good sources of available data. Because producing new data is expensive, we all use available data whenever possible. Here are two examples.

### Example

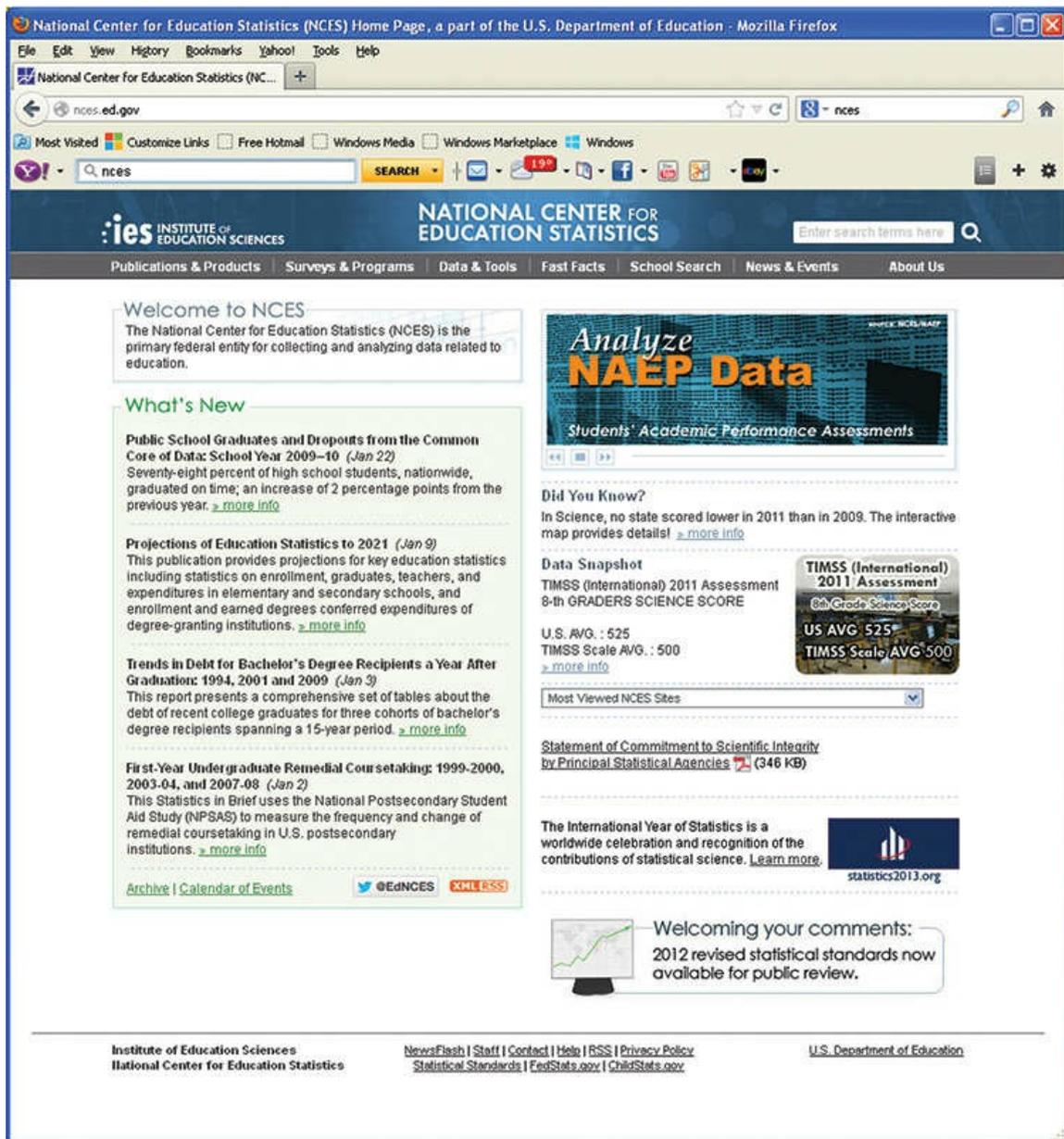
#### 3.1 Wages of U.S. workers.

If you visit the U.S. Bureau of Labor Statistics website, **bls.gov**, you will find many interesting sets of data and statistical summaries. One recent study reported that wages and salary for workers in the United States averaged \$20.36 per hour, and benefits averaged \$8.58 per hour.

### Example

#### 3.2 Math skills.

At the website of the National Center for Education Statistics, **nces.ed.gov**, you will find full details about the math skills of schoolchildren as determined by the latest National Assessment of Educational Progress (Figure 3.1). Mathematics scores have slowly but steadily increased since 1990. All racial/ethnic groups, both boys and girls, and students in most states are getting better in math.



**FIGURE 3.1**

Websites of government statistical offices are prime sources of data. Here is a page from the National Center for Education Statistics website.

Many nations have a single national statistical office, such as Statistics Canada ([statcan.gc.ca](http://statcan.gc.ca)) and Mexico's INEGI ([www.inegi.org.mx](http://www.inegi.org.mx)). More than 70 different U.S. agencies collect data. You can reach most of them through the U.S. government's FedStats site ([fedstats.gov](http://fedstats.gov)).

## USE YOUR KNOWLEDGE

### 3.5 What more do you need?

A website claims that millennial generation consumers are very loyal to the brands that they prefer. What additional information do you need to evaluate this claim?

A survey of college athletes is designed to estimate the percent who gamble. Do restaurant patrons give higher tips when their server repeats their order carefully? The validity of our conclusions from the analysis of data collected to address these issues rests on a foundation of carefully collected data.

In this chapter, we will develop the skills needed to produce trustworthy data and to judge the quality of data produced by others. The techniques for producing data that we will study require no formulas, but they are among the most important ideas in statistics. Statistical designs for producing data rely on either *sampling* or *experiments*.

### Sample surveys and experiments

How have the attitudes of Americans, on issues ranging from abortion to work, changed over time? **Sample surveys** are the usual tool for answering questions like these.

#### Example

##### 3.3 The General Social Survey.

One of the most important sample surveys is the General Social Survey (GSS) conducted by the National Opinion Research Center (NORC), an organization affiliated with the University of Chicago.<sup>1</sup> The GSS interviews about 3000 adult residents of the United States every other year.

The GSS selects a **sample** of adults to represent the larger **population** of all English-speaking adults living in the United States. The idea of *sampling* is to study a part in order to gain information about the whole. Data are often produced by sampling a population of people or things. Opinion polls, for example, may report the views of the entire country based on interviews with a sample of about 1000 people. Government reports on employment and unemployment are produced from a monthly sample of about 60,000 households. The quality of manufactured

items is monitored by inspecting small samples each hour or each shift.

sample

population

## USE YOUR KNOWLEDGE

### 3.6 Check out the General Social Survey.

Visit the General Social Survey website at [www3.norc.org/gss](http://www3.norc.org/gss). Write a short summary of one of their reports, paying particular attention to the methods used to collect the data.

In all our examples, the expense of examining every item in the population makes sampling a practical necessity. Timeliness is another reason for preferring a sample to a **census**, which is an attempt to contact every individual in the population. We want information on current unemployment and public opinion next week, not next year. Moreover, a carefully conducted sample is often more accurate than a census. Accountants, for example, sample a firm's inventory to verify the accuracy of the records. Attempting to count every last item in the warehouse would be not only expensive but also inaccurate. Bored people do not count carefully.

census

If conclusions based on a sample are to be valid for the entire population, a sound design for selecting the sample is required. Sampling designs are the topic of Section 3.3.

A sample survey collects information about a population by selecting and measuring a sample from the population. The goal is a picture of the population, disturbed as little as possible by the act of gathering information. Sample surveys are one kind of *observational study*.

## OBSERVATION VERSUS EXPERIMENT

In an **observational study** we observe individuals and measure variables of interest but do not attempt to influence the responses.

In an **experiment** we deliberately impose some **treatment** on individuals and we observe their responses.

## Example

### 3.4 Baseball players have strong bones in their throwing arms.

A study of young baseball players measured the strength of the bones in their throwing arms. A control group of subjects who were matched with the baseball players based on age were also measured. This is an example of an observational study that is not a sample survey. The study reported that bone strength was 30% higher in the baseball players.<sup>2</sup>

What can we conclude from this study? If you start to play baseball, will you have stronger bones in your throwing arm?

## Example

### 3.5 Is there a cause-and-effect relationship?

Example 3.4 describes an observational study. People choose to participate in baseball or not. Is it possible that those who choose to play baseball have stronger arms than those who do not? The study does not address this question.

We can imagine an experiment that would remove these difficulties. From a large group of subjects, require some to play baseball and forbid the rest from playing. This is an experiment because the treatment (play baseball or not) is imposed on the subjects. Of course, this particular experiment is neither practical nor ethical.

## Example

### 3.6 It's not really about baseball.

Example 3.4 compared the arm bone strengths of baseball players with those of age-matched controls. Although the study tells us something about baseball players, the results are particularly interesting because they suggest that certain kinds of exercise can help us to build strong bones.

## USE YOUR KNOWLEDGE

### 3.7 Available data.

Can available data be from an observational study? Can available data be from an experiment? Explain your answers.

### 3.8 Picky eaters.

A study of 2049 children in grades 4 to 6 in 33 schools recorded their behaviors in the lunchroom. One of the conclusions of the study was that girls discarded more food than boys.<sup>3</sup> Is this an observational study or an experiment? Is it a sample survey? If it is an experiment, what is the treatment? Explain your answers.

### 3.9 Automatic soap dispensers.

A study compared several brands of automatic soap dispensers. For one test, the dispensers were run until their AA batteries failed. The times to failure were compared for the different brands.<sup>4</sup> Is this an observational study or an experiment? Is it a sample survey? If it is an experiment, what is the treatment? Explain your answers.

An observational study, even one based on a carefully chosen sample, is a poor way to determine what will happen if we change something. The best way to see the effects of a change is to do an **intervention**—where we actually impose the change. When our goal is to understand cause and effect, experiments are the only

source of fully convincing data.

### intervention

In Example 3.4, the effect of baseball playing on arm bone strength is **confounded** with (mixed up with) other characteristics of the subjects in the study. Observational studies that examine the effect of a single variable on an outcome can be misleading when the effects of the explanatory variable are confounded with those of other variables. Because experiments allow us to isolate the effects of specific variables, we generally prefer them. Here is an example.

### confounded

## Example

### 3.7 Which web page design sells more?

A company that sells products on the Internet wants to decide which of two possible web page designs to use. During a two-week period they will use both designs and collect data on sales. They randomly select one of the designs to be used on the first day and then alternate the two designs on each of the following days. At the end of this period they compare the sales for the two designs.

Experiments usually require some sort of randomization, as in this example. We begin the discussion of statistical designs for data collection in Section 3.2 with the principles underlying the design of experiments.

## USE YOUR KNOWLEDGE

### 3.10 Software for teaching creative writing.

An educational software company wants to compare the effectiveness of its computer animation for teaching creative writing with that of a textbook presentation. The company tests the creative-writing skills of a number of second-year college students and then randomly divides them

into two groups. One group uses the animation, and the other studies the text. The company retests all the students and compares the increase in creative-writing skills in the two groups. Is this an experiment? Why or why not? What are the explanatory and response variables?

### 3.11 Apples or apple juice?

Food rheologists study different forms of foods and how the form of a food affects how full we feel when we eat it. One study prepared samples of apple juice and samples of apples with the same number of calories. Half of the subjects were fed apples on one day followed by apple juice on a later day; the other half received the apple juice followed by the apples. After eating, the subjects were asked about how full they felt. Is this an experiment? Why or why not? What are the explanatory and response variables?

## SECTION 3.1 Summary

**Anecdotal data** come from stories or reports about cases that do not necessarily represent a larger group of cases.

**Available data** are data that were produced for some other purpose but that may help answer a question of interest.

A **sample survey** collects data from a sample of cases that represent some larger population of cases.

A **census** collects data from all cases in the population of interest.

In an **experiment**, a **treatment** is imposed and the responses are recorded.

**Confounding** occurs when the effects of two or more variables are related in such a way that we need to take care in assigning the effect to one or to the other.

## SECTION 3.1 Exercises

*For Exercises 3.1 to 3.4, see pages 168–169; for Exercise 3.5, see page 170; for Exercise 3.6, see page 171; for Exercises 3.7 to 3.9, see pages 172–173; and for Exercises 3.10 and 3.11, see page 173.*

*In several of the following exercises you are asked to identify the type of data that is described. Possible answers include anecdotal data, available data, observational data that are from sample surveys, observational data that are not from sample surveys, and data that are from experiments. It is possible for some data to be classified in more than one category.*

### 3.12 Not enough tuna.

You like to eat tuna fish sandwiches. Recently you have noticed that there does not seem to be as much tuna as you expect when you open the can. Identify the type of data that this represents and describe how it can or cannot be used to reach a conclusion about the amount of tuna in cans of tuna fish.

### 3.13 More about tuna.

According to a story in *Consumer Reports*, three major producers of canned tuna agreed to pay \$3,300,000 to settle claims in California that the amount of tuna in their cans was less than the amount printed on the label of the cans.<sup>5</sup> What kind of data do you think was used in this situation to convince the producers to pay this amount of money to settle the claims? Explain your answer fully.

### 3.14 Growth of adolescents.

A study was conducted to study the effect of additional milk in the diet of adolescents over a period of 18 months. A control group received no extra milk. Growth rates of total body bone mineral content (TBBMC) over the study period were calculated for each subject. Data for the control group were used to examine the relationship between growth rate of TBBMC and age.

- (a) How would you classify the data used to evaluate the effect of the additional milk in the diet? Explain your answer.
- (b) How would you classify the control group data on growth rate of TBBMC and age for the study of this relationship? Explain your answer.
- (c) Can you classify the variables growth rate of TBBMC and age as explanatory and response? If so, which is the explanatory variable? Give reasons for your answer.

### 3.15 Satisfaction with allocation of concert tickets.

Your college sponsored a concert that sold out.

- (a) After the concert, an article in the student newspaper reported interviews with three students who were unable to get tickets and were very upset with that fact. What kind of data does this represent? Explain your answer.
- (b) A week later the student organization that sponsored the concert set up a website where students could rank their satisfaction with the way that the tickets were allocated using a 5-point scale with values “very satisfied,” “satisfied,” “neither satisfied nor unsatisfied,” “dissatisfied,” and “very dissatisfied.” The website was open to any students who chose to provide their opinion. How would you classify these data? Give reasons for your answer.
- (c) Suppose that the website in part (b) was changed so that only a sample of students from the college were invited by a text message to respond, and those who did not respond within 3 days were sent an additional text message reminding them to respond. How would your answer to part (b) change, if at all?
- (d) Write a short summary contrasting different types of data using your answers to parts (a), (b), and (c) of this exercise.

### 3.16 Does echinacea reduce the severity of the common cold?

In a study designed to evaluate the benefits of taking echinacea when you have a cold, 719 patients were randomly divided into four groups. The groups were (1) no pills, (2) pills that had no echinacea, (3) pills that had echinacea but the subjects did not know whether or not the pills contained echinacea, and (4) pills that had echinacea and the bottle containing the pills stated that the contents included echinacea. The outcome was a measure of the severity of the cold.<sup>6</sup> Identify the type of data collected in this study. Give reasons for your answer.

### **3.17 Are there treatments?**

Refer to Exercises 3.12 to 3.16. For any of these that involve an experiment, describe the treatment that is used.

## 3.2 Design of Experiments

When you complete this section, you will be able to

- Identify experimental units, subjects, treatments, and outcomes for an experiment.
- Identify a comparative experiment.
- Describe a placebo effect in an experiment.
- Identify bias in an experiment.
- Explain the need for a control group in an experiment.
- Explain the need for randomization in an experiment.
- When evaluating an experiment, apply the basic principles of experimental design: compare, randomize, and repeat.
- Use a table of random digits to randomly assign experimental units to treatments in an experiment.
- Use software to randomly assign experimental units to treatments in an experiment.
- Identify a matched pairs design.
- Identify a block design.

An experiment is a study in which we actually do something to people, animals, or objects in order to observe the response. Here is the basic vocabulary of experiments.

### EXPERIMENTAL UNITS, SUBJECTS, TREATMENTS, OUTCOMES

The individuals on which the experiment is done are the **experimental units**. When the units are human beings, they are called **subjects**. Experimental conditions applied to the units are called **treatments**. The **outcomes** are the measured variables that are used to compare the treatments.

Because the purpose of an experiment is to reveal the response of one variable to changes in one or more other variables, the distinction between explanatory and response variables is important. The explanatory variables in an experiment are

often called **factors**. Many experiments study the joint effects of several factors. In such an experiment, each treatment is formed by combining a specific value (often called a **level**) of each of the factors.

factors

level of a factor

## Example

### 3.8 Are smaller class sizes better?

Do smaller classes in elementary school really benefit students in areas such as scores on standard tests, staying in school, and going on to college? We might do an observational study that compares students who happened to be in smaller classes with those who happened to be in larger classes in their early school years. Small classes are expensive, so they are more common in schools that serve richer communities. Students in small classes tend to also have other advantages: their schools have more resources, their parents are better educated, and so on. Confounding makes it impossible to isolate the effects of small classes.

The Tennessee STAR program was an experiment on the effects of class size. It has been called “one of the most important educational investigations ever carried out.” The *subjects* were 6385 students who were beginning kindergarten. Each student was assigned to one of three *treatments*: regular class (22 to 25 students) with one teacher, regular class (22 to 25 students) with a teacher and a full-time teacher’s aide, and small class (13 to 17 students). These treatments are levels of a single *factor*, the type of class. The students stayed in the same type of class for four years, then all returned to regular classes. In later years, students from the small classes had higher scores on the *outcomes*, standard tests. The benefits of small classes were greatest for minority students.<sup>7</sup>

Example 3.8 illustrates the big advantage of experiments over observational studies. **In principle, experiments can give good evidence for causation.** In an experiment, we study the specific factors we are interested in while controlling the effects of **lurking variables**. All the students in the Tennessee STAR program followed the usual curriculum at their schools. Because students were assigned to different class types within their schools, school resources and family backgrounds

were not confounded with class type. The only systematic difference was the type of class. When students from the small classes did better than those in the other two types, we can be confident that class size made the difference.

## Example

### 3.9 Repeated exposure to advertising.



What are the effects of repeated exposure to an advertising message? The answer may depend both on the length of the ad and on how often it is repeated. An experiment investigated this question using undergraduate students as *subjects*. All subjects viewed a 40-minute television program that included ads for a digital camera. Some subjects saw a 30-second commercial; others, a 90-second version. The same commercial was shown either 1, 3, or 5 times during the program.

This experiment has two *factors*: length of the commercial, with 2 levels, and repetitions, with 3 levels. The 6 combinations of one level of each factor form 6 *treatments*. Figure 3.2 shows the layout of the treatments. After viewing the TV program, all the subjects answered questions about their recall of the ad, their attitude toward the camera, and their intention to purchase it. These are the *outcomes*.

Example 3.9 shows how experiments allow us to study the combined effects of more than one factor. The interaction of several factors can produce effects that cannot be predicted from looking at the effects of each factor alone. Perhaps longer commercials increase interest in a product, and more commercials also increase interest, but if we both make a commercial longer and show it more often, viewers get annoyed and their interest in the product drops. The two-factor experiment in Example 3.9 will help us find out.

		Factor B		
		Repetitions		
		1 time	3 times	5 times
Factor A	30 seconds	1	2	3
	90 seconds	4	5	6

**FIGURE 3.2**

The treatments in the study of advertising, for Example 3.9. Combining the levels of the two factors forms six treatments.

## USE YOUR KNOWLEDGE

### 3.18 Does echinacea reduce the severity of the common cold?

In a study designed to evaluate the benefits of taking echinacea when you have a cold, 719 patients were randomly divided into four groups. The groups were (1) no pills, (2) pills that had no echinacea, (3) pills that had echinacea but the subjects did not know whether or not the pills contained echinacea, and (4) pills that had echinacea and the bottle containing the pills stated that the contents included echinacea. The outcome was a measure of the severity of the cold.<sup>8</sup> Identify the experimental units, the treatments, and the outcome. Describe the factor and its levels. The study subjects were aged 12 to 80 years. To what extent do you think the results of this experiment can be generalized to young children?

### 3.19 Can coaching via mobile technology change diet and exercise?

A study was designed to determine the extent to which diet and exercise can be changed through coaching. At the start of the study, the subjects had high saturated-fat intakes, low fruit and vegetable intakes, and low physical activity. Each subject was assigned to one of four coaching groups with different goals: (1) increase fruit and vegetable intake and physical activity, (2) decrease fat intake and sedentary leisure, (3)

decrease fat intake and increase physical activity, and (4) increase fruit and vegetable intake and decrease sedentary leisure. After three weeks of remote coaching, a combined measure of diet and activity improvement was calculated.<sup>9</sup> Explain why this study is an experiment, and identify the experimental units, the treatments, and the response variable. Describe the factor and its levels. What are the outcomes? Of the 204 people who were assigned to mobile coaching, 200 completed the study. Do you think that this fact is important to consider when interpreting the results of the study? Explain your answer.

## Comparative experiments

Laboratory experiments in science and engineering often have a simple design with only a single treatment, which is applied to all experimental units. The design of such an experiment can be outlined as

**Treatment → Observe response**

For example, we may subject a beam to a load (treatment) and measure its deflection (observation). We rely on the controlled environment of the laboratory to protect us from lurking variables. When experiments are conducted outside the laboratory or with living subjects, such simple designs often yield invalid data. That is, we cannot tell whether the response was due to the treatment or to lurking variables.

### Example

#### 3.10 Will writing about it reduce test anxiety?

A study designed to reduce test anxiety had students write an essay about their feelings concerning an upcoming exam.<sup>10</sup> The scores on this exam, the second of the semester, were compared with those on the first exam in the course. The mean scores on the second exam were higher than the mean scores on the first exam.

**Write about feelings → Observe exam scores**

The test anxiety experiment of Example 3.10 was poorly designed to evaluate the effect of the writing exercise. Perhaps exam scores would have increased on the

second exam because the students became more familiar with the exam style of this particular instructor even without the writing exercise. Another possible explanation is that the increase is due to the personal attention that the students received by the person who explained how to write about their feelings regarding the exam.

In medical settings this phenomenon is called the **placebo effect**. In medicine, a placebo is a dummy treatment, such as a sugar pill. People respond favorably to personal attention or to any treatment that they hope will help them. On the other hand, the writing exercise may have been very effective in improving exam scores.

### placebo effect

For this experiment we don't know whether the change was due to writing the essay, to the personal contacts with the study personnel, or to greater familiarity with the way the instructor designed exams.

The test anxiety experiment gave inconclusive results because the effect of writing the essay was confounded with other factors that could have had an effect on exam scores. The best way to avoid confounding is to do a **comparative experiment**. Think about a study in which some students performed the writing exercise and others did not. A comparison of the exam scores of these two groups of students would provide an evaluation of the effect of the writing exercise.

### comparative experiment

In medical settings, it is standard practice to randomly assign patients to either a **control group** or a **treatment group**. All patients are treated the same in every way except that the treatment group receives the product that is being evaluated.

### control group

### treatment group



*Uncontrolled experiments (that is, experiments that don't include a control group) in medicine and the behavioral sciences can be dominated by such influences as the details of the experimental arrangement, the selection of subjects, and the placebo effect. The result is often bias.*

## BIAS

The design of a study is **biased** if it systematically favors certain outcomes.

---

An uncontrolled study of a new medical therapy, for example, is biased in favor of finding the treatment effective because of the placebo effect. Uncontrolled studies in medicine give new therapies a much higher success rate than proper comparative experiments do. Well-designed experiments usually compare several treatments.

## USE YOUR KNOWLEDGE

### 3.20 Are the teacher evaluations biased?

The evaluations of two instructors by their students are compared when it is time to decide raises for the coming year. One teacher always hands out the evaluation forms in class when the grades on the first exam are given to students. The other instructor always hands out the evaluation forms at the end of a class in which a very interesting film clip is shown. Discuss the possibility of bias in this context.

## Randomization

The **design of an experiment** first describes the response variable or variables, the factors (explanatory variables), and the treatments, with comparison as the leading principle. Figure 3.2 (page 177) illustrates this aspect of the design of a study of response to advertising. The second aspect of experimental design is how the experimental units are assigned to the treatments. Comparison of the effects of several treatments is valid only when all treatments are applied to similar groups of experimental units. If one corn variety is planted on more fertile ground, or if one cancer drug is given to more seriously ill patients, comparisons among treatments are meaningless. If groups assigned to treatments are quite different in a comparative experiment, we should be concerned that our experiment will be biased. How can we assign experimental units to treatments in a way that is fair to all treatments?

### experimental design

Experimenters often attempt to match groups by elaborate balancing acts. Medical researchers, for example, try to match the patients in a “new drug” experimental group and a “standard drug” control group by age, sex, physical condition, smoker or not, and so on. Matching is helpful but not adequate—there are too many lurking variables that might affect the outcome. The experimenter is

unable to measure some of these variables and will not think of others until after the experiment.

Some important variables, such as how advanced a cancer patient's disease is, are so subjective that they can't be measured. In other cases, an experimenter might unconsciously bias a study by assigning those patients who seemed the sickest to a promising new treatment in the (unconscious) hope that it would help them.

*The statistician's remedy is to rely on chance to make an assignment that does not depend on any characteristic of the experimental units and that does not rely on the judgment of the experimenter in any way.* The use of chance can be combined with matching, but the simplest experimental design creates groups by chance alone. Here is an example.

## Example

### 3.11 Which smartphone should be marketed?



Two teams have each prepared a prototype for a new smartphone. Before deciding which one will be marketed, the smartphone will be evaluated by college students. Forty students will receive a new phone. They will use it for two weeks and then answer some questions about how well they like the phone. The 40 students will be randomized with 20 receiving each phone.

This experiment has a single factor (prototype) with two levels. The researchers must divide the 40 student subjects into two groups of 20. To do this in a completely unbiased fashion, put the names of the 40 students in a

hat, mix them up, and draw 20. These students will receive Phone 1, and the remaining 20 will receive Phone 2. Figure 3.3 outlines the design of this experiment.

The use of chance to divide experimental units into groups is called **randomization**. The design in Figure 3.3 combines comparison and randomization to arrive at the simplest randomized comparative design. This “flowchart” outline presents all the essentials: randomization, the sizes of the groups and which treatment they receive, and the response variable. There are, as we will see later, statistical reasons for using treatment groups that are about equal in size.

randomization

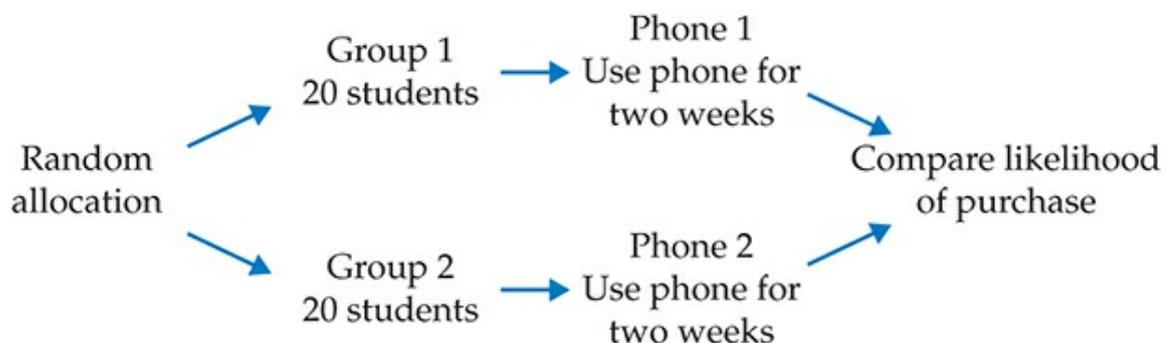
## USE YOUR KNOWLEDGE

### 3.21 Diagram the echinacea experiment.

Refer to Exercise 3.18 (page 177). Draw a diagram similar to Figure 3.3 that describes the experiment.

### 3.22 Diagram the coaching via mobile technology experiment.

Refer to Exercise 3.19 (page 177). Draw a diagram similar to Figure 3.3 that describes the experiment.



**FIGURE 3.3**

Outline of a randomized comparative experiment, for Example 3.11.

## Randomized comparative experiments

The logic behind the randomized comparative design in Figure 3.3 is as follows:

- Randomization produces two groups of subjects that we expect to be similar in all respects before the treatments are applied.
- Comparative design helps ensure that influences other than the characteristics of the smartphone operate equally on both groups.
- Therefore, differences in the satisfaction with the smartphone must be due either to the characteristics of the phone or to the chance assignment of subjects to the two groups.

That “either-or” deserves more comment. We cannot say that *all* the difference in the satisfaction with the two smartphones is caused by the characteristics of the phones. There would be some difference even if both groups used the same phone. Some students would be more likely to be highly favorable of any new phone. Chance can assign more of these students to one of the phones, so that there is a chance difference between the groups. We would not trust an experiment with just one subject in each group, for example. The results would depend too much on which phone got lucky and received the subject who was more likely to be highly satisfied. If we assign many students to each group, however, the effects of chance will average out. There will be little difference in the satisfaction between the two groups unless the phone characteristics causes a difference. “Use enough subjects to reduce chance variation” is the third big idea of statistical design of experiments.

## PRINCIPLES OF EXPERIMENTAL DESIGN

The basic principles of statistical design of experiments are

1. **Compare** two or more treatments. This will control the effects of lurking variables on the response.
2. **Randomize**—use impersonal chance to assign experimental units to treatments.
3. **Repeat** each treatment on many units to reduce chance variation in the results.

### How to randomize



The idea of randomization is to assign subjects to treatments by drawing names from a hat. In practice, experimenters use software to carry out randomization. Most statistical software will choose 20 out of a list of 40 at random, for example.

The list might contain the names of 40 human subjects. The 20 chosen form one group, and the 20 that remain form the second group. The *Simple Random Sample* applet on the text website makes it particularly easy to choose treatment groups at random.

You can randomize without software by using a *table of random digits*. Thinking about random digits helps you to understand randomization even if you will use software in practice. Table B at the back of the book is a table of random digits.

## RANDOM DIGITS

A **table of random digits** is a list of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 that has the following properties:

1. The digit in any position in the list has the same chance of being any one of 0, 1, 2, 3, 4, 5, 6, 7, 8, 9.
2. The digits in different positions are independent in the sense that the value of one has no influence on the value of any other.

You can think of Table B as the result of asking an assistant (or a computer) to mix the digits 0 to 9 in a hat, draw one, then replace the digit drawn, mix again, draw a second digit, and so on. The assistant's mixing and drawing save us the work of mixing and drawing when we need to randomize. Table B begins with the digits 19223950340575628713. To make the table easier to read, the digits appear in groups of five and in numbered rows. The groups and rows have no meaning—the table is just a long list of digits having Properties 1 and 2 described above.

Our goal is to use random digits for experimental randomization. We need the following facts about random digits, which are consequences of Properties 1 and 2:

- Any *pair* of random digits has the same chance of being any of the 100 possible pairs: 00, 01, 02, . . . , 98, 99.
- Any *triple* of random digits has the same chance of being any of the 1000 possible triples: 000, 001, 002, . . . , 998, 999.
- . . . and so on for groups of four or more random digits.

### Example

### 3.12 Randomize the students.

In the smartphone experiment of Example 3.11, we must divide 40 students at random into two groups of 20 students each.

*Step 1: Label.* Give each student a numerical label, using as few digits as possible. Two digits are needed to label 40 students, so we use labels

01, 02, 03, . . . , 39, 40

It is also correct to use labels 00 to 39 or some other choice of 40 two-digit labels.

*Step 2: Table.* Start anywhere in Table B and read two-digit groups. Suppose we begin at line 130, which is

69051 64817 87174 09517 84534 06489 87201 97245

The first 10 two-digit groups in this line are

69 05 16 48 17 87 17 40 95 17

Each of these two-digit groups is a label. The labels 00 and 41 to 99 are not used in this example, so we ignore them. The first 20 labels between 01 and 40 that we encounter in the table choose students for the first phone. Of the first 10 labels in line 130, we ignore four because they are too high (over 40). The others are 05, 16, 17, 17, 40, and 17. The students labeled 05, 16, 17, and 40 will evaluate the first phone. Ignore the second and third 17s because that student is already in the group. Run your finger across line 130 (and continue to the following lines) until you have chosen 20 students. They are the students labeled

05, 16, 17, 40, 20, 19, 32, 04, 25, 29, 37, 39, 31, 18, 07, 13, 33, 02, 36, 23

You should check at least the first few of these. These students will receive the first phone. The remaining 20 will evaluate the second phone.

As Example 3.12 illustrates, randomization requires two steps: assign labels to the experimental units and then use Table B to select labels at random. Be sure that all labels are the same length so that all have the same chance to be chosen. Use the shortest possible labels—one digit for 10 or fewer individuals, two digits for 11 to 100 individuals, and so on. Don't try to scramble the labels as you assign them. Table B will do the required randomizing, so assign labels in any convenient manner, such as in alphabetical order for human subjects. You can read digits from Table B in any order—along a row, down a column, and so on—because the table has no order. As an easy standard practice, we recommend reading along rows.

It is easy to use statistical software or Excel to randomize. Here are the steps:

*Step 1: Label.* The first step in assigning labels to the experimental units is similar to the procedure we described previously. One difference, however, is that we

are not restricted to using numerical labels. Any system where each experimental unit has a unique label identifier will work.

*Step 2: Use the computer.* Once we have the labels, we then create a data set with the labels and generate a random number for each label. In Excel, this can be done with the RAND() function. Finally, we sort the entire data set based on the random numbers. Groups are formed by selecting units in order from the sorted list.

This process is essentially the same as writing the labels on a deck of cards, shuffling the cards, and dealing them out one at a time.

## **Example**

### **3.13 Using software for the randomization.**

Let's do a randomization similar to the one we did in Example 3.12, but this time using Excel. Here we will use 10 experimental units. We will assign 5 to the treatment group and 5 to the control group. We first create a data set with the numbers 1 to 10 in the first column. See Figure 3.4(a). Then we use RAND() to generate 10 random numbers in the second column. See Figure 3.4(b). Finally, we sort the data set based on the numbers in the second column. See Figure 3.4(c). The first 5 labels (7, 1, 10, 4, and 6) are assigned to the experimental group. The remaining 5 labels (8, 2, 3, 9, and 5) correspond to the control group.

	A	B
1	1	
2	2	
3	3	
4	4	
5	5	
6	6	
7	7	
8	8	
9	9	
10	10	

(a)

	A	B
1	1	0.145489
2	2	0.818926
3	3	0.906712
4	4	0.536023
5	5	0.967995
6	6	0.635631
7	7	0.060332
8	8	0.725457
9	9	0.922880
10	10	0.494231

(b)

	A	B
1	7	0.060332
2	1	0.145489
3	10	0.494231
4	4	0.536023
5	6	0.635631
6	8	0.725457
7	2	0.818926
8	3	0.906712
9	9	0.922880
10	5	0.967995

(c)

### FIGURE 3.4

Randomization of 10 experimental units using an Excel spreadsheet, for Example 3.13.  
 (a) Labels. (b) Random numbers. (c) Sorted list of labels.

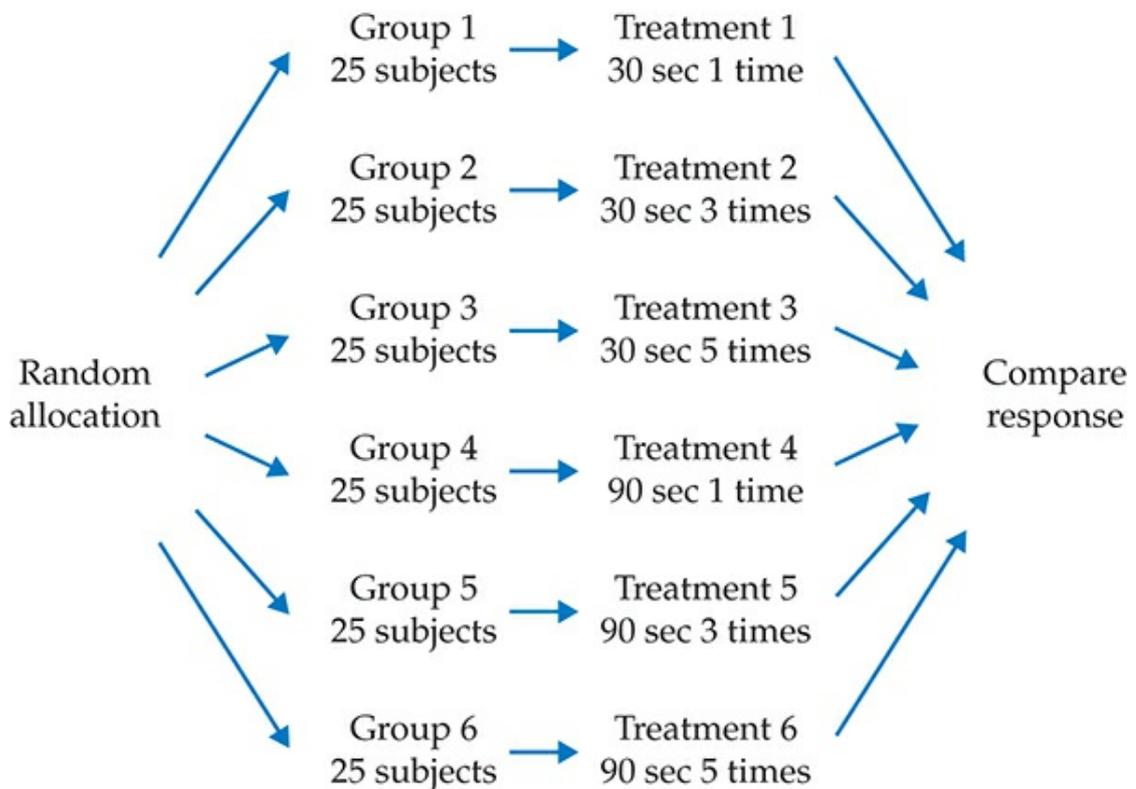
When all experimental units are allocated at random among all treatments, as in Example 3.13, the experimental design is **completely randomized**. Completely randomized designs can compare any number of treatments. The treatments can be formed by levels of a single factor or by more than one factor.

## Example

### 3.14 Randomization for the TV commercial experiment.

Figure 3.2 (page 177) displays six treatments formed by the two factors in an experiment on response to a TV commercial. Suppose that we have 150 students who are willing to serve as subjects. We must assign 25 students at random to each group. Figure 3.5 outlines the completely randomized design.

To carry out the random assignment, label the 150 students 001 to 150. (Three digits are needed to label 150 subjects.) Enter Table B and read three-digit groups until you have selected 25 students to receive Treatment 1 (a 30-second ad shown once). If you start at line 140, the first few labels for Treatment 1 subjects are 129, 048, and 003.



**FIGURE 3.5**

Outline of a completely randomized design comparing six treatments, for Example 3.14.

Continue in Table B to select 25 more students to receive Treatment 2 (a 30-second ad shown 3 times). Then select another 25 for Treatment 3 and so

on until you have assigned 125 of the 150 students to Treatments 1 through 5. The 25 students who remain get Treatment 6.



The randomization is straightforward but very tedious to do by hand. We recommend the *Simple Random Sample* applet. Exercise 3.41 (page 191) shows how to use the applet to do the randomization for this example.

## USE YOUR KNOWLEDGE

### 3.23 Do the randomization.

Use computer software to carry out the randomization in Example 3.14.

## Cautions about experimentation

The logic of a randomized comparative experiment depends on our ability to treat all the experimental units identically in every way except for the actual treatments being compared. Good experiments therefore require careful attention to details. The ideal situation is where a study is **double-blind**—neither the subjects themselves nor the experimenters know which treatment any subject has received. The double-blind method avoids unconscious bias by, for example, a doctor who doesn't think that “just a placebo” can benefit a patient.

### double-blind



*Many—perhaps most—experiments have some weaknesses in detail. The environment of an experiment can influence the outcomes in unexpected ways. Although experiments are the gold standard for evidence of cause and effect, really convincing evidence usually requires that a number of studies in different places with different details produce similar results. Here are some brief examples of what can go wrong.*

## Example

### 3.15 Placebo for a marijuana experiment.

A study of the effects of marijuana recruited young men who used marijuana. Some were randomly assigned to smoke marijuana cigarettes, while others were given placebo cigarettes. This failed: the control group recognized that their cigarettes were phony and complained loudly. It may be quite common for blindness to fail because the subjects can tell which treatment they are receiving.<sup>11</sup>

The most serious potential weakness of experiments is **lack of realism**. The subjects or treatments or setting of an experiment may not realistically duplicate the conditions we really want to study. Here is an example.

lack of realism

## Example

### 3.16 Layoffs and feeling bad.

How do layoffs at a workplace affect the workers who remain on the job? To try to answer this question, psychologists asked student subjects to proofread text for extra course credit, then “let go” some of the workers (who were actually accomplices of the experimenters). Some subjects were told that those let go had performed poorly (Treatment 1). Others were told that not all could be kept and that it was just luck that they were kept and others let go (Treatment 2). We can’t be sure that the reactions of the students are the same as those of workers who survive a layoff in which other workers lose their jobs. Many behavioral science experiments use student subjects in a campus setting. Do the conclusions apply to the real world?



Lack of realism can limit our ability to apply the conclusions of an experiment to the settings of greatest interest. Most experimenters want to generalize their

conclusions to some setting wider than that of the actual experiment. *Statistical analysis of an experiment cannot tell us how far the results will generalize to other settings.* Nonetheless, the randomized comparative experiment, because of its ability to give convincing evidence for causation, is one of the most important ideas in statistics.

## Matched pairs designs

Completely randomized designs are the simplest statistical designs for experiments. They illustrate clearly the principles of control, randomization, and repetition. However, completely randomized designs are often inferior to more elaborate statistical designs. In particular, matching the subjects in various ways can produce more precise results than simple randomization.

The simplest use of matching is a **matched pairs design**, which compares just two treatments. The subjects are matched in pairs. For example, an experiment to compare two advertisements for the same product might use pairs of subjects with the same age, sex, and income. The idea is that matched subjects are more similar than unmatched subjects, so that comparing responses within a number of pairs is more efficient than comparing the responses of groups of randomly assigned subjects. Randomization remains important: which one of a matched pair sees the first ad is decided at random. One common variation of the matched pairs design imposes both treatments on the same subjects, so that each subject serves as his or her own control. Here is an example.

### matched pairs design

## Example

### 3.17 Matched pairs for the smartphone prototype experiment.

Example 3.11 describes an experiment to compare two prototypes of a new smartphone. The experiment compared two treatments: Phone 1 and Phone 2. The response variable is the satisfaction of the college student participant with the new smartphone. In Example 3.11, 40 student subjects were assigned at random, 20 students to each phone. This is a completely randomized design, outlined in Figure 3.3. Subjects differ in how satisfied they are with smartphones in general. The completely randomized design relies on chance to create two similar groups of subjects.

If we wanted to do a matched pairs version of this experiment, we would

have each college student use each phone for two weeks. An effective design would randomize the *order* in which the phones are evaluated by each student. This will eliminate bias due to the possibility that the first phone evaluated will be systematically evaluated higher or lower than the second phone evaluated.

The completely randomized design uses chance to decide which subjects will evaluate each smartphone prototype. The matched pairs design uses chance to decide which 20 subjects will evaluate Phone 1 first. The other 20 will evaluate Phone 2 first.

## Block designs

The matched pairs design of Example 3.17 uses the principles of comparison of treatments, randomization, and repetition on several experimental units. However, the randomization is not complete (all subjects randomly assigned to treatment groups) but is restricted to assigning the order of the treatments for each subject. *Block designs* extend the use of “similar subjects” from pairs to larger groups.

### BLOCK DESIGN

A **block** is a group of experimental units or subjects that are known before the experiment to be similar in some way that is expected to affect the response to the treatments. In a **block design**, the random assignment of units to treatments is carried out separately within each block.

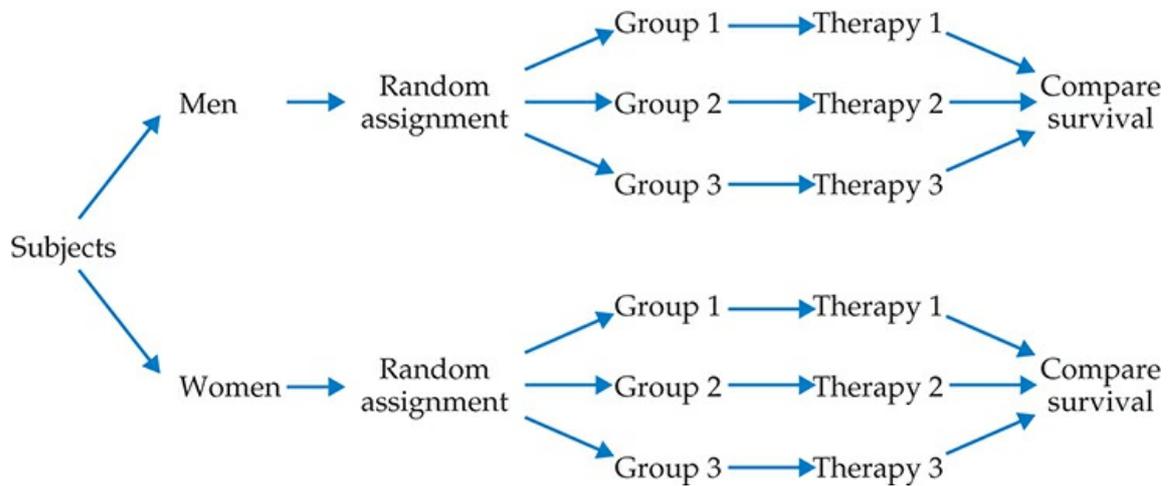
Block designs can have blocks of any size. A block design combines the idea of creating equivalent treatment groups by matching with the principle of forming treatment groups at random. Blocks are another form of *control*. They control the effects of some outside variables by bringing those variables into the experiment to form the blocks. Here are some typical examples of block designs.

### Example

#### 3.18 Blocking in a cancer experiment.

The progress of a type of cancer differs in women and men. A clinical experiment to compare three therapies for this cancer therefore treats sex as a

blocking variable. Two separate randomizations are done, one assigning the female subjects to the treatments and the other assigning the male subjects. Figure 3.6 outlines the design of this experiment. Note that there is no randomization involved in making up the blocks. They are groups of subjects who differ in some way (sex in this case) that is apparent before the experiment begins.



**FIGURE 3.6**

Outline of a block design, for Example 3.18. The blocks consist of male and female subjects. The treatments are the three therapies for cancer.

## Example

### 3.19 Blocking in an agriculture experiment.

The soil type and fertility of farmland differ by location. Because of this, a test of the effect of tillage type (two types) and pesticide application (three application schedules) on soybean yields uses small fields as blocks. Each block is divided into six plots, and the six treatments are randomly assigned to plots separately within each block.

## Example

### 3.20 Blocking in an education experiment.

The Tennessee STAR class size experiment (Example 3.8, page 176) used a block design. It was important to compare different class types in the same school because the children in a school come from the same neighborhood, follow the same curriculum, and have the same school environment outside class. In all, 79 schools across Tennessee participated in the program. That is, there were 79 blocks. New kindergarten students were randomly placed in the three types of class separately within each school.

Blocks allow us to draw separate conclusions about each block, for example, about men and women in the cancer study in Example 3.18. Blocking also allows more precise overall conclusions because the systematic differences between men and women can be removed when we study the overall effects of the three therapies. The idea of blocking is an important additional principle of statistical design of experiments. A wise experimenter will form blocks based on the most important unavoidable sources of variability among the experimental units. Randomization will then average out the effects of the remaining variation and allow an unbiased comparison of the treatments.

### SECTION 3.2 Summary

In an experiment, one or more **treatments** are imposed on the **experimental units** or **subjects**. Each treatment is a combination of **levels** of the explanatory variables, which we call **factors**. **Outcomes** are the measured variables that are used to compare the treatments.

The **design** of an experiment refers to the choice of treatments and the manner in which the experimental units or subjects are assigned to the treatments.

The basic principles of statistical design of experiments are **compare, randomization, and repetition**.

The simplest form of control is **comparison**. Experiments should compare two or more treatments in order to prevent **confounding** the effect of a treatment with other influences, such as lurking variables.

**Randomization** uses chance to assign subjects to the treatments. Randomization creates treatment groups that are similar (except for chance variation) before the treatments are applied. Randomization and comparison together prevent **bias**, or systematic favoritism, in experiments.

You can carry out randomization by giving numerical labels to the experimental units and using a **table of random digits** to choose treatment groups.

**Repetition** of the treatments on many units reduces the role of chance variation and makes the experiment more sensitive to differences among the treatments.

Good experiments require attention to detail as well as good statistical design. Many behavioral and medical experiments are **double-blind**. **Lack of realism** in an experiment can prevent us from generalizing its results.

In addition to comparison, a second form of control is to restrict randomization by forming **blocks** of experimental units that are similar in some way that is important to the response. Randomization is then carried out separately within each block.

**Matched pairs** are a common form of blocking for comparing just two treatments. In some matched pairs designs, each subject receives both treatments in a random order. In others, the subjects are matched in pairs as closely as possible, and one subject in each pair receives each treatment.

## SECTION 3.2 Exercises

*For Exercises 3.18 and 3.19, see page 177; for Exercise 3.20, see page 179; for Exercises 3.21 and 3.22, see page 180; and for Exercise 3.23, see page 185.*

### 3.24 How to relax.

An experiment compared three ways to relax before an exam. Sixty college students were randomly assigned to use one of three methods to relax before an exam. The methods were (1) take a slow walk for 15 minutes, (2) do a yoga exercise for 15 minutes, and (3) lie down and listen to soothing music. What are the experimental units, the treatments, and the outcomes for this experiment? Can we use the term “subjects” for the experimental units? Explain your answers.

### 3.25 Online homework.

Thirty students participated in a study designed to evaluate a new online homework system. None of the students had used an online homework system in the past. After using the system for a month, they were asked to rate their satisfaction with the system using a 5-point scale.

- What are the experimental units, the treatment, and the outcome for this experiment? Can we use the term “subjects” for the experimental units? Explain your answers.
- Is this a comparative experiment? If your answer is Yes, explain why. If your answer is No, describe how you would change the design so that it would be a comparative experiment.
- Suggest some different outcomes that you think would be appropriate for this experiment.

### 3.26 Coaching using mobile technology.

Refer to Exercise 3.19 (page 177), where an experiment using mobile technology to improve diet and exercise behavior is described.

- Why would a control group with a placebo treatment be useful in this experiment?
- Explain what the placebo effect is in this setting.
- Describe a treatment that would serve as a placebo for this experiment.

### 3.27 Online sales of running shoes.

A company that sells running shoes online wants to compare two new marketing strategies. They will test

the strategies on 10 weekdays. In the morning of each day, a web page describing the comfort of the running shoes will be displayed. In the afternoon of each day, a web page describing the discounted price for the shoes will be displayed. Sales of the featured running shoes will be compared at the end of the experiment.

- (a) What are the experimental units, the treatments, and the outcomes for this experiment? Explain your answers.
- (b) Is this a comparative experiment? Why or why not?
- (c) Could the experiment be improved by using randomization? Explain your answer.
- (d) Could the experiment be improved by using a placebo treatment? Explain your answer.

### **3.28 Online sales of running shoes.**

Refer to the previous exercise. Suppose that for each day, you randomized the web pages, showing one in the morning and the other in the afternoon. Can you view this experiment as a block design? Explain your answer.

### **3.29 Randomize the web pages for the running shoes.**

Refer to the previous exercise. Use Table B to randomize the treatments. Report the place in the table where you started, and list the random numbers that you used to determine when to display the web pages.

### **3.30 Randomize the web pages for the running shoes.**

Refer to the previous exercise. Use software to carry out the randomization.

### **3.31 Online sales of running shoes.**

Refer to Exercise 3.27. Here is another way in which the experiment could be designed. Suppose that you alternate the display each time a customer visits the website. Can you view this experiment as a matched pairs design? Explain your answer.

### **3.32 The *Sports Illustrated* jinx.**

Some people believe that teams or individual athletes who appear on the cover of *Sports Illustrated* magazine will experience bad luck soon after they appear. Can you evaluate this belief with an experiment? Explain your answer.

### **3.33 What is needed?**

Explain what is deficient in each of the following proposed experiments and explain how you would improve the experiment.

(a) Two product promotion offers are to be compared. The first, which offers two items for \$2 will be used in a store on Friday. The second, which offers three items for \$3, will be used in the same store on Saturday.

(b) A study compares two marketing campaigns to encourage individuals to eat more fruits and vegetables. The first campaign is launched in Florida at the same time that the second campaign is launched in Minnesota.

(c) You want to evaluate the effectiveness of a new investment strategy. You try the strategy for one year and evaluate the performance of the strategy.

### **3.34 What is wrong?**

Explain what is wrong with each of the following randomization procedures and describe how you would do the randomization correctly.

(a) Twenty students are to be used to evaluate a new treatment. Ten men are assigned to receive the treatment, and 10 women are assigned to be the controls.

(b) Ten subjects are to be assigned to two treatments, 5 to each. For each subject, a coin is tossed. If the coin comes up heads, the subject is assigned to the first treatment; if the coin comes up tails, the subject is assigned to the second treatment.

(c) An experiment will assign 40 rats to four different treatment conditions. The rats arrive from the supplier in batches of 10 and the treatment lasts two weeks. The first batch of 10 rats is randomly assigned to one of the four treatments, and data for these rats are collected. After a one-week break, another batch of 10 rats arrives and is assigned to one of the three remaining treatments. The process continues until the last batch of rats is given the treatment that has not been assigned to the three previous batches.

### **3.35 Evaluate a new orientation program.**

Your company runs a two-day orientation program Monday and Tuesday each week for new employees. A new program is to be compared with the current one. Set up an experiment to compare the new program with the old. Be sure to provide details regarding randomization and what outcome variables you will measure.

### **3.36 Do magnets reduce pain?**

Some claim that magnets can be used to reduce pain. Design a double-blind experiment to test this claim. Write a proposal requesting funding for your study giving all the important details, including the number of subjects, issues concerning randomization, and how you will make the study double-blind.

### **3.37 Calcium and vitamin D.**

Vitamin D is needed for the body to use calcium. An experiment is designed to study the effects of calcium and vitamin D supplements on the bones of first-year college students. The outcome measure is the total body bone mineral content (TBBMC), a measure of bone health. Three doses of calcium will be used: 0, 200, and 400 milligrams per day (mg/day). The doses of vitamin D will be 0, 50, and 100 international units (IU) per day. The calcium and vitamin D will be given in a single tablet. All tablets, including those with no calcium and no vitamin D, will look identical. Subjects for the study will be 90 men and 90 women.

(a) What are the factors and the treatments for this experiment?

- (b) Draw a picture explaining how you would randomize the 180 college students to the treatments.
- (c) Use a spreadsheet to carry out the randomization.
- (d) Is there a placebo in this experiment? Explain your answer.

### 3.38 Does oxygen help football players?

We often see players on the sidelines of a football game inhaling oxygen. Their coaches think this will speed their recovery. We might measure recovery from intense exercise as follows: Have a football player run 100 yards three times in quick succession. Then allow three minutes to rest before running 100 yards again. Time the final run. Because players vary greatly in speed, you plan a matched pairs experiment using 30 football players as subjects. Describe the design of such an experiment to investigate the effect of inhaling oxygen during the rest period. Why should each player's two trials be on different days? Use Table B at line 135 to decide which players will get oxygen on their first trial.

### 3.39 Five-digit zip codes and delivery time of mail.

Does adding the five-digit postal zip code to an address really speed up delivery of letters? Does adding the four more digits that make up “zip + 4” speed delivery yet more? What about mailing a letter on Monday, Thursday, or Saturday? Describe the design of an experiment on the speed of first-class mail delivery. For simplicity, suppose that all letters go from you to a friend, so that the sending and receiving locations are fixed.



### 3.40 Which coffee is preferred?

A coffeehouse wants to compare two new varieties of coffee.

- (a) Describe an experiment in which different customers evaluate each variety. Be sure to provide details, including how many customers you will use, issues related to randomization, and what evaluation data you will collect.
- (b) Do the same for an experiment in which each customer evaluates both varieties of coffee.
- (c) Which experiment do you prefer? Give reasons for your answer.



### 3.41 Use the *Simple Random Sample* applet.

The *Simple Random Sample* applet allows you to randomly assign experimental units to more than two groups without difficulty. Example 3.14 (page 184) describes a randomized comparative experiment in which 150 students are randomly assigned to six groups of 25.

- (a) Use the applet to randomly choose 25 out of 150 students to form the first group. Which students are in this group?
- (b) The “Population hopper” now contains the 125 students who were not chosen, in scrambled order. Click “Sample” again to choose 25 of these remaining students to make up the second group. Which students were chosen?
- (c) Click “Sample” three more times to choose the third, fourth, and fifth groups. Don't take the time to write down these groups. Check that there are only 25 students remaining in the “Population hopper.” These subjects get Treatment 6. Which students are they?



### 3.42 Use the *Simple Random Sample* applet.

You can use the *Simple Random Sample* applet to choose a group at random once you have labeled the subjects. Example 3.12 (page 182) uses Table B to choose 20 students from a group of 40 for a study of smartphone preferences. Use the applet to choose 20 students. Which students did you choose?



### 3.43 Health benefits of bee pollen.

“Bee pollen is effective for combating fatigue, depression, cancer, and colon disorders.” So says a website that offers the pollen for sale. We wonder if bee pollen really does prevent colon disorders. Here are two ways to study this question. Explain why the first design will produce more trustworthy data.

- (a) Find 400 women who do not have colon disorders. Randomly assign 200 to take bee pollen capsules and the other 200 to take placebo capsules that are identical in appearance. Follow both groups for 5 years.
- (b) Find 200 women who take bee pollen regularly. Match each with a woman of the same age, race, and occupation who does not take bee pollen. Follow both groups for 5 years.



### 3.44 Use the *Simple Random Sample* applet.

The *Simple Random Sample* applet can demonstrate how randomization works to create similar groups for comparative experiments. Suppose that (unknown to the experimenters) the 20 even-numbered students among the 40 subjects for the smartphone study in Example 3.12 (page 182) tend to send more text messages than the odd-numbered students. We would like the two groups to be similar with respect to text messaging. Use the applet to choose 10 samples of size 20 from the 40 students. (Be sure to click “Reset” after each sample.) Record the counts of even-numbered students in each of your 10 samples. You see that there is considerable chance variation but no systematic bias in favor of one or the other group in assigning the fast-reacting students. Larger samples from larger populations will on the average do a better job of making the two groups equivalent.



### 3.45 Calcium and the bones of young girls.

Calcium is important to the bone development of young girls. To study how the bodies of young girls process calcium, investigators used the setting of a summer camp. Calcium was given in punch at either a high or a low level. The camp diet was otherwise the same for all girls. Suppose that there are 40 campers.

- (a) Outline a completely randomized design for this experiment.
- (b) Describe a matched pairs design in which each girl receives both levels of calcium (with a “washout period” in which no calcium supplementation was given between the two treatment periods). What is the advantage of the matched pairs design over the completely randomized design?
- (c) The same randomization can be used in different ways for both designs. Label the subjects 01 to 40. You must choose 20 of the 40. Use Table B at line 120 to choose just the first 5 of the 20. How are the 20 subjects chosen treated in the completely randomized design? How are they treated in the matched pairs design?



### 3.46 Random digits.

Table B is a table of random digits. Which of the following statements are true of a table of random digits, and which are false? Explain your answers.

- (a) There are exactly four 0s in each row of 40 digits.
- (b) Each pair of digits has chance  $1/100$  of being 00.
- (c) The digits 0000 can never appear as a group, because this pattern is not random.

### 3.47 Measuring water quality in streams and lakes.

Water quality of streams and lakes is an issue of concern to the public. Although trained professionals typically are used to take reliable measurements, many volunteer groups are gathering and distributing information based on data that they collect.<sup>12</sup> You are part of a team to train volunteers to collect accurate water quality data. Design an experiment to evaluate the effectiveness of the training. Write a summary of your proposed design to present to your team. Be sure to include all the details that they will need to evaluate your proposal.

## 3.3 Sampling Design

When you complete this section, you will be able to

- Distinguish between a population and a sample.
- Use the response rate to evaluate a survey.
- Use Table B to generate a simple random sample (SRS).
- Use software to generate a simple random sample.
- Construct a stratified random sample, using Table B or software to select the samples from the strata.
- Identify sample designs as voluntary response samples, simple random samples, stratified random samples, or multistage random samples.
- Identify characteristics of samples that limit their usefulness, including undercoverage, nonresponse, response bias, and the wording of questions.

A political scientist wants to know what percent of college-age adults consider themselves conservatives. An automaker hires a market research firm to learn what percent of adults aged 18 to 35 recall seeing television advertisements for a new sport utility vehicle. Government economists inquire about average household income.

In all these cases, we want to gather information about a large group of individuals. We will not, as in an experiment, impose a treatment in order to observe the response. Also, time, cost, and inconvenience forbid contacting every individual. In such cases, we gather information about only part of the group—a *sample*—in order to draw conclusions about the whole. **Sample surveys** are an important kind of observational study.

sample survey

### POPULATION AND SAMPLE

The entire group of individuals that we want information about is called the **population**.

A **sample** is a part of the population that we actually examine in order to gather information.

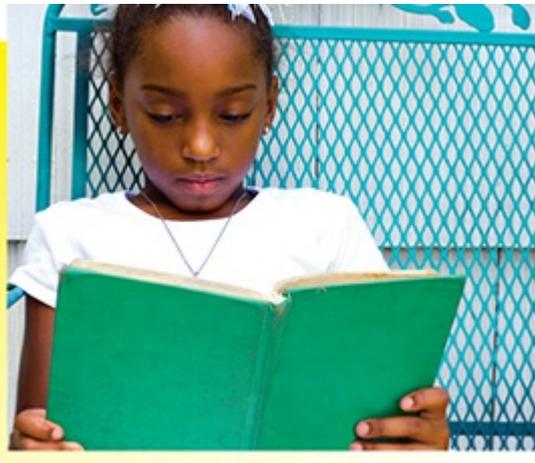
Notice that “population” is defined in terms of our desire for knowledge. If we wish to draw conclusions about all U.S. college students, that group is our

population even if only local students are available for questioning. The sample is the part from which we draw conclusions about the whole. The **design** of a sample survey refers to the method used to choose the sample from the population.

### sample design

## Example

### 3.21 The Reading Recovery program.



The Reading Recovery (RR) program has specially trained teachers work one-on-one with at-risk first-grade students to help them learn to read. A study was designed to examine the relationship between the RR teachers' beliefs about their ability to motivate students and the progress of the students whom they teach.<sup>13</sup> The Reading Recovery International Data Evaluation Center website ([www.idecweb.us](http://www.idecweb.us)) says that there are 13,823 RR teachers. The researchers send a questionnaire to a random sample of 200 of these. The population consists of all 13,823 RR teachers, and the sample is the 200 that were randomly selected.

Unfortunately, our idealized framework of population and sample does not exactly correspond to the situations that we face in many cases. In Example 3.21, the list of teachers was prepared at a particular time in the past. It is very likely that some of the teachers on the list are no longer working as RR teachers today. New teachers have been trained in RR methods and are not on the list. Despite these difficulties, we still view the list as the population. Also, we may have out-of-date addresses for some who are still working as RR teachers, and some teachers may

choose not to respond to our survey questions.

In reporting the results of a sample survey, it is important to include all details regarding the procedures used. Follow-up mailings or phone calls to those who do not initially respond can help increase the response rate. The proportion of the original sample who actually provide usable data is called the **response rate** and should be reported for all surveys. If only 150 of the teachers who were sent questionnaires provided usable data, the response rate would be 150/200, or 75%.

response rate

## USE YOUR KNOWLEDGE

### 3.48 Are they satisfied?

An educational research team wanted to examine the relationship between faculty participation in decision making and job satisfaction in Mongolian public universities. They are planning to randomly select 300 faculty members from a list of 2500 faculty members in these universities. The Job Descriptive Index will be used to measure job satisfaction, and the Conway Adaptation of the Alutto-Belasco Decisional Participation Scale will be used to measure decision participation. Describe the population and the sample for this study. Can you determine the response rate?

### 3.49 What is the impact of the taxes?

A study was designed to assess the impact of taxes on forestland usage in part of the Upper Wabash River Watershed in Indiana.<sup>14</sup> A survey was sent to 772 forest owners from this region and 348 were returned. Consider the population, the sample, and the response rate for this study. Describe these on the basis of the information given and indicate any additional information that you would need to assess the impact of taxes.

Poor sample designs can produce misleading conclusions. Here is an example.

## Example

### 3.22 Sampling pieces of steel.

A mill produces large coils of thin steel for use in manufacturing home appliances. The quality engineer wants to submit a sample of 5-centimeter squares to detailed laboratory examination. She asks a technician to cut a sample of 10 such squares. Wanting to provide “good” pieces of steel, the technician carefully avoids the visible defects in the coil material when cutting the sample. The laboratory results are wonderful, but the customers complain about the material they are receiving.

In Example 3.22, the sample was selected in a manner that guaranteed that it would not be representative of the entire population. This sampling scheme displays *bias*, or systematic error, in favoring some parts of the population over others.

Online opinion polls use *voluntary response samples*, a particularly common form of biased sample. The sample who respond are not representative of the population at large. People who take the trouble to respond to an open invitation are not representative of the entire population.

## VOLUNTARY RESPONSE SAMPLE

A **voluntary response sample** consists of people who choose themselves by responding to a general appeal. Voluntary response samples are biased because people with strong opinions, especially negative opinions, are most likely to respond.

The remedy for bias in choosing a sample is to allow impersonal chance to do the choosing, so that there is neither favoritism by the sampler (Example 3.22) nor voluntary response (online opinion polls). Random selection of a sample eliminates bias by giving all individuals an equal chance to be chosen, just as randomization eliminates bias in assigning experimental units.

## Simple random samples

The simplest sampling design amounts to placing names in a hat (the population) and drawing out a handful (the sample). This is *simple random sampling*.

## SIMPLE RANDOM SAMPLE

A **simple random sample (SRS)** of size  $n$  consists of  $n$  individuals from the population chosen in such a way that every set of  $n$  individuals has an equal chance to be the sample actually selected.

Each treatment group in a completely randomized experimental design is an SRS drawn from the available experimental units. We select an SRS by labeling all the individuals in the population and using software or a table of random digits to select a sample of the desired size, just as in experimental randomization. Notice that an SRS not only gives every possible sample an equal chance to be chosen but also gives each individual an equal chance to be chosen. There are other random sampling designs that give each individual, but not each sample, an equal chance. One such design, systematic random sampling, is described in Exercise 3.70 (page 204).

### Example

#### 3.23 Brands.



A brand is a symbol or images that are associated with a company. An

effective brand identifies the company and its products. Using a variety of measures, dollar values for brands can be calculated. In Exercise 1.67 (page 78), you examined the distribution of the values of the top 100 brands.

Suppose that you want to write a research report on some of the characteristics of the companies in this elite group. You decide to look carefully at the websites of 10 companies from the list. One way to select the companies is to use a simple random sample. Here are some details about how to do this using Table B. We start with a list of the companies with the top 100 brands. This is given in the data file BRANDS. Next we need to label the companies. In the data file they are listed with their ranks, 1 to 100. Let's assign the labels 01 to 09 to the first nine companies, and 00 to the company with rank 100. With these labels, we can use Table B to select the SRS.

Let's start with line 156 of Table B. This line has the entries 55494 67690 88131 81800 11188 28552 25752 21953. These are grouped in sets of five digits, but we need to use sets of two digits for our randomization. Here is line 156 of Table B in sets of two digits: 55 49 46 76 90 88 13 18 18 00 11 18 82 85 52 25 75 22 19 53.

Using these random digits, we select Audi (55), Dell (49), Heinz (46), Santander (76), Smirnoff (90), Starbucks (88), Disney (13), Oracle (18; we skip the second 18 because we have already selected Oracle to be in our SRS), Gap (00; recoded from rank 100), and Mercedes-Benz (11).



Most statistical software will select an SRS for you, eliminating the need for Table B. The *Simple Random Sample* applet on the text website is another convenient way to automate this task.

Excel can do the job in a way similar to how we randomized experimental units to treatments in designed experiments. There are four steps:

1. Create a data set with all the elements of the population in one column.
2. Generate a random number for each element of the population; put these in another column.
3. Sort the data set by the random number column.
4. The simple random sample is obtained by taking elements in the sorted list in order until the desired sample size is reached.

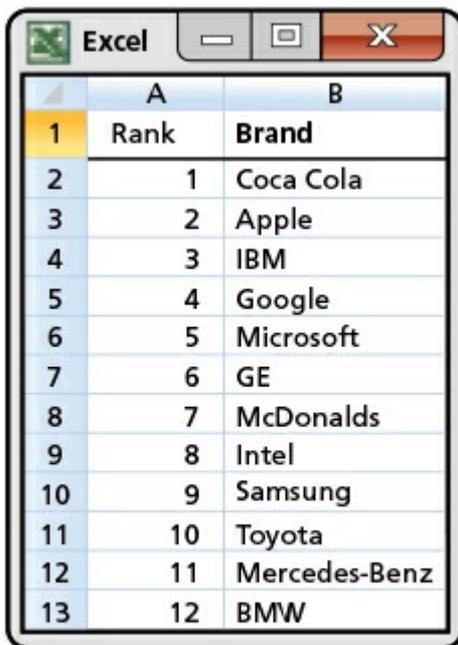
We illustrate the procedure with the brands in Example 3.24.

## Example

### 3.24 Select a random sample.

Figure 3.7(a) gives the spreadsheet with the company names in column B. Only the first 11 of the 100 companies in the top 100 brands list are shown.

The random numbers generated by the RAND() function are given in the next column in Figure 3.7(b). The sorted (smallest to largest) data set is given in Figure 3.7(c). The 10 resorts that were selected for our random sample are HP, Dell, Philips, Nestle, Hermes, AXA, Gucci, SAP, Nescafe, and Pizza Hut.



	A	B
1	Rank	Brand
2	1	Coca Cola
3	2	Apple
4	3	IBM
5	4	Google
6	5	Microsoft
7	6	GE
8	7	McDonalds
9	8	Intel
10	9	Samsung
11	10	Toyota
12	11	Mercedes-Benz
13	12	BMW

(a)

	A	B	C
1	Rank	Brand	Random
2	1	Coca Cola	0.811570218
3	2	Apple	0.264747625
4	3	IBM	0.962913044
5	4	Google	0.336853738
6	5	Microsoft	0.216197915
7	6	GE	0.645539433
8	7	McDonalds	0.388802686
9	8	Intel	0.654192165
10	9	Samsung	0.655038459
11	10	Toyota	0.815300556
12	11	Mercedes-Benz	0.950368088
13	12	BMW	0.615746964

(b)

	A	B	C
1	Rank	Brand	Random
2	15	HP	0.005952107
3	49	Dell	0.026236225
4	41	Philips	0.032582368
5	57	Nestle	0.059717930
6	63	Hermes	0.104594242
7	58	AXA	0.109335084
8	38	Gucci	0.150611277
9	25	SAP	0.161301562
10	35	Nescafe	0.187316329
11	86	Pizza Hut	0.210133012
12	5	Microsoft	0.216197915

(c)

**FIGURE 3.7**

Selection of a simple random sample of companies in the list of top 100 brands, for Example 3.24.

## USE YOUR KNOWLEDGE

### 3.50 Ringtones for cell phones.

You decide to change the ringtones for your cell phone by choosing 2 from a list of the 10 most popular ringtones.<sup>15</sup> Here is the list:

Gangnam Style	Cruise	Girl on Fire	I'm Different
I Knew You Were Trouble	Diamonds	Pontoon	Better Dig Two
Locked Out of Heaven	No Worries		

Select your two ringtones using a simple random sample. Show your work.

### 3.51 Listen to three songs.

The walk to your statistics class takes about 10 minutes, about the amount of time needed to listen to three songs on your iPod. You decide to take a simple random sample of songs from a Billboard List of Rock Songs.<sup>16</sup> Here is the list:

Ho Hey	Home	It's Time	Some Nights
The A Team	Little Talks	I Will Wait	Radioactive
Too Close	Madness		

Select the three songs for your iPod using a simple random sample. Show your work.

## Stratified random samples

The general framework for designs that use chance to choose a sample is a *probability sample*.

### PROBABILITY SAMPLE

A **probability sample** is a sample chosen by chance. We must know what samples are possible and what chance, or probability, each possible sample has.

Some probability sampling designs (such as an SRS) give each member of the population an *equal* chance to be selected. This may not be true in more elaborate sampling designs. In every case, however, the use of chance to select the sample is the essential principle of statistical sampling.

Designs for sampling from large populations spread out over a wide area are usually more complex than an SRS. For example, it is common to sample important groups within the population separately, then combine these samples.

This is the idea of a *stratified sample*.

## STRATIFIED RANDOM SAMPLE

To select a **stratified random sample**, first divide the population into groups of similar individuals, called **strata**. Then choose a separate SRS in each stratum and combine these SRSs to form the full sample.

Choose the strata based on facts known before the sample is taken. For example, a population of election districts might be divided into urban, suburban, and rural strata.

A stratified design can produce more exact information than an SRS of the same size by taking advantage of the fact that individuals in the same stratum are similar to one another. Think of the extreme case in which all individuals in each stratum are identical: just one individual from each stratum is then enough to completely describe the population.

Strata for sampling are similar to blocks in experiments. We have two names because the idea of grouping similar units before randomizing arose separately in sampling and in experiments.

### Example

#### 3.25 A stratified sample of companies from the top 100 brands list.

In Examples 3.23 and 3.24, you selected SRSs of size 10 from the companies in the list of the top 100 brands. Let's think about using a stratified sample. You still want to select 10 companies to examine for your report.

Let's classify the population of the 100 top brand companies into five strata based on the value of their brand ranks. The first stratum contains the companies with ranks 1 to 20, the second has ranks 21 to 40, the third has ranks 41 to 60, the fourth has ranks 61 to 80, and the fifth has ranks 81 to 100.

We have five strata and we want a total of 10 companies to study for a report. Therefore, we need to sample 2 companies from each stratum. We take an SRS of size 2 from each of these strata.

### Multistage random samples

Another common means of restricting random selection is to choose the sample in stages. These designs are called **multistage designs**. They are widely used in national samples of households or people. For example, data on employment and unemployment are gathered by the government's Current Population Survey, which conducts interviews in about 60,000 households each month. The cost of sending interviewers to the widely scattered households in an SRS would be too high. Moreover, the government wants data broken down by states and large cities.

### multistage random sample

The Current Population Survey therefore uses a multistage random sampling design. The final sample consists of groups of nearby households, called **clusters**, that an interviewer can easily visit. Most opinion polls and other national samples are also multistage, though interviewing in most national samples today is done by telephone rather than in person, eliminating the economic need for clustering. The Current Population Survey sampling design is roughly as follows:<sup>17</sup>

### clusters

**Stage 1.** Divide the United States into 2007 geographical areas called Primary Sampling Units, or PSUs. PSUs do not cross state lines. Select a sample of 754 PSUs. This sample includes the 428 PSUs with the largest population and a stratified sample of 326 of the others.

**Stage 2.** Divide each PSU selected into smaller areas called "blocks." Stratify the blocks using ethnic and other information and take a stratified sample of the blocks in each PSU.

**Stage 3.** Sort the housing units in each block into clusters of four nearby units. Interview the households in a probability sample of these clusters.

Analysis of data from sampling designs more complex than an SRS takes us beyond basic statistics. But the SRS is the building block of more elaborate designs, and analysis of other designs differs more in complexity of detail than in fundamental concepts.

## Cautions about sample surveys

Random selection eliminates bias in the choice of a sample from a list of the population. Sample surveys of large human populations, however, require much more than a good sampling design.<sup>18</sup> To begin, we need an accurate and complete list of the population. Because such a list is rarely available, most samples suffer from some degree of *undercoverage*. A sample survey of households, for example, will miss not only homeless people but also prison inmates and students in dormitories. An opinion poll conducted by telephone will miss the large number of American households without residential phones. The results of national sample surveys therefore have some bias if the people not covered—who most often are

poor people—differ from the rest of the population.

A more serious source of bias in most sample surveys is *nonresponse*, which occurs when a selected individual cannot be contacted or refuses to cooperate. Nonresponse to sample surveys often reaches 50% or more, even with careful planning and several callbacks. Because nonresponse is higher in urban areas, most sample surveys substitute other people in the same area to avoid favoring rural areas in the final sample. If the people contacted differ from those who are rarely at home or who refuse to answer questions, some bias remains.

## UNDERCOVERAGE AND NONRESPONSE

**Undercoverage** occurs when some groups in the population are left out of the process of choosing the sample.

**Nonresponse** occurs when an individual chosen for the sample can't be contacted or does not cooperate.

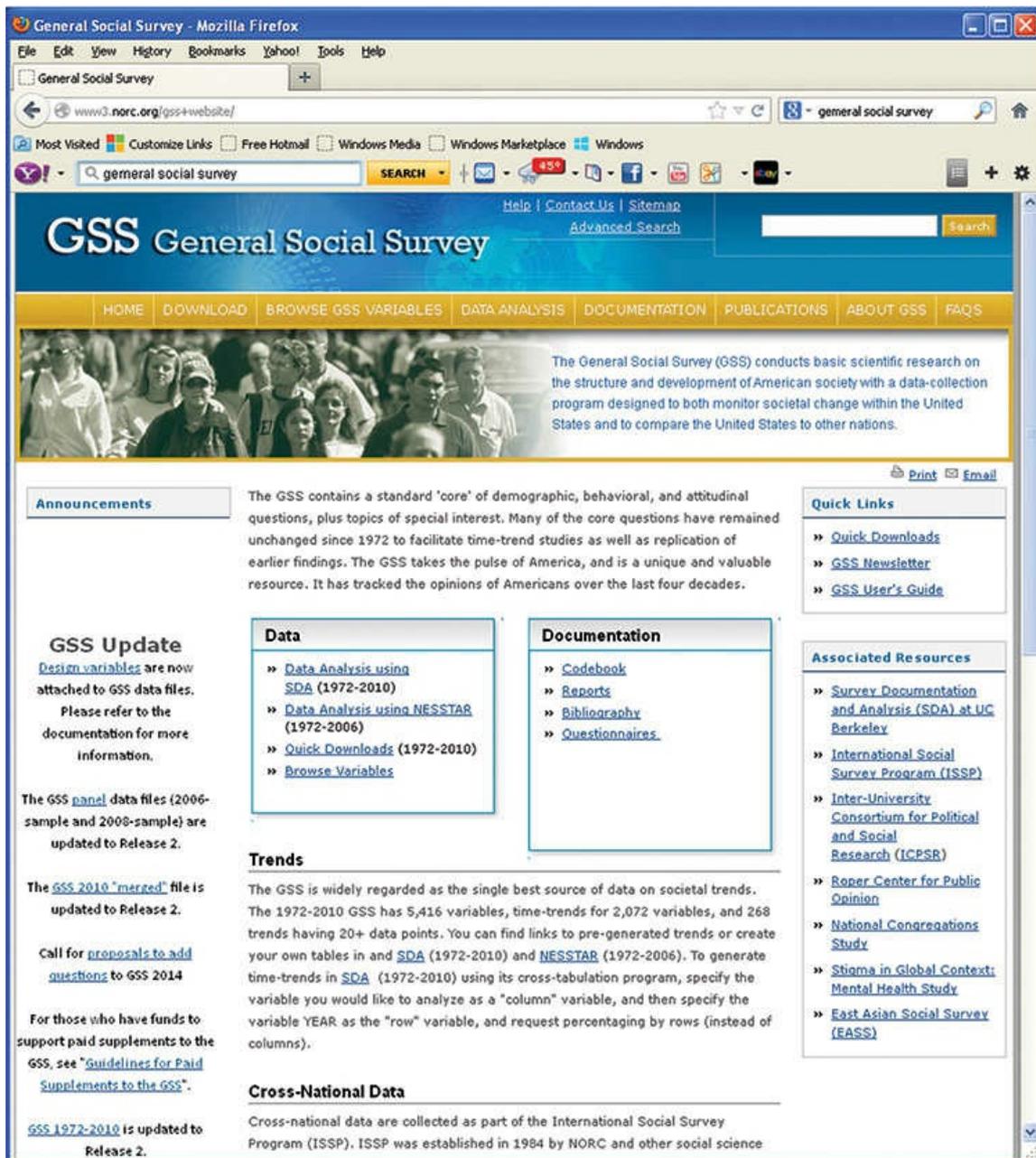
### Example

#### 3.26 Nonresponse in the Current Population Survey.

How bad is nonresponse? The Current Population Survey (CPS) has the lowest nonresponse rate of any poll we know: only about 5% of the households in the CPS sample refuse to take part, and another 2% or 3% can't be contacted. People are more likely to respond to a government survey such as the CPS, and the CPS contacts its sample in person before doing later interviews by phone.

The General Social Survey (Figure 3.8) is the nation's most important social science research survey. The GSS also contacts its sample in person, and it is run by a university. Despite these advantages, its most recent survey had a 30% rate of nonresponse.<sup>20</sup>

What about polls done by the media and by market research and opinion-polling firms? We don't know their rates of nonresponse, because they won't say. That itself is a bad sign.



**FIGURE 3.8**

Part of the home page for the General Social Survey (GSS). The GSS has assessed attitudes on a wide variety of topics since 1972. Its continuity over time makes the GSS a valuable source for studies of changing attitudes.

## Example

### 3.27 Change in nonresponse in Pew surveys.

The Pew Research Center conducts research using surveys on a variety of issues, attitudes, and trends.<sup>21</sup> A study by the center examined the decline in the response rates to their surveys over time. The changes are dramatic and there is a consistent pattern over time. Here are some data from the report:<sup>22</sup>

Year	1997	2000	2003	2006	2009	2012
Nonresponse rate	64%	72%	75%	79%	85%	91%

The center is devising alternative methods that show some promise of improving the response rates of their surveys.

Most sample surveys, and almost all opinion polls, are now carried out by telephone. This and other details of the interview method can affect the results. When presented with several options for a reply, such as “completely agree,” “mostly agree,” “mostly disagree,” and “completely disagree,” people tend to be a little more likely to respond to the first one or two options presented.

The behavior of the respondent or of the interviewer can cause **response bias** in sample results. Respondents may lie, especially if asked about illegal or unpopular behavior. The race or sex of the interviewer can influence responses to questions about race relations or attitudes toward feminism. Answers to questions that ask respondents to recall past events are often inaccurate because of faulty memory. For example, many people “telescope” events in the past, bringing them forward in memory to more recent time periods. “Have you visited a dentist in the last 6 months?” will often elicit a “Yes” from someone who last visited a dentist 8 months ago.

### response bias

The **wording of questions** is the most important influence on the answers given to a sample survey. Confusing or leading questions can introduce strong bias, and even minor changes in wording can change a survey’s outcome. Here are some examples.

### wording of questions

## Example

### 3.28 The form of the question is important.

In response to the question “Are you heterosexual, homosexual, or bisexual?”

in a social science research survey, one woman answered, “It’s just me and my husband, so bisexual.” The issue is serious, even if the example seems silly: reporting about sexual behavior is difficult because people understand and misunderstand sexual terms in many ways.

How do Americans feel about government help for the poor? Only 13% think we are spending too much on “assistance to the poor,” but 44% think we are spending too much on “welfare.” How do the Scots feel about the movement to become independent from England? Well, 51% would vote for “independence for Scotland,” but only 34% support “an independent Scotland separate from the United Kingdom.” It seems that “assistance to the poor” and “independence” are nice, hopeful words. “Welfare” and “separate” are negative words.<sup>23</sup>



The statistical design of sample surveys is a science, but this science is only part of the art of sampling. Because of nonresponse, response bias, and the difficulty of posing clear and neutral questions, you should hesitate to fully trust reports about complicated issues based on surveys of large human populations. *Insist on knowing the exact questions asked, the rate of nonresponse, and the date and method of the survey before you trust a poll result.*

### SECTION 3.3 Summary

A sample survey selects a **sample** from the **population** of all individuals about which we desire information. We base conclusions about the population on data about the sample.

The **design** of a sample refers to the method used to select the sample from the population. **Probability sampling designs** use impersonal chance to select a sample.

The basic probability sample is a **simple random sample (SRS)**. An SRS gives every possible sample of a given size the same chance to be chosen.

Choose an SRS by labeling the members of the population and using a **table of random digits** to select the sample. Software can automate this process.

To choose a **stratified random sample**, divide the population into **strata**, groups of individuals that are similar in some way that is important to the response. Then choose a separate SRS from each stratum and combine them to form the full sample.

**Multistage random samples** select successively smaller groups within the population in stages, resulting in a sample consisting of clusters of individuals. Each stage may employ an SRS, a stratified sample, or another type of sample.

Failure to use probability sampling often results in **bias**, or systematic errors in

the way the sample represents the population. **Voluntary response** samples, in which the respondents choose themselves, are particularly prone to large bias.

In human populations, even probability samples can suffer from bias due to **undercoverage** or **nonresponse**, from **response bias** due to the behavior of the interviewer or the respondent, or from misleading results due to **poorly worded questions**.

### SECTION 3.3 Exercises

*For Exercises 3.48 and 3.49, see page 193; and for Exercises 3.50 and 3.51, see pages 196.*

#### 3.52 What population and sample?

Twenty fourth-year students from your college who are majoring in English are randomly selected to be on a committee to evaluate changes in the statistics requirement for the major. There are 76 fourth-year English majors at your college.

- (a) Describe the population for this setting.
- (b) What is the sample?
- (c) Discuss the rationale for using fourth-year students for this study. Do you think that another group—for example, first-year students—would be better? Explain your answer.

#### 3.53 Response rate?

A survey designed to assess satisfaction with food items sold at a college's football games was sent to 150 fans who had season tickets. The total number of fans who have season tickets is 5674. Responses to the survey were received from 98 fans.

- (a) Describe the population for this survey.
- (b) What is the sample?
- (c) What is the response rate?
- (d) What is the nonresponse rate?
- (e) Suggest some ways that could be used in a future survey to increase the response rate.

#### 3.54 Who gets the dinner?

You are a member of a student organization that volunteers to work with third-grade students in your community who need help with their reading. Your organization will receive an award for its work. Three members of your organization will attend a dinner and ceremony where they will be given the award. There are 18 students in the organization.

- (a) What is the population for this setting?
- (b) What is the sample?
- (c) You have a spreadsheet with the names of the students. Explain how you would use the spreadsheet to select the students who will attend the dinner and ceremony. Give details.

(d) Use Table B to select the students. Give details.

### 3.55 Who gets the dinner using software?

Refer to the previous exercise.

(a) Use software to select the students. Explain the steps that you used in sufficient detail so that another person could repeat your work.

(b) Compare the use of Table B with software for selecting the students. Which do you prefer? Give reasons for your answer.

### 3.56 What kind of sample?

In each of the following situations, identify the sample as either an SRS, a stratified random sample, a multistage random sample, or a voluntary response sample. Explain your answers.

(a) There are seven sections of an introductory statistics course. A random sample of three sections is chosen, and then random samples of 8 students from each of these sections are chosen.

(b) A student organization has 55 members. A table of random numbers is used to select a sample of 5.

(c) An online poll asks people who visit this site to choose their favorite television show.

(d) Separate random samples of male and female first-year college students in an introductory psychology course are selected to receive a one-week alternative instructional method.

### 3.57 What's wrong?

Explain what is wrong in each of the following scenarios.

(a) The population consists of all individuals selected in a simple random sample.

(b) In a poll of an SRS of residents in a local community, respondents are asked to indicate the level of their concern about the dangers of dihydrogen monoxide, a substance that is a major component of acid rain and in its gaseous state can cause severe burns. (*Hint*: Ask a friend who is majoring in chemistry about this substance or search the Internet for information about it.)

(c) Students in a class are asked to raise their hands if they have cheated on an exam one or more times within the past year.

### 3.58 What's wrong?

Explain what is wrong with each of the following random selection procedures and explain how you would do the randomization correctly.

(a) To determine the reading level of an introductory statistics text, you evaluate all the written material in the third chapter.

(b) You want to sample student opinions about a proposed change in procedures for changing majors. You hand out questionnaires to 100 students as they arrive for class at 7:30A.M.

(c) A population of subjects is put in alphabetical order and a simple random sample of size 10 is taken by

selecting the first 10 subjects in the list.

### 3.59 Importance of students as customers.

A committee on community relations in a college town plans to survey local businesses about the importance of students as customers. From telephone book listings, the committee chooses 160 businesses at random. Of these, 72 return the questionnaire mailed by the committee. What is the population for this sample survey? What is the sample? What is the rate (percent) of nonresponse?

### 3.60 Consumer spending.

A Gallup Poll used telephone interviews to collect data on consumer spending on different days of the week.<sup>24</sup> Here are the averages (in dollars) for each day of the week:

Monday	59	Friday	63
Tuesday	56	Saturday	73
Wednesday	55	Sunday	76
Thursday	59		

- (a) Display the data graphically and write a short paragraph describing these averages.
- (b) The data were collected between January 2 and October 21, 2009. Discuss how this choice may have affected the results.

### 3.61 Which channel do you watch for news?

A Pew Research Center survey asked people what channel they regularly watch for news and their political party identification.<sup>25</sup> For one analysis they focused on those who regularly watch the Fox News Channel, CNN, MSNBC, and nightly network news. Here are the political profiles (in percents) for each of these news sources:

Party	Fox	CNN	MSNBC	Network
Republican	39	18	18	22
Democratic	33	51	45	45
Independent	22	23	27	26
Other/Don't know	6	8	10	7

Display the data graphically and write a report summarizing the results.

### 3.62 Identify the populations.

For each of the following sampling situations, identify the population as exactly as possible. That is, say what kind of individuals the population consists of and say exactly which individuals fall in the population. If the information given is not complete, complete the description of the population in a reasonable way.

(a) A college has changed its core curriculum and wants to obtain detailed feedback information from the students during each of the first 12 weeks of the coming semester. Each week, a random sample of 5 students will be selected to be interviewed.

(b) The American Community Survey (ACS) replaced the census “long form” starting with the 2010 census. The ACS contacts 250,000 addresses by mail each month, with follow-up by phone and in person if there is no response. Each household answers questions about their housing, economic, and social status.

(c) An opinion poll contacts 1161 adults and asks them, “Which political party do you think has better ideas for leading the country in the twenty-first century?”

### 3.63 Interview residents of apartment complexes.

You are planning a report on apartment living in a college town. You decide to select 5 apartment complexes at random for in-depth interviews with residents. Select a simple random sample of 5 of the following apartment complexes. If you use Table B, start at line 126.  RESIDEN

Ashley Oaks	Country View	Mayfair Village
Bay Pointe	Country Villa	Nobb Hill
Beau Jardin	Crestview	Pemberly Courts
Bluffs	Del-Lynn	Peppermill
Brandon Place	Fairington	Pheasant Run
Briarwood	Fairway Knolls	Richfield
Brownstone	Fowler	Sagamore Ridge
Burberry	Franklin Park	Salem Courthouse
Cambridge	Georgetown	Village Manor
Chauncey Village	Greenacres	Waterford Court
Country Squire	Lahr House	Williamsburg

### 3.64 Using GIS to identify mint field conditions.

A Geographic Information System (GIS) is to be used to distinguish different conditions in mint fields. Ground observations will be used to classify regions of each field as either healthy mint, diseased mint, or weed-infested mint. The GIS divides mint-growing areas into regions called pixels. An experimental area contains 200 pixels. For a random sample of 20 pixels, ground measurements will be made to determine the status of the mint, and these observations will be compared with information obtained by the GIS. Select the random sample. If you use Table B, start at line 122 and choose only the first 10 pixels in the sample.



### 3.65 Use the *Simple Random Sample* applet.

After you have labeled the individuals in a population, the *Simple Random Sample* applet automates the task of choosing an SRS. Use the applet to choose the sample in the previous exercise.

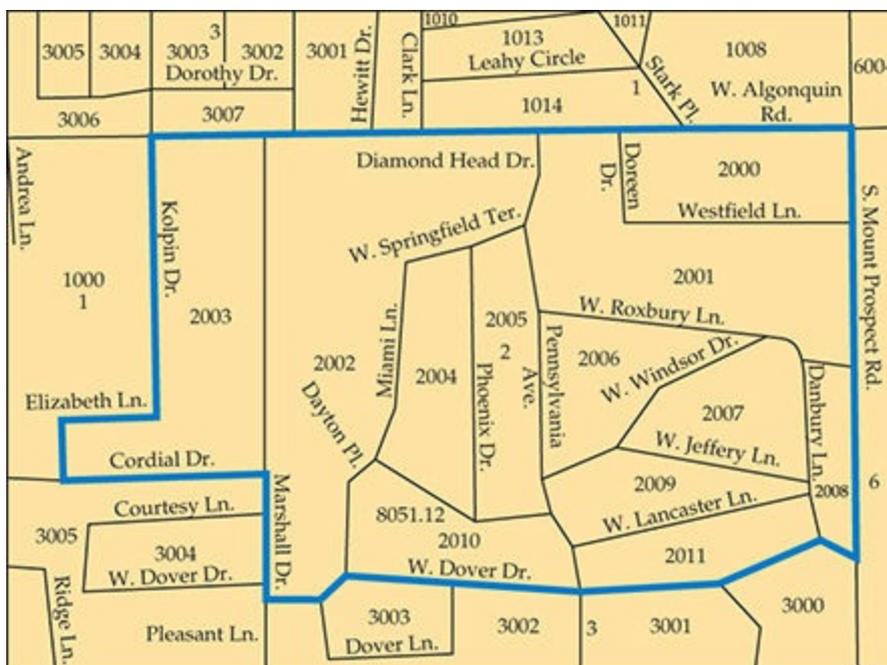


### 3.66 Use the *Simple Random Sample* applet.

There are approximately 405 active telephone area codes covering Canada, the United States, and some Caribbean areas. (More are created regularly.) You want to choose an SRS of 30 of these area codes for a study of available telephone numbers. Label the codes 001 to 405 and use the *Simple Random Sample* applet or software to choose your sample. (If you use Table B, start at line 121 and choose only the first 8 codes in the sample.)

### 3.67 Census tracts.

The U.S. Census Bureau divides the entire country into “census tracts” that contain about 4000 people. Each tract is in turn divided into small “blocks,” which in urban areas are bounded by local streets. An SRS of blocks from a census tract is often the next-to-last stage in a multistage sample. Figure 3.9 shows part of census tract 8051.12, in Cook County, Illinois, west of Chicago. The 44 blocks in this tract are divided into three “block groups.” Group 1 contains 6 blocks numbered 1000 to 1005; Group 2 (outlined in Figure 3.9) contains 12 blocks numbered 2000 to 2011; Group 3 contains 26 blocks numbered 3000 to 3025. Use software or Table B, beginning at line 125, to choose an SRS of 9 of the 44 blocks in this census tract. Explain carefully how you labeled the blocks.



**FIGURE 3.9**

Census blocks in Cook County, Illinois, for Exercises 3.67 and 3.69. The outlined area is a block group.

### 3.68 Repeated use of Table B.

In using Table B repeatedly to choose samples or do randomization for experiments, you should not always begin at the same place, such as line 101. Why not?

### 3.69 A stratified sample.

Exercise 3.67 asks you to choose an SRS of blocks from the census tract pictured in Figure 3.9. You might instead choose a stratified sample of two blocks from the 6 blocks in Group 1, two from the 12 blocks in Group 2, and three from the 26

blocks in Group 3. Choose such a sample, explaining carefully how you labeled blocks and used software or Table B.

### 3.70 Systematic random samples.

**Systematic random samples** are often used to choose a sample of apartments in a large building or dwelling units in a block at the last stage of a multistage sample. An example will illustrate the idea of a systematic sample. Suppose that we must choose 5 addresses out of 125. Because  $125/5 = 25$ , we can think of the list as five lists of 25 addresses. Choose 1 of the first 25 at random, using software or Table B. The sample contains this address and the addresses 25, 50, 75, and 100 places down the list from it. If 13 is chosen, for example, then the systematic random sample consists of the addresses numbered 13, 38, 63, 88, and 113.

(a) A study of dating among college students wanted a sample of 200 of the 8000 single male students on campus. The sample consisted of every 40th name from a list of the 8000 students. Explain why the survey chooses every 40th name.

(b) Use software or Table B at line 112 to choose the starting point for this systematic sample.



### 3.71 Systematic random samples versus simple random samples.

The previous exercise introduces systematic random samples. Explain carefully why a systematic random sample *does* give every individual the same chance to be chosen but is *not* a simple random sample.

### 3.72 Random digit telephone dialing.

An opinion poll in California uses random digit dialing to choose telephone numbers at random. Numbers are selected separately within each California area code. The size of the sample in each area code is proportional to the population living there.  **AREACOD**

(a) What is the name for this kind of sampling design?

(b) California area codes, in rough order from north to south, are

209	213	310	323	341	369	408	415	424	442
510	530	559	562	619	626	627	628	650	657
661	669	707	714	747	752	760	764	805	818
831	858	909	916	925	935	949	951		

Another California survey does not call numbers in all area codes but starts with an SRS of 10 area codes. Choose such an SRS. If you use Table B, start at line 122.

### 3.73 Stratified samples of forest areas.

Stratified samples are widely used to study large areas of forest. Based on satellite images, a forest area in the Amazon basin is divided into 14 types. Foresters studied the four most commercially valuable types: alluvial climax forests of quality levels 1, 2, and 3, and mature secondary forest. They divided the area of each type into large parcels, chose parcels of each type at random, and counted tree species in a 20- by 25-meter rectangle randomly placed within each parcel selected. Here is some detail:

Forest type	Total parcels	Sample size
-------------	---------------	-------------

Climax 1	46	5
Climax 2	62	6
Climax 3	32	3
Secondary	43	4

Choose the stratified sample of 18 parcels. Be sure to explain how you assigned labels to parcels. If you use Table B, start at line 130.

### 3.74 Select club members to go to a convention.

A club has 30 student members and 10 faculty members. The students are

Abel	Fisher	Huber	Moran	Reinmann
Carson	Golomb	Jimenez	Moskowitz	Santos
Chen	Griswold	Jones	Neyman	Shaw
David	Hein	Kiefer	O'Brien	Thompson
Deming	Hernandez	Klotz	Pearl	Utts
Elashoff	Holland	Liu	Potter	Vlasic

and the faculty members are

Andrews	Fernandez	Kim	Moore	Rabinowitz
Besicovitch	Gupta	Lightman	Phillips	Yang

The club can send 6 students and 2 faculty members to a convention and decides to choose those who will go by random selection. Select a stratified random sample of 6 students and 2 faculty members.

### 3.75 Stratified samples for attitudes about alcohol.

At a party there are 32 students over age 21 and 16 students under age 21. You choose at random 4 of those over 21 and separately choose at random 2 of those under 21 to interview about attitudes toward alcohol. You have given every student at the party the same chance to be interviewed: what is that chance? Why is your sample not an SRS?

### 3.76 Stratified samples for accounting audits.

Accountants use stratified samples during audits to verify a company's records of such things as accounts receivable. The stratification is based on the dollar amount of the item and often includes 100% sampling of the largest items. One company reports 5000 accounts receivable. Of these, 100 are in amounts over \$50,000; 500 are in amounts between \$1000 and \$50,000; and the remaining 4400 are in amounts under \$1000. Using these groups as strata, you decide to verify all of the largest accounts and to sample 5% of the midsize accounts and 1% of the small accounts. How would you label the two strata from which you will sample? Use software or Table B, starting at line 125, to select the first 6 accounts from each of these strata.

### 3.77 The sampling frame.

The list of individuals from which a sample is actually selected is called the **sampling frame**. Ideally, the frame should list every individual in the population, but in practice this is often difficult. A frame that leaves out part of the population is a common source of undercoverage.

(a) Suppose that a sample of households in a community is selected at random from the telephone directory. What households are omitted from this frame? What types of people do you think are likely to live in these households? These people will probably be underrepresented in the sample.

(b) It is usual in telephone surveys to use random digit dialing equipment that selects the last four digits of a telephone number at random after being given the area code and the exchange. The exchange is the first three digits of the telephone number. Which of the households that you mentioned in your answer to (a) will be included in the sampling frame by random digit dialing?

### **3.78 Survey questions.**

Comment on each of the following as a potential sample survey question. Is the question clear? Is it slanted toward a desired response?

(a) "Some cell phone users have developed brain cancer. Should all cell phones come with a warning label explaining the danger of using cell phones?"

(b) "Do you agree that a national system of health insurance should be favored because it would provide health insurance for everyone and would reduce administrative costs?"

(c) "In view of escalating environmental degradation and incipient resource depletion, would you favor economic incentives for recycling of resource-intensive consumer goods?"

## 3.4 Toward Statistical Inference

When you complete this section, you will be able to

- Identify parameters, populations, statistics, and samples and the relationships among these items.
- Use simulation to study a sampling distribution.
- Interpret and use a sampling distribution to describe a property of a statistic.
- Identify bias in a statistic by examining its sampling distribution, and characterize an unbiased estimator of a parameter.
- Describe the relationship between the sample size and the variability of a statistic.
- Identify ways to reduce bias and variability of a statistic.
- Use the margin of error to describe the variability of a statistic.

A market research firm interviews a random sample of 2500 adults. Result: 66% find shopping for clothes frustrating and time-consuming. That's the truth about the 2500 people in the sample. What is the truth about the 235 million American adults who make up the population? Because the sample was chosen at random, it's reasonable to think that these 2500 people represent the entire population fairly well. So the market researchers turn the *fact* that 66% of the *sample* find shopping frustrating into an *estimate* that about 66% of *all adults* feel this way.

That's a basic idea in statistics: use a fact about a sample to estimate the truth about the whole population. We call this **statistical inference** because we infer conclusions about the wider population from data on selected individuals. To think about inference, we must keep straight whether a number describes a sample or a population. Here is the vocabulary we use.

statistical inference

### PARAMETERS AND STATISTICS

A **parameter** is a number that describes the **population**. A parameter is a fixed number, but in practice we do not know its value.

A **statistic** is a number that describes a **sample**. The value of a statistic is known when we have taken

a sample, but it can change from sample to sample. We often use a statistic to estimate an unknown parameter.

## Example

### 3.29 Building a customer base.

The Futures Company provides clients with research about maintaining and improving their business. They use a web interface to collect data from between 1000 and 2500 potential customers using 30- to 40-minute surveys.<sup>26</sup> Let's assume that 1650 out of 2500 potential customers in a sample show strong interest in a product. The proportion of the sample who are interested is

$$\hat{p} = \frac{1650}{2500} = 0.66 = 66\%$$

The number  $\hat{p} = 0.66$  is a *statistic*. The corresponding *parameter* is the proportion (call it  $p$ ) of all potential customers who would have expressed interest in this product if they had been asked. We don't know the value of the parameter  $p$ , so we use the statistic  $\hat{p}$  to estimate it.

## USE YOUR KNOWLEDGE

### 3.79 Sexual harassment of college students.

A recent survey of undergraduate college students reports that 62% of female college students and 61% of male college students say they have encountered some type of sexual harassment at their college.<sup>27</sup> Describe the samples and the populations for the survey.

### 3.80 Web polls.

If you connect to the website [boston.cbslocal.com/wbz-daily-poll](http://boston.cbslocal.com/wbz-daily-poll), you will be given the opportunity to give your opinion about a different question of public interest each day. Can you apply the ideas about populations and samples that we have just discussed to this poll? Explain why or why not.

## Sampling variability

If the Futures Company took a second random sample of 2500 customers, the new sample would have different people in it. It is almost certain that there would not be exactly 1650 positive responses. That is, the value of the statistic  $\hat{p}$  will vary from sample to sample. This basic fact is called **sampling variability**: the value of a statistic varies in repeated random sampling. Could it happen that one random sample finds that 66% of potential customers are interested in this product and a second random sample finds that only 42% expressed interest?

### sampling variability

Random samples eliminate *bias* from the act of choosing a sample, but they can still be wrong because of the *variability* that results when we choose at random. If the variation when we take repeat samples from the same population is too great, we can't trust the results of any one sample.

We are saved by the second great advantage of random samples. The first advantage is that choosing at random eliminates favoritism. That is, random sampling attacks bias. The second advantage is that if we take lots of random samples of the same size from the same population, the variation from sample to sample will follow a predictable pattern. **All statistical inference is based on one idea: to see how trustworthy a procedure is, ask what would happen if we repeated it many times.**

To understand why sampling variability is not fatal, we ask, "What would happen if we took many samples?" Here's how to answer that question:

- Take a large number of samples from the same population.
- Calculate the sample proportion  $\hat{p}$  for each sample.
- Make a histogram of the values of  $\hat{p}$ .
- Examine the distribution displayed in the histogram for shape, center, and spread, as well as outliers or other deviations.

In practice it is too expensive to take many samples from a large population such as all adult U.S. residents. But we can imitate taking many samples by using random digits. Using random digits from a table or computer software to imitate chance behavior is called **simulation**.

### simulation

---

## Example

### 3.30 Simulate a random sample.

We will simulate drawing simple random samples (SRSs) of size 100 from the population of potential customers. Suppose that in fact 60% of the population have interest in the product. Then the true value of the parameter we want to estimate is  $p = 0.6$ . (Of course, we would not sample in practice if we already knew that  $p = 0.6$ . We are sampling here to understand how sampling behaves.)

We can imitate the population by a table of random digits, with each entry standing for a person. Six of the 10 digits (say 0 to 5) stand for people who have interest in the product. The remaining four digits, 6 to 9, stand for those who do not. Because all digits in a random number table are equally likely, this assignment produces a population proportion of potential customers equal to  $p = 0.6$ . We then simulate an SRS of 100 people from the population by taking 100 consecutive digits from Table B. The statistic  $\hat{p}$  is the proportion of 0s to 5s in the sample.

Here are the first 100 entries in Table B, with digits 0 to 5 highlighted:

19223	95034	05756	28713	96409	12531	42544	82853
73676	47150	99400	01927	27754	42648	82425	36290
45467	71709	77558	00095				

There are 64 digits between 0 and 5, so  $\hat{p} = 64/100 = 0.64$ . A second SRS based on the second 100 entries in Table B gives a different result,  $\hat{p} = 0.55$ . The two sample results are different, and neither is equal to the true population value  $p = 0.6$ . That's sampling variability.

## Sampling distributions

Simulation is a powerful tool for studying chance. Now that we see how simulation works, it is faster to abandon Table B and to use a computer programmed to generate random numbers.

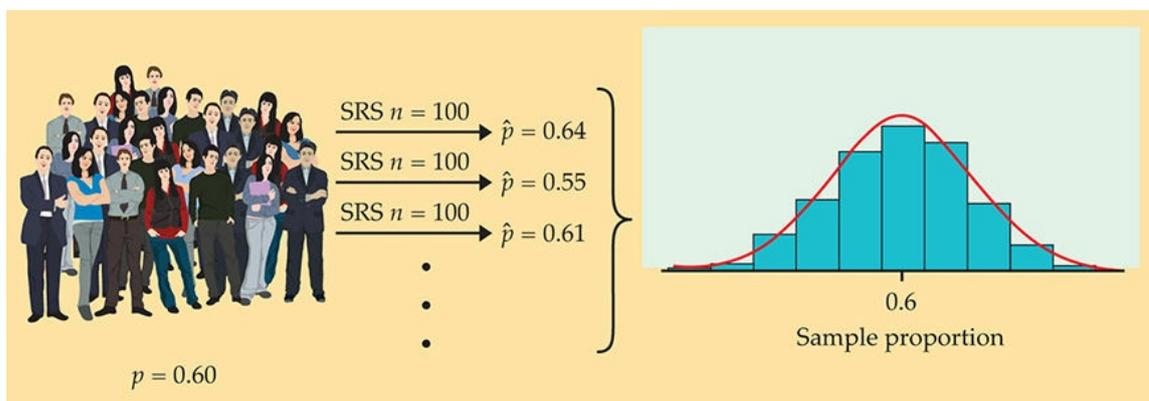
### Example

#### 3.31 Take many random samples.

Figure 3.10 illustrates the process of choosing many samples and finding the sample proportion  $\hat{p}$  for each one. Follow the flow of the figure from the

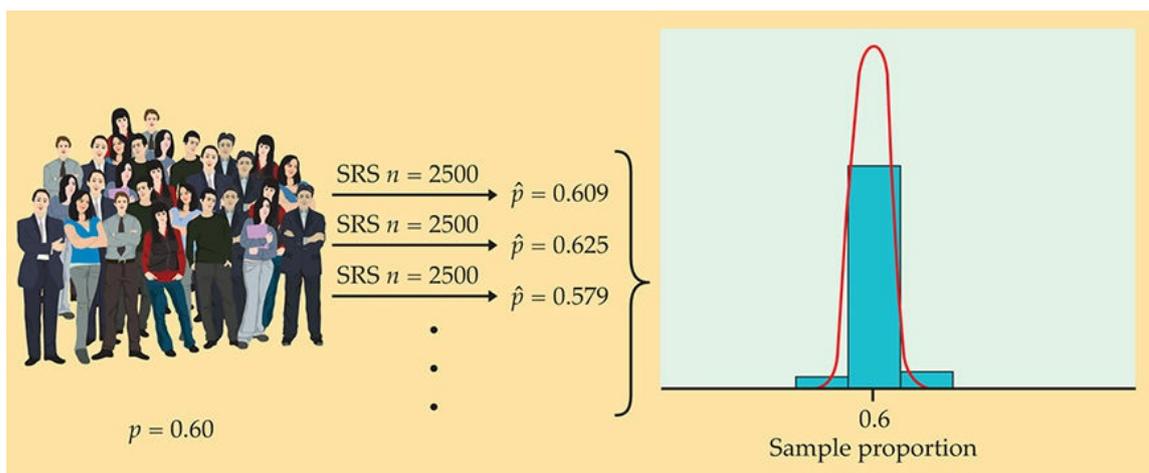
population at the left, to choosing an SRS and finding the  $\hat{p}$  for this sample, to collecting together the  $\hat{p}$ 's from many samples. The histogram at the right of the figure shows the distribution of the values of  $\hat{p}$  from 1000 separate SRSs of size 100 drawn from a population with  $p = 0.6$ .

Of course, the Futures Company samples 2500 people, not just 100. Figure 3.11 is parallel to Figure 3.10. It shows the process of choosing 1000 SRSs, each of size 2500, from a population in which the true proportion is  $p = 0.6$ . The 1000 values of  $\hat{p}$  from these samples form the histogram at the right of the figure. Figures 3.10 and 3.11 are drawn on the same scale. Comparing them shows what happens when we increase the size of our samples from 100 to 2500. These histograms display the *sampling distribution* of the statistic  $\hat{p}$  for two sample sizes.



**FIGURE 3.10**

The results of many SRSs have a regular pattern. Here we draw 1000 SRSs of size 100 from the same population. The population proportion is  $p = 0.60$ . The histogram shows the distribution of 1000 sample proportions.



**FIGURE 3.11**

The distribution of sample proportions for 1000 SRSs of size 2500 drawn from the same population as in Figure 3.10. The two histograms have the same scale. The statistic from the larger sample is less variable.

## SAMPLING DISTRIBUTION

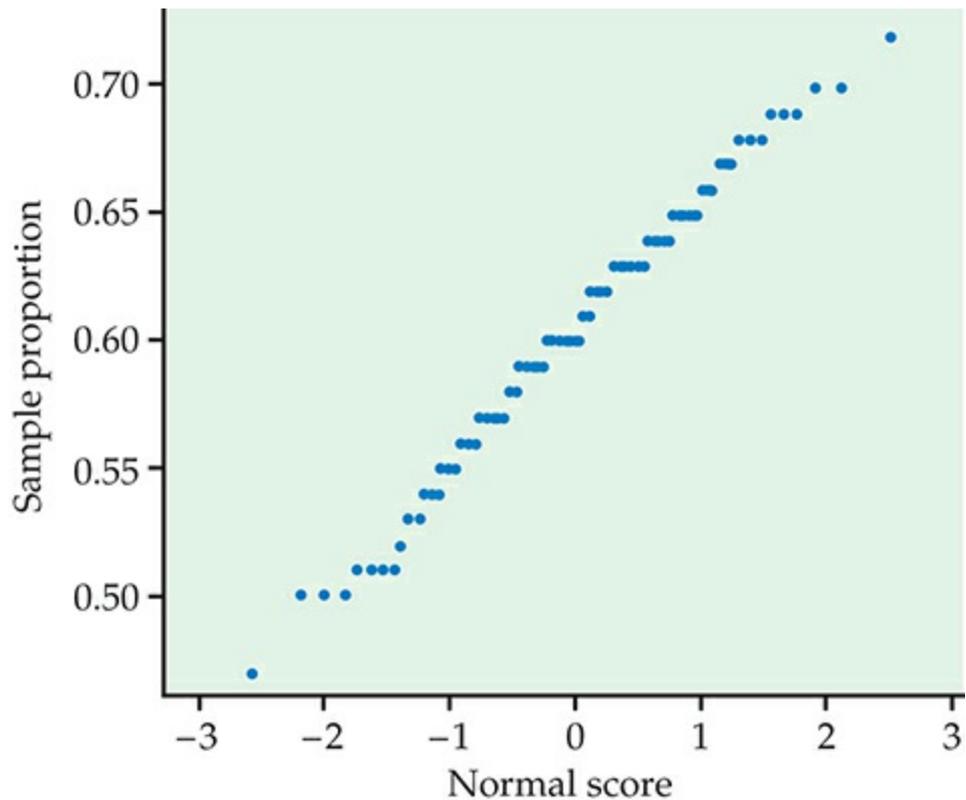
The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

Strictly speaking, the sampling distribution is the ideal pattern that would emerge if we looked at all possible samples of the same size from our population. A distribution obtained from a fixed number of trials, like the 1000 trials in Figures 3.10 and 3.11, is only an approximation to the sampling distribution.

We will see that probability theory, the mathematics of chance behavior, can sometimes describe sampling distributions exactly. The interpretation of a sampling distribution is the same, however, whether we obtain it by simulation or by the mathematics of probability.

We can use the tools of data analysis to describe any distribution. Let's apply those tools to Figures 3.10 and 3.11.

- **Shape:** The histograms look Normal. Figure 3.12 is a Normal quantile plot of the values  $\hat{p}$  for our samples of size 100. It confirms that the distribution in Figure 3.10 is close to Normal. The 1000 values for samples of size 2500 in Figure 3.11 are even closer to Normal. The Normal curves drawn through the histograms describe the overall shapes quite well.
- **Center:** In both cases, the values of the sample proportion  $\hat{p}$  vary from sample to sample, but the values are centered at 0.6. Recall that  $p = 0.6$  is the true population parameter. Some samples have a  $\hat{p}$  less than 0.6 and some greater, but there is no tendency to be always low or always high. That is,  $\hat{p}$  has no **bias** as an estimator of  $p$ . This is true for both large and small samples. (Want the details? The mean of the 1000 values of  $\hat{p}$  is 0.598 for samples of size 100 and 0.6002 for samples of size 2500. The median value of  $\hat{p}$  is exactly 0.6 for samples of both sizes.)



**FIGURE 3.12**

Normal quantile plot of the sample proportions in Figure 3.10. The distribution is close to Normal except for some clustering due to the fact that sample proportions from a sample size of 100 can take only values that are multiples of 0.01. Because a plot of 1000 points is hard to read, this plot presents only every 10th value.

- **Spread:** The values of  $\hat{p}$  from samples of size 2500 are much less spread out than the values from samples of size 100. In fact, the standard deviations are 0.051 for Figure 3.10 and 0.0097, or about 0.01, for Figure 3.11.

Although these results describe just two sets of simulations, they reflect facts that are true whenever we use random sampling.

## USE YOUR KNOWLEDGE

### 3.81 Should you choose 300 observations or 700 observations?

You are planning a study and are considering taking an SRS of either 300 or 700 observations. Explain how the sampling distribution would differ for these two scenarios.

## Bias and variability

Our simulations show that a sample of size 2500 will almost always give an estimate  $\hat{p}$  that is close to the truth about the population. Figure 3.11 illustrates this fact for just one value of the population proportion, but it is true for any proportion. Samples of size 100, on the other hand, might give an estimate of 50% or 70% when the truth is 60%.

Thinking about Figures 3.10 and 3.11 helps us restate the idea of bias when we use a statistic like  $\hat{p}$  to estimate a parameter like  $p$ . It also reminds us that variability matters as much as bias.

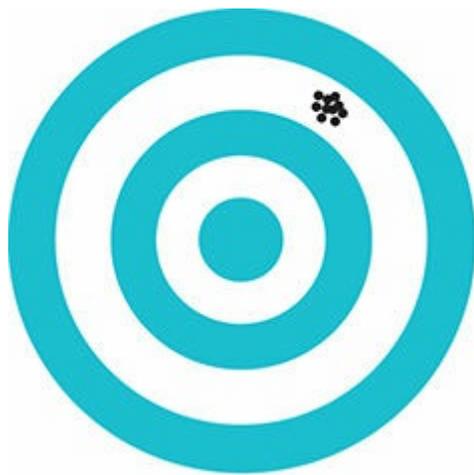
### BIAS AND VARIABILITY OF A STATISTIC

**Bias** concerns the center of the sampling distribution. A statistic used to estimate a parameter is an **unbiased estimator** if the mean of its sampling distribution is equal to the true value of the parameter being estimated.

The **variability of a statistic** is described by the spread of its sampling distribution. This spread is determined by the sampling design and the sample size  $n$ . Statistics from larger probability samples have smaller spreads.

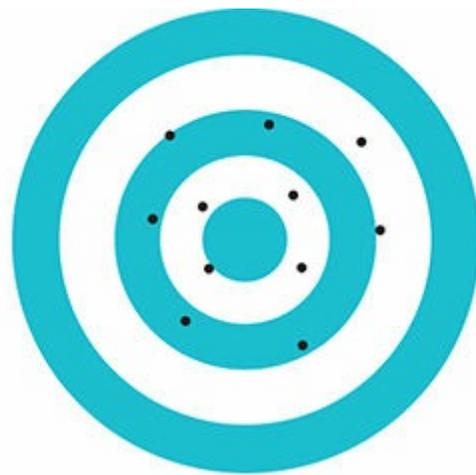
The **margin of error** is a numerical measure of the spread of a sampling distribution. It can be used to set bounds on the size of the likely error in using the statistic as an estimator of a population parameter.

We can think of the true value of the population parameter as the bull's-eye on a target, and of the sample statistic as an arrow fired at the bull's-eye. Bias and variability describe what happens when an archer fires many arrows at the target. *Bias* means that the aim is off, and the arrows land consistently off the bull's-eye in the same direction. The sample values do not center about the population value. Large *variability* means that repeated shots are widely scattered on the target. Repeated samples do not give similar results but differ widely among themselves. Figure 3.13 shows this target illustration of the two types of error.



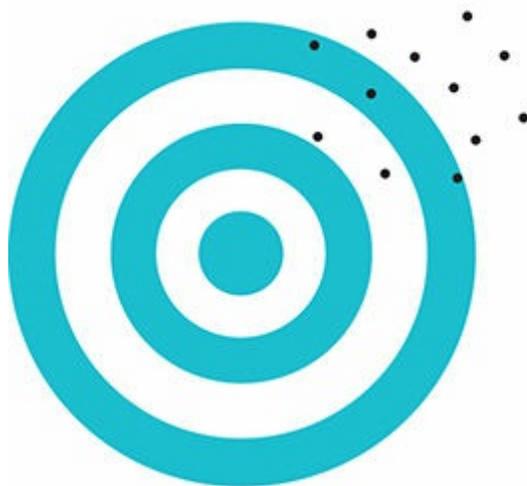
High bias, low variability

(a)



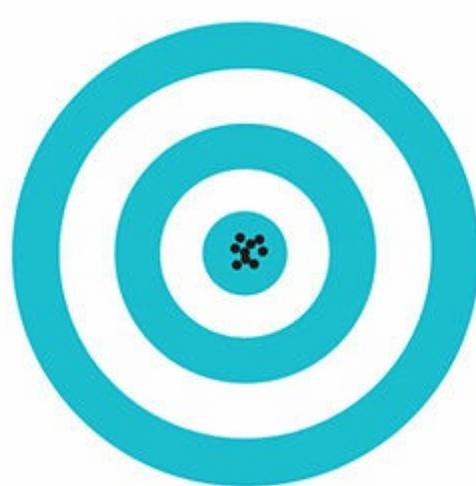
Low bias, high variability

(b)



High bias, high variability

(c)



The ideal: low bias, low variability

(d)

**FIGURE 3.13**

Bias and variability in shooting arrows at a target. Bias means the archer systematically misses in the same direction. Variability means that the arrows are scattered.

Notice that small variability (repeated shots are close together) can accompany large bias (the arrows are consistently away from the bull's-eye in one direction). And small bias (the arrows center on the bull's-eye) can accompany large variability (repeated shots are widely scattered). A good sampling scheme, like a good archer, must have both small bias and small variability. Here's how we do this.

## MANAGING BIAS AND VARIABILITY

**To reduce bias**, use random sampling. When we start with a list of the entire population, simple random sampling produces unbiased estimates—the values of a statistic computed from an SRS neither consistently overestimate nor consistently underestimate the value of the population parameter.

**To reduce the variability** of a statistic from an SRS, use a larger sample. You can make the variability as small as you want by taking a large enough sample.

In practice, the Futures Company takes only one sample. We don't know how close to the truth an estimate from this one sample is because we don't know what the truth about the population is. But *large random samples almost always give an estimate that is close to the truth*. Looking at the pattern of many samples shows that we can trust the result of one sample.

The Current Population Survey's sample of 60,000 households estimates the national unemployment rate very accurately. Of course, only probability samples carry this guarantee. Using a probability sampling design and taking care to deal with practical difficulties reduce bias in a sample.

The size of the sample then determines how close to the population truth the sample result is likely to fall. Results from a sample survey usually come with a margin of error that sets bounds on the size of the likely error. The margin of error directly reflects the variability of the sample statistic, so it is smaller for larger samples. We will describe the details in later chapters.

In many areas where statistical methods are used, there is another way to reduce variability. The trick is to redefine the problem. Here is an example.

## Example

### 3.32 Study protocols.

Many studies that involve people start with a very clear definition of the characteristics of the individuals who are eligible to be subjects in the study. So, for example, a study designed to evaluate a new treatment for high blood pressure might exclude persons who have had a heart attack and persons under the age of 30 or over the age of 60. Exclusions of this type reduce the variability of the population and thereby generally reduce the variability of the statistics that are computed from the data provided by the subjects in the study.

## Sampling from large populations

The Futures Company's sample of 2500 adults is only about 1 out of every 94,000 adults in the United States. Does it matter whether we sample 1-in-100 individuals in the population or 1-in-94,000?

## POPULATION SIZE DOESN'T MATTER

The variability of a statistic from a random sample does not depend on the size of the population, as long as the population is at least 100 times larger than the sample.

Why does the size of the population have little influence on the behavior of statistics from random samples? To see why this is plausible, imagine sampling harvested corn by thrusting a scoop into a lot of corn kernels. The scoop doesn't know whether it is surrounded by a bag of corn or by an entire truckload. As long as the corn is well mixed (so that the scoop selects a random sample), the variability of the result depends only on the size of the scoop.

The fact that the variability of sample results is controlled by the size of the sample has important consequences for sampling design. An SRS of size 2500 from the 235 million adult residents of the United States gives results as precise as an SRS of size 2500 from the 665,000 adult inhabitants of San Francisco. This is good news for designers of national samples but bad news for those who want accurate information about the citizens of San Francisco. If both use an SRS, both must use the same size sample to obtain equally trustworthy results.

## Why randomize?

Why randomize? The act of randomizing guarantees that the results of analyzing our data are subject to the laws of probability. The behavior of statistics is described by a sampling distribution. The form of the distribution is known, and in many cases is approximately Normal. Often the center of the distribution lies at the true parameter value, so that the notion that randomization eliminates bias is made more precise. The spread of the distribution describes the variability of the statistic and can be made as small as we wish by choosing a large enough sample. In a randomized experiment, we can reduce variability by choosing larger groups of subjects for each treatment.

These facts are at the heart of formal statistical inference. Later chapters will have much to say in more technical language about sampling distributions and the way statistical conclusions are based on them. What any user of statistics must understand is that all the technical talk has its basis in a simple question: *What would happen if the sample or the experiment were repeated many times?* The reasoning applies not only to an SRS but also to the complex sampling designs actually used by opinion polls and other national sample surveys. The same

conclusions hold as well for randomized experimental designs. The details vary with the design but the basic facts are true whenever randomization is used to produce data.



Remember that proper statistical design is not the only aspect of a good sample or experiment. *The sampling distribution shows only how a statistic varies due to the operation of chance in randomization. It reveals nothing about possible bias due to undercoverage or nonresponse in a sample or to lack of realism in an experiment.* The actual error in estimating a parameter by a statistic can be much larger than the sampling distribution suggests. What is worse, there is no way to say how large the added error is. The real world is less orderly than statistics textbooks imply.

---

## BEYOND THE BASICS

### Capture-recapture sampling

Sockeye salmon return to reproduce in the river where they were hatched four years earlier. How many salmon survived natural perils and heavy fishing to make it back this year? How many mountain sheep are there in Colorado? Are migratory songbird populations in North America decreasing or holding their own? These questions concern the size of animal populations. Biologists address them with a special kind of repeated sampling, called *capture-recapture sampling*.

---

### Example

#### 3.33 Estimate the number of least flycatchers.



You are interested in the number of least flycatchers migrating along a major route in the north-central United States. You set up “mist nets” that capture the birds but do not harm them. The birds caught in the net are fitted with a small aluminum leg band and released. Last year you banded and released 200 least flycatchers. This year you repeat the process. Your net catches 120 least flycatchers, 12 of which have tags from last year’s catch.

The proportion of your second sample that have bands should estimate the proportion in the entire population that are banded. So if  $N$  is the unknown number of least flycatchers, we should have approximately

$$\text{proportion banded in sample} = \text{proportion banded in population}$$

$$\frac{12}{120} = \frac{200}{N}$$

Solve for  $N$  to estimate that the total number of flycatchers migrating while your net was up this year is approximately

$$N = 200 \times \frac{120}{12} = 2000$$

The capture-recapture idea extends the use of a sample proportion to estimate a population proportion. The idea works well if both samples are SRSs from the population and the population remains unchanged between samples. In practice, complications arise because, for example, some of the birds tagged last year died before this year’s migration.

Variations on capture-recapture samples are widely used in wildlife studies and are now finding other applications. One way to estimate the census undercount in a district is to consider the census as “capturing and marking” the households that respond. Census workers then visit the district, take an SRS of households, and see how many of those counted by the census show up

in the sample. Capture-recapture estimates the total count of households in the district. As with estimating wildlife populations, there are many practical pitfalls. Our final word is as before: the real world is less orderly than statistics textbooks imply.

### SECTION 3.4 Summary

A number that describes a population is a **parameter**. A number that can be computed from the data is a **statistic**. The purpose of sampling or experimentation is usually **inference**: use sample statistics to make statements about unknown population parameters.

A statistic from a probability sample or randomized experiment has a **sampling distribution** that describes how the statistic varies in repeated data production. The sampling distribution answers the question “What would happen if we repeated the sample or experiment many times?” Formal statistical inference is based on the sampling distributions of statistics.

A statistic as an estimator of a parameter may suffer from **bias** or from high **variability**. Bias means that the center of the sampling distribution is not equal to the true value of the parameter. The variability of the statistic is described by the spread of its sampling distribution. Variability is usually reported by giving a **margin of error** for conclusions based on sample results.

Properly chosen statistics from randomized data production designs have no bias resulting from the way the sample is selected or the way the experimental units are assigned to treatments. We can reduce the variability of the statistic by increasing the size of the sample or the size of the experimental groups.

### SECTION 3.4 Exercises

For Exercises 3.79 and 3.80, see page 206; and for Exercise 3.81, see page 210.

#### 3.82 What population and sample?

Twenty fourth-year students from your college who are majoring in English are randomly selected to be on a committee to evaluate changes in the statistics requirement for the major. There are 76 fourth-year English majors at your college. The current rules say that a statistics course is one of four options for a quantitative competency requirement. The proposed change would be to require a statistics course. Each of the committee members is asked to vote Yes or No on the new requirement.

- Describe the population for this setting.
- What is the sample?
- Describe the statistic and how it would be calculated.
- What is the population parameter?
- Write a short summary based on your answers to parts (a) to (d) using this setting to explain population, sample, parameter, statistic, and the relationships among these items.

### 3.83 Simulate a sampling distribution.

In Exercise 1.122 (page 74), you examined the density curve for a uniform distribution. Let's simulate taking samples of size 2 from this distribution.

- (a) Use the RAND() function in Excel or similar software to generate 100 samples from this distribution. Put these in the first column. Generate another 100 samples from this distribution and put these in the second sample in the second column. Calculate the mean of the entries in the first and second columns and put these in the third column. Now, you have 100 samples of the mean of two uniform variables in the third column of your spreadsheet.
- (b) Examine the distribution of the means of samples of size 2 from the uniform distribution using your simulation of 100 samples. Using the graphical and numerical summaries that you learned in Chapter 1, describe the shape, center, and spread of this distribution.
- (c) The theoretical (population) mean for this distribution is 0.5. How close is your simulation estimate to this parameter value?
- (d) The theoretical (population) standard deviation for this distribution is the square root of  $1/24$ . How close is your simulation estimate to this parameter value?

### 3.84 What is the effect of increasing the number of simulations?

Refer to the previous exercise. Increase the number of simulations from 100 to 500. Compare your results with those that you found in the previous exercise. Write a report summarizing your findings. Include a comparison with the results from the previous exercise and a recommendation regarding whether or not a larger number of simulations is needed to answer the questions that we have regarding this sampling distribution.

### 3.85 Change the sample size to 12.

Refer to Exercise 3.83. Change the sample size to 12 and answer parts (a) through (d) of that exercise. Note that the population mean is still 0.5 but the population standard deviation is 1. Explain the effect of increasing the sample size from 2 to 12 using the results from Exercise 3.83 and what you have found in this exercise.

### 3.86 Increase the number of simulations.

Refer to the previous exercise and to Exercise 3.84. Use 500 simulations to study the sampling distribution of the mean of a sample of size 12 from a uniform distribution. Write a summary of what you have found.

### 3.87 Normal distributions.

Many software packages generate standard Normal variables by taking the sum of 12 uniform variables and subtracting 6.

- (a) Simulate this distribution.
- (b) Use numerical and graphical summaries to assess how well this distribution approximates the standard Normal distribution.
- (c) Write a short summary of your work. Include details of your simulation.

### 3.88 Is it unbiased?

A statistic has a sampling distribution that is somewhat skewed. The median is 5 and the quartiles are 2 and 10. The mean is 8.

- (a) If the population parameter is 5, is the estimator unbiased?
- (b) If the population parameter is 10, is the estimator unbiased?
- (c) If the population parameter is 8, is the estimator unbiased?
- (d) Write a short summary of your results in parts (a) to (c) and include a discussion of bias and unbiased estimators.

### 3.89 The effect of the sample size.

Refer to Exercise 3.83 where you simulated the sampling distribution of the mean of two uniform variables and Exercise 3.85 where you simulated the sampling distribution of the mean of 12 uniform variables.

- (a) Based on what you know about the effect of the sample size on the sampling distribution, which simulation should have the smaller variability?
- (b) Did your simulations confirm your answer in part (a)? Write a short paragraph about the effect of the sample size on the variability of a sampling distribution using these simulations to illustrate the basic idea. Be sure to include how you assessed the variability of the sampling distributions.

### 3.90 What's wrong?

State what is wrong in each of the following scenarios.

- (a) A parameter describes a sample.
- (b) Bias and variability are two names for the same thing.
- (c) Large samples are always better than small samples.
- (d) A sampling distribution is something generated by a computer.

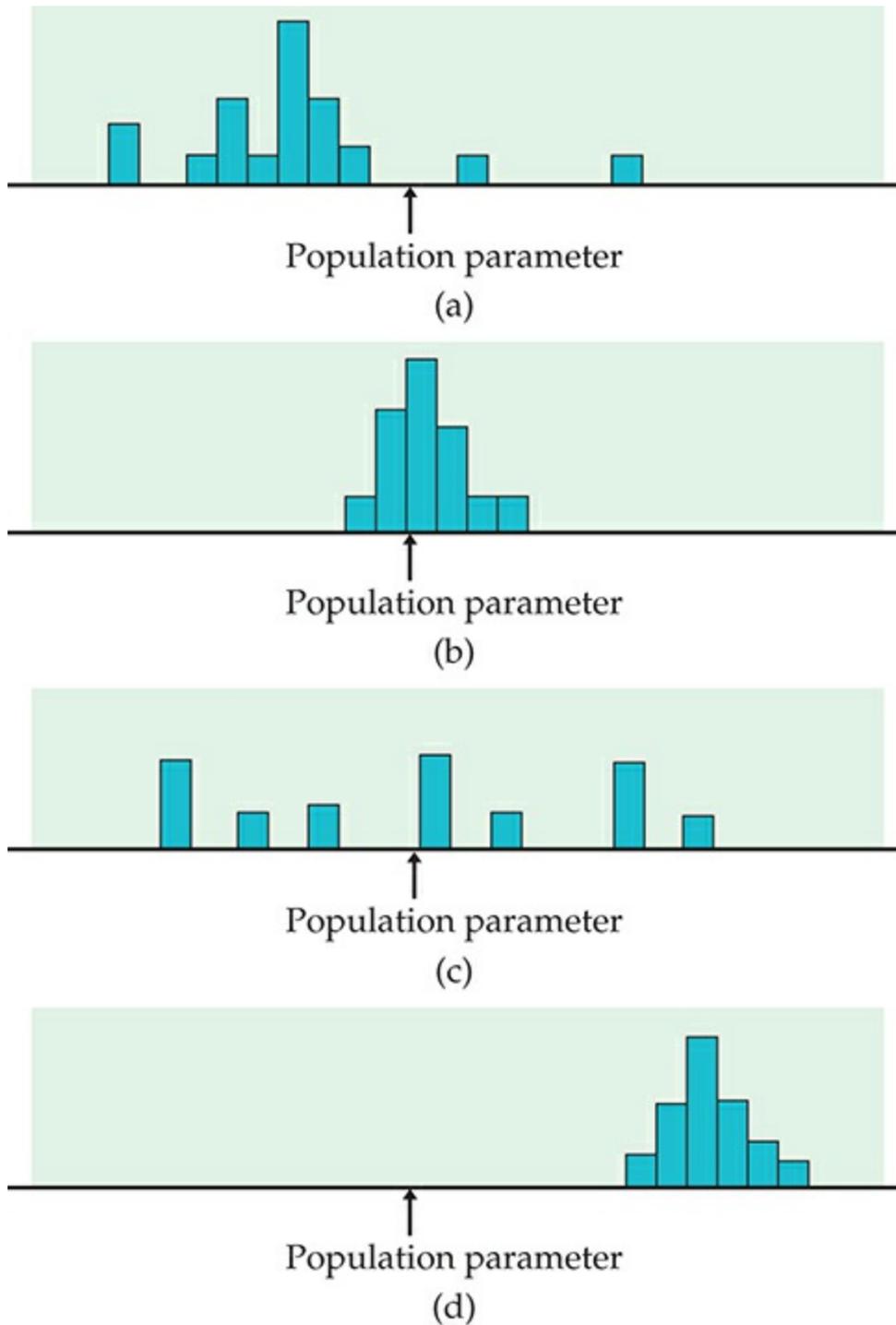
### 3.91 Describe the population and the sample.

For each of the following situations, describe the population and the sample.

- (a) A survey of 17,096 students in U.S. four-year colleges reported that 19.4% were binge drinkers.
- (b) In a study of work stress, 100 restaurant workers were asked about the impact of work stress on their personal lives.
- (c) A tract of forest has 584 longleaf pine trees. The diameters of 40 of these trees were measured.

### 3.92 Bias and variability.

Figure 3.14 shows histograms of four sampling distributions of statistics intended to estimate the same parameter. Label each distribution relative to the others as high or low bias and as high or low variability.



**FIGURE 3.14**

Determine which of these sampling distributions displays high or low bias and high or low variability, for Exercise 3.92.



**3.93 Use the *Probability* applet.**

The *Probability* applet simulates tossing a coin, with the advantage that you can choose the true long-term proportion, or probability, of a head. Suppose that we have a population in which proportion  $p = 0.6$  (the parameter) approve of legal gambling. Tossing a coin with probability  $p = 0.6$  of a head simulates this situation: each head is a person who approves of legal gambling, and each tail is a person who does not.

Set the “Probability of heads” in the applet to 0.6 and the number of tosses to 25. This simulates an SRS of size 25 from this population. By alternating between “Toss” and “Reset” you can take many samples quickly.

(a) Take 50 samples, recording the number of heads in each sample. Make a histogram of the 50 sample proportions (count of heads divided by 25). You are constructing the sampling distribution of this statistic.

(b) Another population contains only 20% who approve of legal gambling. Take 50 samples of size 25 from this population, record the number in each sample who approve, and make a histogram of the 50 sample proportions. How do the centers of your two histograms reflect the differing truths about the two populations?

### 3.94 Use statistical software for simulations.

Statistical software can speed simulations. We are interested in the sampling distribution of the proportion  $\hat{p}$  of people who find shopping frustrating in an SRS from a population in which proportion  $p$  find shopping frustrating. Here,  $p$  is a parameter and  $\hat{p}$  is a statistic used to estimate  $p$ . We will see in Chapter 5 that “binomial” is the key word to look for in the software menus.

(a) Set  $n = 50$  and  $p = 0.6$  and generate 100 binomial observations. These are the counts for 100 SRSs of size 50 when 60% of the population find shopping frustrating. Save these counts and divide them by 50 to get values of  $\hat{p}$  from 100 SRSs. Make a stemplot of the 100 values of  $\hat{p}$ .

(b) Repeat this process with  $p = 0.3$ , representing a population in which only 30% of people find shopping frustrating. Compare your two stemplots. How does changing the parameter  $p$  affect the center and spread of the sampling distribution?

(c) Now generate 100 binomial observations with  $n = 200$  and  $p = 0.6$ . This simulates 100 SRSs, each of size 200. Obtain the 100 sample proportions  $\hat{p}$  and make a stemplot. Compare this with your stemplot from (a). How does changing the sample size  $n$  affect the center and spread of the sampling distribution?



### 3.95 Use the *Simple Random Sample* applet.

The *Simple Random Sample* applet can illustrate the idea of a sampling distribution. Form a population labeled 1 to 100. We will choose an SRS of 10 of these numbers. That is, in this exercise, the numbers themselves are the population, not just labels for 100 individuals. The mean of the whole numbers 1 to 100 is 50.5. This is the parameter, the mean of the population.

(a) Use the applet to choose an SRS of size 10. Which 10 numbers were chosen? What is their mean? This is a statistic, the sample mean  $\bar{x}$ .

(b) Although the population and its mean 50.5 remain fixed, the sample mean changes as we take more samples. Take another SRS of size 10. (Use the “Reset” button to return to the original population before taking the second sample.) What are the 10 numbers in your sample? What is their mean? This is another value of  $\bar{x}$ .

(c) Take 8 more SRSs from this same population and record their means. You now have 10 values of the sample mean  $\bar{x}$  from 10 SRSs of the same size from the same population. Make a histogram of the 10 values and mark the population mean 50.5 on the horizontal axis. Are your 10 sample values roughly centered at the population value? (If you kept going forever, your  $\bar{x}$ -values would form the sampling distribution of the sample mean; the population mean would indeed be the center of this distribution.)

## 3.5 Ethics

When you complete this section, you will be able to

- Describe the purpose of an institutional review board and what kinds of expertise its members require.
- Describe informed consent and evaluate whether or not it has been given in specific examples.
- Determine when data have been kept confidential in a study.
- Evaluate a clinical trial from the viewpoint of ethics.

The production and use of data, like all human endeavors, raise ethical questions. We won't discuss the telemarketer who begins a telephone sales pitch with "I'm conducting a survey." Such deception is clearly unethical. It enrages legitimate survey organizations, which find the public less willing to talk with them. Neither will we discuss those few researchers who, in the pursuit of professional advancement, publish fake data. There is no ethical question here—faking data to advance your career is just wrong. It will end your career when uncovered.

But just how honest must researchers be about real, unfaked data? Here is an example that suggests the answer is "More honest than they often are."

### Example

#### 3.34 Provide all the critical information.

Papers reporting scientific research are supposed to be short, with no extra baggage. But brevity can allow the researchers to avoid complete honesty about their data. Did they choose their subjects in a biased way? Did they report data on only some of their subjects? Did they try several statistical analyses and report only the ones that looked best? The statistician John Bailar screened more than 4000 medical papers in more than a decade as consultant to the *New England Journal of Medicine*. He says, "When it came to the statistical review, it was often clear that critical information was lacking, and the gaps nearly always had the practical effect of making the authors' conclusions look stronger than they should have."<sup>28</sup> The situation is no doubt worse in fields that screen published work less carefully.

The most complex issues of data ethics arise when we collect data from people.

The ethical difficulties are more severe for experiments that impose some treatment on people than for sample surveys that simply gather information. Trials of new medical treatments, for example, can do harm as well as good to their subjects. Here are some basic standards of data ethics that must be obeyed by any study that gathers data from human subjects, whether sample survey or experiment.

## BASIC DATA ETHICS

The organization that carries out the study must have an **institutional review board** that reviews all planned studies in advance in order to protect the subjects from possible harm.

All individuals who are subjects in a study must give their **informed consent** before data are collected.

All individual data must be kept **confidential**. Only statistical summaries for groups of subjects may be made public.

The law requires that studies funded by the federal government obey these principles. But neither the law nor the consensus of experts is completely clear about the details of their application.

## Institutional review boards

The purpose of an institutional review board is not to decide whether a proposed study will produce valuable information or whether it is statistically sound. The board's purpose is, in the words of one university's board, "to protect the rights and welfare of human subjects (including patients) recruited to participate in research activities."

The board reviews the plan of the study and can require changes. It reviews the consent form to be sure that subjects are informed about the nature of the study and about any potential risks. Once research begins, the board monitors its progress at least once a year.

The most pressing issue concerning institutional review boards is whether their workload has become so large that their effectiveness in protecting subjects drops. There are shorter review procedures for projects that involve only minimal risks to subjects, such as most sample surveys. When a board is overloaded, there is a temptation to put more proposals in the minimal-risk category to speed the work.

## USE YOUR KNOWLEDGE

*The exercises in this section on ethics are designed to help you think about the issues that we are discussing and to formulate some opinions. In general, there are no wrong or right answers, but you need to give reasons for your answers.*

### **3.96 Do these proposals involve minimal risk?**

You are a member of your college's institutional review board. You must decide whether several research proposals qualify for lighter review because they involve only minimal risk to subjects. Federal regulations say that "minimal risk" means that the risks are no greater than "those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests." That's vague. Which of these do you think qualifies as "minimal risk"?

- (a) Draw a drop of blood by pricking a finger in order to measure blood sugar.
- (b) Draw blood from the arm for a full set of blood tests.
- (c) Insert a tube that remains in the arm, so that blood can be drawn regularly.

### **3.97 Who should be on an institutional review board?**

Government regulations require that institutional review boards consist of at least five people, including at least one scientist, one nonscientist, and one person from outside the institution. Most boards are larger, but many contain just one outsider.

- (a) Why should review boards contain people who are not scientists?
- (b) Do you think that one outside member is enough? How would you choose that member? (For example, would you prefer a medical doctor? A member of the clergy? An activist for patients' rights?)

## **Informed consent**

Both words in the phrase "informed consent" are important, and both can be controversial. Subjects must be *informed* in advance about the nature of a study and any risk of harm it may bring. In the case of a sample survey, physical harm is not possible. The subjects should be told what kinds of questions the survey will ask and about how much of their time it will take. Experimenters must tell subjects the nature and purpose of the study and outline possible risks. Subjects must then *consent* in writing.

## Example

### 3.35 Who can give informed consent?

Are there some subjects who can't give informed consent? It was once common, for example, to test new vaccines on prison inmates who gave their consent in return for good-behavior credit. Now we worry that prisoners are not really free to refuse, and the law forbids medical experiments in prisons.

Young children can't give fully informed consent, so the usual procedure is to ask their parents. A study of new ways to teach reading is about to start at a local elementary school, so the study team sends consent forms home to parents. Many parents don't return the forms. Can their children take part in the study because the parents did not say "No," or should we allow only children whose parents returned the form and said "Yes"?

What about research into new medical treatments for people with mental disorders? What about studies of new ways to help emergency room patients who may be unconscious or have suffered a stroke? In most cases, there is not time even to get the consent of the family. Does the principle of informed consent bar realistic trials of new treatments for unconscious patients?

These are questions without clear answers. Reasonable people differ strongly on all of them. There is nothing simple about informed consent.<sup>29</sup>

The difficulties of informed consent do not vanish even for capable subjects. Some researchers, especially in medical trials, regard consent as a barrier to getting patients to participate in research. They may not explain all possible risks; they may not point out that there are other therapies that might be better than those being studied; they may be too optimistic in talking with patients even when the consent form has all the right details.

On the other hand, mentioning every possible risk leads to very long consent forms that really are barriers. "They are like rental car contracts," one lawyer said. Some subjects don't read forms that run five or six printed pages. Others are frightened by the large number of possible (but unlikely) disasters that might happen and so refuse to participate. Of course, unlikely disasters sometimes happen. When they do, lawsuits follow and the consent forms become yet longer and more detailed.

## Confidentiality

Ethical problems do not disappear once a study has been cleared by the review board, has obtained consent from its subjects, and has actually collected data about the subjects. **Confidentiality** means that only the researchers can identify responses of individual subjects. The report of an opinion poll may say what percent of the 1500 respondents felt that legal immigration should be reduced. It

may not report what *you* said about this or any other issue.

### confidentiality

Confidentiality is not the same as **anonymity**. Anonymity means that subjects are anonymous—their names are not known even to the director of the study. Anonymity is rare in statistical studies. Even where anonymity is possible (mainly in surveys conducted by mail), it prevents any follow-up to improve nonresponse or inform subjects of results.

### anonymity

Any breach of confidentiality is a serious violation of data ethics. The best practice is to separate the identity of the subjects from the rest of the data at once. Sample surveys, for example, use the identification only to check on who did or did not respond. In an era of advanced technology, however, it is no longer enough to be sure that each individual set of data protects people's privacy.

The government, for example, maintains a vast amount of information about citizens in many separate databases—census responses, tax returns, Social Security information, data from surveys such as the Current Population Survey, and so on. Many of these databases can be searched by computers for statistical studies.

A clever computer search of several databases might be able, by combining information, to identify you and learn a great deal about you even if your name and other identification have been removed from the data available for search. A colleague from Germany once remarked that “female full professor of statistics with a PhD from the United States” was enough to identify her among all the citizens of Germany. Privacy and confidentiality of data are hot issues among statisticians in the computer age.

## Example

### 3.36 Data collected by the government.

Citizens are required to give information to the government. Think of tax returns and Social Security contributions. The government needs these data for administrative purposes—to see if we paid the right amount of tax and how large a Social Security benefit we are owed when we retire. Some people feel that individuals should be able to forbid any other use of their data, even with all identification removed. This would prevent using government records to study, say, the ages, incomes, and household sizes of Social Security

recipients. Such a study could well be vital to debates on reforming Social Security.

## USE YOUR KNOWLEDGE

### 3.98 How can we obtain informed consent?

A researcher suspects that traditional religious beliefs tend to be associated with an authoritarian personality. She prepares a questionnaire that measures authoritarian tendencies and also asks many religious questions. Write a description of the purpose of this research to be read by subjects in order to obtain their informed consent. You must balance the conflicting goals of not deceiving the subjects as to what the questionnaire will tell about them and of not biasing the sample by scaring off religious people.

### 3.99 Should we allow this personal information to be collected?

In which of the following circumstances would you allow collecting personal information without the subjects' consent?

- (a) A government agency takes a random sample of income tax returns to obtain information on the average income of people in different occupations. Only the incomes and occupations are recorded from the returns, not the names.
- (b) A social psychologist attends public meetings of a religious group to study the behavior patterns of members.
- (c) A social psychologist pretends to be converted to membership in a religious group and attends private meetings to study the behavior patterns of members.

## Clinical trials

Clinical trials are experiments that study the effectiveness of medical treatments on actual patients. Medical treatments can harm as well as heal, so clinical trials spotlight the ethical problems of experiments with human subjects. Here are the starting points for a discussion:

- Randomized comparative experiments are the only way to see the true effects of new treatments. Without them, risky treatments that are no better than placebos will become common.
- Clinical trials produce great benefits, but most of these benefits go to future

patients. The trials also pose risks, and these risks are borne by the subjects of the trial. So we must balance future benefits against present risks.

- Both medical ethics and international human rights standards say that “the interests of the subject must always prevail over the interests of science and society.”

The quoted words are from the 1964 Helsinki Declaration of the World Medical Association, the most respected international standard. The most outrageous examples of unethical experiments are those that ignore the interests of the subjects.

## Example

### 3.37 The Tuskegee study.



In the 1930s, syphilis was common among black men in the rural South, a group that had almost no access to medical care. The Public Health Service Tuskegee study recruited 399 poor black sharecroppers with syphilis and 201

others without the disease in order to observe how syphilis progressed when no treatment was given. Beginning in 1943, penicillin became available to treat syphilis. The study subjects were not treated. In fact, the Public Health Service prevented any treatment until word leaked out and forced an end to the study in the 1970s.

The Tuskegee study is an extreme example of investigators following their own interests and ignoring the well-being of their subjects. A 1996 review said, “It has come to symbolize racism in medicine, ethical misconduct in human research, paternalism by physicians, and government abuse of vulnerable people.” In 1997, President Clinton formally apologized to the surviving participants in a White House ceremony.<sup>30</sup>

Because “the interests of the subject must always prevail,” medical treatments can be tested in clinical trials only when there is reason to hope that they will help the patients who are subjects in the trials. Future benefits aren’t enough to justify experiments with human subjects. Of course, if there is already strong evidence that a treatment works and is safe, it is unethical *not* to give it.

Here are the words of Dr. Charles Hennekens of the Harvard Medical School, who directed the large clinical trial that showed that aspirin reduces the risk of heart attacks:

*There’s a delicate balance between when to do or not do a randomized trial. On the one hand, there must be sufficient belief in the agent’s potential to justify exposing half the subjects to it. On the other hand, there must be sufficient doubt about its efficacy to justify withholding it from the other half of subjects who might be assigned to placebos.*<sup>31</sup>

Why is it ethical to give a control group of patients a placebo? Well, we know that placebos often work. What is more, placebos have no harmful side effects. So in the state of balanced doubt described by Dr. Hennekens, the placebo group may be getting a better treatment than the drug group. If we *knew* which treatment was better, we would give it to everyone. When we don’t know, it is ethical to try both and compare them.

The idea of using a control or placebo is a fundamental principle to be considered in designing experiments. In many situations, deciding what to use as an appropriate control requires some careful thought.



*The choice of the control can have a substantial impact on how the results of an experiment are interpreted. Here is an example.*

## Example

### 3.38 Attentiveness improves by nearly 20%.

The manufacturer of a breakfast cereal designed for children claims that eating this cereal has been clinically shown to improve attentiveness by nearly 20%. The study used two groups of children who were tested before and after breakfast. One group received the cereal for breakfast, while breakfast for the control group was water. The results of the tests taken three hours after breakfast were used in the claim.

The Federal Trade Commission investigated the marketing of this product. They charged that the claim was false and violated federal law. The charges were settled and the company agreed to not use misleading claims in their advertising.<sup>32</sup>

It is not sufficient to obtain appropriate controls. The data must be collected from all groups in the same way. Here is an example of this type of flawed design:

## Example

### 3.39 Accurate identification of ovarian cancer.

Two scientists published a paper claiming to have developed a very exciting new method to detect ovarian cancer using blood samples. When other scientists were unable to reproduce the results in different labs, the original work was examined more carefully. In the original study there were samples for women with ovarian cancer and for healthy controls. The blood samples were all analyzed using a mass spectrometer. The control samples were analyzed on one day, and the cancer samples were analyzed on the next day. This design was flawed in that it could not control for changes over time in the measuring instrument.<sup>33</sup>

## USE YOUR KNOWLEDGE

### **3.100 Is this study ethical?**

Researchers on aging proposed to investigate the effect of supplemental health services on the quality of life of older people. Eligible patients on the rolls of a large medical clinic were to be randomly assigned to treatment and control groups. The treatment group would be offered hearing aids, dentures, transportation, and other services not available without charge to the control group. The review board felt that providing these services to some but not other persons in the same institution raised ethical questions. Do you agree?

### **3.101 Should the treatments be given to everyone?**

Effective drugs for treating AIDS are very expensive, so most African nations cannot afford to give them to large numbers of people. Yet AIDS is more common in parts of Africa than anywhere else. Several clinical trials are looking at ways to prevent pregnant mothers infected with HIV from passing the infection to their unborn children, a major source of HIV infections in Africa. Some people say these trials are unethical because they do not give effective AIDS drugs to their subjects, as would be required in rich nations. Others reply that the trials are looking for treatments that can work in the real world in Africa and that they promise benefits at least to the children of their subjects. What do you think?

## **Behavioral and social science experiments**

When we move from medicine to the behavioral and social sciences, the direct risks to experimental subjects are less acute, but so are the possible benefits to the subjects. Consider, for example, the experiments conducted by psychologists in their study of human behavior.

### **Example**

#### **3.40 Personal space.**

Psychologists observe that people have a “personal space” and get annoyed if others come too close to them. We don’t like strangers to sit at our table in a

coffee shop if other tables are available, and we see people move apart in elevators if there is room to do so. Americans tend to require more personal space than people in most other cultures. Can violations of personal space have physical, as well as emotional, effects?

Investigators set up shop in a men's public rest room. They blocked off urinals to force men walking in to use either a urinal next to an experimenter (treatment group) or a urinal separated from the experimenter (control group). Another experimenter, using a periscope from a toilet stall, measured how long the subject took to start urinating and how long he kept at it.<sup>34</sup>

This personal space experiment illustrates the difficulties facing those who plan and review behavioral studies.

- There is no risk of harm to the subjects, although they would certainly object to being watched through a periscope. What should we protect subjects from when physical harm is unlikely? Possible emotional harm? Undignified situations? Invasion of privacy?
- What about informed consent? The subjects in Example 3.40 did not even know they were participating in an experiment. Many behavioral experiments rely on hiding the true purpose of the study. The subjects would change their behavior if told in advance what the investigators were looking for. Subjects are asked to consent on the basis of vague information. They receive full information only after the experiment.

The "Ethical Principles" of the American Psychological Association require consent unless a study merely observes behavior in a public place. They allow deception only when it is necessary to the study, does not hide information that might influence a subject's willingness to participate, and is explained to subjects as soon as possible. The personal space study (from the 1970s) does not meet current ethical standards.

We see that the basic requirement for informed consent is understood differently in medicine and psychology. Here is an example of another setting with yet another interpretation of what is ethical. The subjects get no information and give no consent. They don't even know that an experiment may be sending them to jail for the night.

## **Example**

### **3.41 Domestic violence.**

How should police respond to domestic-violence calls? In the past, the usual practice was to remove the offender and order him to stay out of the household overnight. Police were reluctant to make arrests because the victims rarely pressed charges. Women's groups argued that arresting offenders would help prevent future violence even if no charges were filed. Is there evidence that arrest will reduce future offenses? That's a question that experiments have tried to answer.

A typical domestic-violence experiment compares two treatments: arrest the suspect and hold him overnight, or warn the suspect and release him. When police officers reach the scene of a domestic-violence call, they calm the participants and investigate. Weapons or death threats require an arrest. If the facts permit an arrest but do not require it, an officer radios headquarters for instructions. The person on duty opens the next envelope in a file prepared in advance by a statistician. The envelopes contain the treatments in random order. The police either arrest the suspect or warn and release him, depending on the contents of the envelope. The researchers then watch police records and visit the victim to see if the domestic violence reoccurs.

The first such experiment appeared to show that arresting domestic-violence suspects does reduce their future violent behavior. As a result of this evidence, arrest has become the common police response to domestic violence.

The domestic-violence experiments shed light on an important issue of public policy. Because there is no informed consent, the ethical rules that govern clinical trials and most social science studies would forbid these experiments. They were cleared by review boards because, in the words of one domestic-violence researcher, "These people became subjects by committing acts that allow the police to arrest them. You don't need consent to arrest someone."

### SECTION 3.5 Summary

Approval of an **institutional review board** is required for studies that involve humans or animals as subjects.

Human subjects must give **informed consent** if they are to participate in experiments.

Data on human subjects must be kept **confidential**.

### SECTION 3.5 Exercises

*For Exercises 3.96 and 3.97, see page 219; for Exercises 3.98 and 3.99, see pages 221–222; and for Exercises 3.100 and 3.101, see page 224.*

#### 3.102 Apply for the IRB.

You have been asked to apply to become a member of the institutional review board (IRB) of your college. Write a short essay explaining your understanding of the purpose of the IRB and how your perspective as a

student would be a valuable addition to the IRB in accomplishing its mission.

### **3.103 Did you give informed consent?**

You were asked to participate in a study by a friend who is recruiting subjects. You trust your friend and you tell her that you are willing to do whatever is needed for the study. Have you given informed consent? Explain your answer.

### **3.104 Are the data confidential?**

You have participated in a study, and the results were published in an article in a very prestigious journal. Only summary information was published. The policy of the journal requires that all data used in the articles they publish be available to the public, and they archive the data on a website. When you examine the data, you realize that you have a unique set of characteristics that would allow someone who knows you very well to identify which data are from you. Someone who does not know you would not be able to do this. Are the data confidential? Explain your answer.

### **3.105 One of the subjects died.**

A subject in a corrective gene study died during the study. A lawsuit, *Gelsinger v. Trustees of the University of Pennsylvania*, was filed claiming wrongful death, assault and battery linked to a lack of informed consent, and common-law fraud linked to the informed-consent process.<sup>35</sup> Discuss this case from the point of view of ethics. Describe any additional information that you would need to form your opinion.

### **3.106 Is the IRB responsible?**

An institutional review board (IRB) approved an experimental cancer vaccine for use in a clinical trial. The subjects were patients who had advanced disease and had received standard treatments with no success. Of the 94 subjects who received the vaccine, 26 died during the study. Their deaths were not due to the vaccine. Some family members of the subjects sued the hospital, the study director, the company that made the vaccine, a university official, individual members of the IRB, and the university bioethicist who consulted with the IRB.<sup>36</sup> Discuss this case from the point of view of ethics. Discuss any additional information that you would need to form your opinion.

### **3.107 Facebook and academic performance.**

*First Monday* is a peer-reviewed journal on the Internet. It published two articles concerning Facebook and academic performance. Visit their website, [firstmonday.org](http://firstmonday.org), and look at the first three articles in Volume 14, Number 5 (May 4, 2009). Identify the key controversial issues that involve the use of statistics addressed in these articles, and write a report summarizing the facts as you see them. Be sure to include your opinions regarding ethical issues related to this work.

### **3.108 What is wrong?**

Explain what is wrong in each of the following scenarios.

- (a) Clinical trials are always ethical as long as they randomly assign patients to the treatments.
- (b) The job of an institutional review board is complete when they decide to allow a study to be conducted.

(c) A treatment that has no risk of physical harm to subjects is always ethical.

### **3.109 How should the samples have been analyzed?**

Refer to the ovarian cancer diagnostic test study in Example 3.39 (page 223). Describe how you would process the samples through the mass spectrometer.

### **3.110 The Vytorin controversy.**

Vytorin is a combination pill designed to lower cholesterol. It consists of a relatively inexpensive and widely used drug, Zocor, and a newer drug called Zetia. Early study results suggested that Vytorin was no more effective than Zetia. Critics claimed that the makers of the drugs tried to change the response variable for the study, and two congressional panels investigated why there was a two-year delay in the release of the results. Use the Internet to search for more information about this controversy and write a report about what you find. Include an evaluation in the framework of ethical use of experiments and data. A good place to start your search would be to look for the phrase “Vytorin’s shortcomings.”

### **3.111 The General Social Survey.**

One of the most important nongovernment surveys in the United States is the National Opinion Research Center’s General Social Survey. The GSS regularly monitors public opinion on a wide variety of political and social issues. Interviews are conducted in person in the subject’s home. Are a subject’s responses to GSS questions anonymous, confidential, or both? Explain your answer.

### **3.112 Anonymity and confidentiality in health screening.**

Texas A&M, like many universities, offers free screening for HIV, the virus that causes AIDS. The announcement says, “Persons who sign up for the HIV Screening will be assigned a number so that they do not have to give their name.” They can learn the results of the test by telephone, still without giving their name. Does this practice offer *anonymity* or just *confidentiality*?

### **3.113 Anonymity and confidentiality in mail surveys.**

Some common practices may appear to offer anonymity while actually delivering only confidentiality. Market researchers often use mail surveys that do not ask the respondent’s identity but contain hidden codes on the questionnaire that identify the respondent. A false claim of anonymity is clearly unethical. If only confidentiality is promised, is it also unethical to say nothing about the identifying code, perhaps causing respondents to believe their replies are anonymous?

### **3.114 Use of stored blood.**

Long ago, doctors drew a blood specimen from you as part of treating minor anemia. Unknown to you, the sample was stored. Now researchers plan to use stored samples from you and many other people to look for genetic factors that may influence anemia. It is no longer possible to ask your consent. Modern technology can read your entire genetic makeup from the blood sample.

(a) Do you think it violates the principle of informed consent to use your blood sample if your name is on it but you were not told that it might be saved and studied later?

(b) Suppose that your identity is not attached. The blood sample is known only to come from (say) “a 20-

year-old white female being treated for anemia.” Is it now OK to use the sample for research?

(c) Perhaps we should use biological materials such as blood samples only from patients who have agreed to allow the material to be stored for later use in research. It isn't possible to say in advance what kind of research, so this falls short of the usual standard for informed consent. Is it nonetheless acceptable, given complete confidentiality and the fact that using the sample can't physically harm the patient?

### **3.115 Political polls.**

The presidential election campaign is in full swing, and the candidates have hired polling organizations to take regular polls to find out what the voters think about the issues. What information should the pollsters be required to give out?

- (a) What does the standard of informed consent require the pollsters to tell potential respondents?
- (b) The standards accepted by polling organizations also require giving respondents the name and address of the organization that carries out the poll. Why do you think this is required?
- (c) The polling organization usually has a professional name such as “Samples Incorporated,” so respondents don't know that the poll is being paid for by a political party or candidate. Would revealing the sponsor to respondents bias the poll? Should the sponsor always be announced whenever poll results are made public?

### **3.116 Should poll results be made public?**

Some people think that the law should require that all political poll results be made public. Otherwise, the possessors of poll results can use the information to their own advantage. They can act on the information, release only selected parts of it, or time the release for best effect. A candidate's organization replies that they are paying for the poll in order to gain information for their own use, not to amuse the public. Do you favor requiring complete disclosure of political poll results? What about other private surveys, such as market research surveys of consumer tastes?

### **3.117 Informed consent to take blood samples.**

Researchers from Yale, working with medical teams in Tanzania, wanted to know how common infection with the AIDS virus is among pregnant women in that country. To do this, they planned to test blood samples drawn from pregnant women.

Yale's institutional review board insisted that the researchers get the informed consent of each woman and tell her the results of the test. This is the usual procedure in developed nations. The Tanzanian government did not want to tell the women why blood was drawn or tell them the test results. The government feared panic if many people turned out to have an incurable disease for which the country's medical system could not provide care. The study was canceled. Do you think that Yale was right to apply its usual standards for protecting subjects?

## CHAPTER 3 Exercises

### 3.118 Experiments and surveys.

Write a short report describing the differences and similarities between experiments and surveys. Include a discussion of the advantages and disadvantages of each.

### 3.119 Online behavioral advertising.

The Federal Trade Commission Staff Report “Self-Regulatory Principles for Online Behavioral Advertising” defines behavioral advertising as “the tracking of a consumer’s online activities over time—including the searches the consumer has conducted, the Web pages visited and the content viewed—in order to deliver advertising targeted to the individual consumer’s interests.” The report suggests four governing concepts for their proposals. These are (1) transparency and control: when companies collect information from consumers for advertising, they should tell consumers how the data will be collected, and consumers should be given a choice about whether to allow the data to be collected; (2) security and data retention: data should be kept secure and should be retained only as long as they are needed; (3) privacy: before data are used in a way that differs from promises made when they were collected, consent should be obtained from the consumer; and (4) sensitive data: affirmative express consent should be obtained before using any sensitive data.<sup>37</sup> Write a report discussing your opinions concerning online behavioral advertising and the four governing concepts. Pay particular attention to issues related to the ethical collection and use of statistical data.

### 3.120 Confidentiality at NORC.

The National Opinion Research Center conducts a large number of surveys and has established procedures for protecting the confidentiality of their survey participants. For their Survey of Consumer Finances, they provide a pledge to participants regarding confidentiality. This pledge is available at [norc.org](http://norc.org). Review the pledge and summarize its key parts. Do you think that the pledge adequately addresses issues related to the ethical collection and use of data? Explain your answer.

### 3.121 Make it an experiment!

In the following observational studies, describe changes that could be made to the data collection process that would result in an experiment rather than an observational study. Also, offer suggestions about unseen biases or lurking variables that may be present in the studies as they are described here.

(a) A friend of yours likes to play Texas hold ’em. Every time that he tells you about his playing, he says that he won.

(b) In an introductory statistics class you notice that the students who sit in the first two rows of seats had higher scores on the first exam than the other students in the class.

### 3.122 Name the designs.

What is the name for each of these study designs?

- (a) A study to compare two methods of preserving wood started with boards of southern white pine. Each board was ripped from end to end to form two edge-matched specimens. One was assigned to Method A; the other, to Method B.
- (b) A survey on youth and smoking contacted by telephone 300 smokers and 300 nonsmokers, all 14 to 22 years of age.
- (c) Does air pollution induce DNA mutations in mice? Starting with 40 male and 40 female mice, 20 of each sex were housed in a polluted industrial area downwind from a steel mill. The other 20 of each sex were housed at an unpolluted rural location 30 kilometers away.

### **3.123 Price promotions and consumer expectations.**

A researcher studying the effect of price promotions on consumer expectations makes up two different histories of the store price of a hypothetical brand of laundry detergent for the past year. Students in a marketing course view one or the other price history on a computer. Some students see a steady price, while others see regular promotions that temporarily cut the price. Then the students are asked what price they would expect to pay for the detergent. Is this study an experiment? Why? What are the explanatory and response variables?

### **3.124 Calcium and healthy bones.**

Adults need to eat foods or supplements that contain enough calcium to maintain healthy bones. Calcium intake is generally measured in milligrams per day (mg/d), and one measure of healthy bones is total body bone mineral density measured in grams per centimeter squared (TBBMD,  $g/cm^2$ ). Suppose that you want to study the relationship between calcium intake and TBBMD.

- (a) Design an observational study to study the relationship.
- (b) Design an experiment to study the relationship.
- (c) Compare the relative merits of your two designs. Which do you prefer? Give reasons for your answer.

### **3.125 Choose the type of study.**

Give an example of a question about pets and their owners, their behavior, or their opinions that would best be answered by

- (a) a sample survey.
- (b) an observational study that is not a sample survey.
- (c) an experiment.

### **3.126 Compare the fries.**

Do consumers prefer the fries from Burger King or from McDonald's? Design a blind test in which the source of the fries is not identified. Describe briefly the design of a matched pairs experiment to investigate this question. How will you use randomization?

### 3.127 Bicycle gears.

How does the time it takes a bicycle rider to travel 100 meters depend on which gear is used and how steep the course is? It may be, for example, that higher gears are faster on level ground but lower gears are faster on steep inclines. Discuss the design of a two-factor experiment to investigate this issue, using one bicycle with three gears and one rider. How will you use randomization?

### 3.128 Design an experiment.

The previous two exercises illustrate the use of statistically designed experiments to answer questions that arise in everyday life. Select a question of interest to you that an experiment might answer and carefully discuss the design of an appropriate experiment.

### 3.129 Design a survey.

You want to investigate the attitudes of students at your school about the faculty's commitment to teaching. The student government will pay the costs of contacting about 500 students.

- (a) Specify the exact population for your study; for example, will you include part-time students?
- (b) Describe your sample design. Will you use a stratified sample?
- (c) Briefly discuss the practical difficulties that you anticipate; for example, how will you contact the students in your sample?

### 3.130 Compare two doses of a drug.

A drug manufacturer is studying how a new drug behaves in patients. Investigators compare two doses: 5 milligrams (mg) and 10 mg. The drug can be administered by injection, by a skin patch, or by intravenous drip. Concentration in the blood after 30 minutes (the response variable) may depend both on the dose and on the method of administration.

- (a) Make a sketch that describes the treatments formed by combining dose and method. Then use a diagram to outline a completely randomized design for this two-factor experiment.
- (b) "How many subjects?" is a tough issue. We will explain the basic ideas in Chapter 6. What can you say now about the advantage of using larger groups of subjects?

### 3.131 Would the results be different for men and women?

The drug that is the subject of the experiment in Exercise 3.130 may behave differently in men and women. How would you modify your experimental design to take this into account?

### 3.132 Informed consent.

The requirement that human subjects give their informed consent to participate in an experiment can greatly reduce the number of available subjects. For example, a study of new teaching methods asks the consent of parents for their children to be randomly assigned to be taught by either a new method or the standard method. Many parents do not return the forms, so their children must continue to be taught by the standard method. Why is it not correct to consider these children as part of the control group along with children who are randomly assigned to the standard method?

### 3.133 Two ways to ask sensitive questions.

Sample survey questions are usually read from a computer screen. In a Computer Aided Personal Interview (CAPI), the interviewer reads the questions and enters the responses. In a Computer Aided Self Interview (CASI), the interviewer stands aside and the respondent reads the questions and enters responses. One method almost always shows a higher percent of subjects admitting use of illegal drugs. Which method? Explain why.

### 3.134 Your institutional review board.

Your college or university has an institutional review board that screens all studies that use human subjects. Get a copy of the document that describes this board (you can probably find it online).

(a) According to this document, what are the duties of the board?

(b) How are members of the board chosen? How many members are not scientists? How many members are not employees of the college? Do these members have some special expertise, or are they simply members of the “general public”?

### 3.135 Use of data produced by the government.

Data produced by the government are often available free or at low cost to private users. For example, satellite weather data produced by the U.S. National Weather Service are available free to TV stations for their weather reports and to anyone on the Internet. *Opinion 1:* Government data should be available to everyone at minimal cost. *Opinion 2:* The satellites are expensive, and the TV stations are making a profit from their weather services, so they should share the cost. European governments, for example, charge TV stations for weather data. Which opinion do you support, and why?

### 3.136 Should we ask for the consent of the parents?

The Centers for Disease Control and Prevention, in a survey of teenagers, asked the subjects if they were sexually active. Those who said “Yes” were then asked, “How old were you when you had sexual intercourse for the first time?” Should consent of parents be required to ask minors about sex, drugs, and other such issues, or is consent of the minors themselves enough? Give reasons for your opinion.

### 3.137 A theft experiment.

Students sign up to be subjects in a psychology experiment. When they arrive, they are told that interviews are running late and are taken to a waiting room. The experimenters then stage a theft of a valuable object left in the waiting room. Some subjects are alone with the thief, and others are in pairs—these are the treatments being compared. Will the subject report the theft? The students had agreed to take part in an unspecified study, and the true nature of the experiment is explained to them afterward. Do you think this study is ethically OK?

### 3.138 A cheating experiment.

A psychologist conducts the following experiment. She measures the attitude of subjects toward cheating and then has them play a game rigged so that winning without cheating is impossible. The computer that organizes the game also records—unknown to the subjects—whether or not they

cheat. Then attitude toward cheating is retested. Subjects who cheat tend to change their attitudes to find cheating more acceptable. Those who resist the temptation to cheat tend to condemn cheating more strongly on the second test of attitude. These results confirm the psychologist's theory. This experiment tempts subjects to cheat. The subjects are led to believe that they can cheat secretly when in fact they are observed. Is this experiment ethically objectionable? Explain your position.

# **4 Probability: The Study of Randomness**

CHAPTER



- 4.1 Randomness
- 4.2 Probability Models
- 4.3 Random Variables
- 4.4 Means and Variances of Random Variables
- 4.5 General Probability Rules

## **Introduction**

The reasoning of statistical inference rests on asking, “How often would this method give a correct answer if I used it very many times?” When we produce data by random sampling or randomized comparative experiments, the laws of probability answer the question “What would happen if we did this many times?” Games of chance like Texas hold ’em are exciting because the outcomes are determined by the rules of probability.

## 4.1 Randomness

When you complete this section, you will be able to

- Identify random phenomena.
- Interpret the term “probability” for particular examples.
- Identify trials as independent or not.

Toss a coin, or choose an SRS. The result can't be predicted in advance, because the result will vary when you toss the coin or choose the sample repeatedly. But there is nonetheless a regular pattern in the results, a pattern that emerges clearly only after many repetitions. This remarkable fact is the basis for the idea of probability.



sampling distributions, p. 209

### Example

#### 4.1 Toss a coin 5000 times.

When you toss a coin, there are only two possible outcomes, heads or tails. Figure 4.1 shows the results of tossing a coin 5000 times twice. For each number of tosses from 1 to 5000, we have plotted the proportion of those tosses that gave a head. Trial A (red line) begins tail, head, tail, tail.

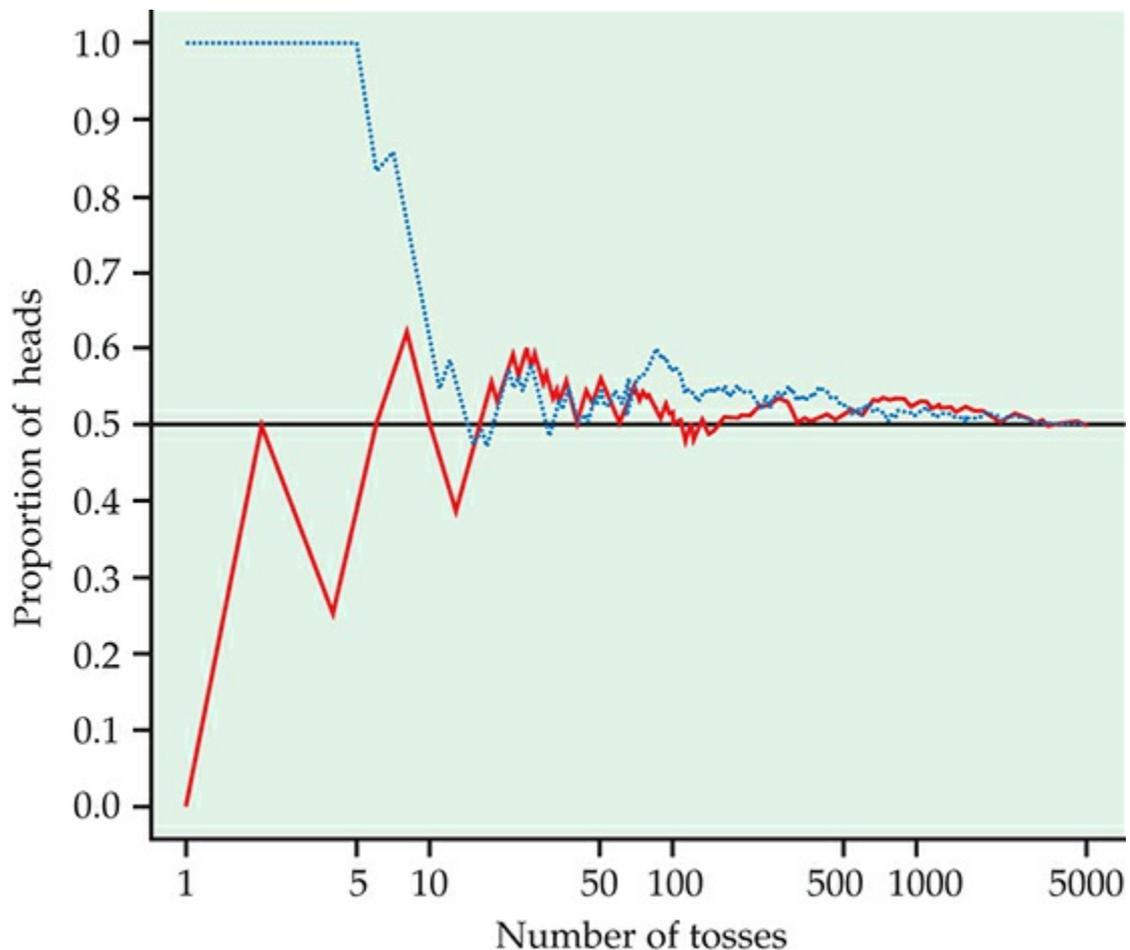
You can see that the proportion of heads for Trial A starts at 0 on the first toss, rises to 0.5 when the second toss gives a head, then falls to 0.33 and 0.25 as we get two more tails. Trial B (blue dotted line), on the other hand, starts with five straight heads, so the proportion of heads is 1 until the sixth toss.

The proportion of tosses that produce heads is quite variable at first. Trial A starts low and Trial B starts high. As we make more and more tosses, however, the proportions of heads for both trials get close to 0.5 and stay there.

If we made yet a third trial at tossing the coin a great many times, the proportion of heads would again settle down to 0.5 in the long run. We say that 0.5 is the **probability** of a head. The probability 0.5 appears as a horizontal line on the graph.



The *Probability* applet on the text website animates Figure 4.1. It allows you to choose the probability of a head and simulate any number of tosses of a coin with that probability. Try it. You will see that the proportion of heads gradually settles down close to the chosen probability. Equally important, you will also see that the proportion in a small or moderate number of tosses can be far from the probability. *Probability describes only what happens in the long run. Most people expect chance outcomes to show more short-term regularity than is actually true.*



**FIGURE 4.1**

The proportion of tosses of a coin that give a head varies as we make more tosses. Eventually, however, the proportion approaches 0.5, the probability of a head. This figure shows the results of two trials of 5000 tosses each, for Example 4.1.

## Example

### 4.2 Significance testing and Type I errors.

In Chapter 6 we will learn about significance testing and Type I errors. When we perform a significance test, we have the possibility of making a Type I error under certain circumstances. The significance-testing procedure is set up so that the probability of making this kind of error is small, usually 5%. If we perform a large number of significance tests under this set of circumstances, the proportion of times that we will make a Type I error is 0.05.

In the coin toss setting, the probability of a head is a characteristic of the coin being tossed. A coin is called **fair** if the probability of a head is 0.5; that is, it is equally likely to come up heads or tails. If we toss a coin five times and it comes up heads for all five tosses, we suspect that the coin is not fair. Is this outcome likely if, in fact, the coin is fair? We will learn a lot more about significance testing in later chapters. For now, we are content with some very general ideas.

#### fair coin

When the Type I error of a statistical significance procedure is set at 0.05, this probability is a characteristic of the procedure. If we roll a pair of dice once, we do not know whether the sum of the faces will be seven or not. Similarly, if we perform a significance test once, we do not know if we will make a Type I error or not. However, if the procedure is designed to have a Type I error probability of 0.05, then we are much less likely than not to make a Type I error.

## The language of probability

“Random” in statistics is not a synonym for “haphazard” but a description of a kind of order that emerges in the long run. We often encounter the unpredictable side of randomness in our everyday experience, but we rarely see enough repetitions of the same random phenomenon to observe the long-term regularity that probability describes. You can see that regularity emerging in Figure 4.1. In the very long run, the proportion of tosses that give a head is 0.5. This is the intuitive idea of probability. Probability 0.5 means “occurs half the time in a very large number of trials.”

## RANDOMNESS AND PROBABILITY

We call a phenomenon **random** if individual outcomes are uncertain but there is nonetheless a regular distribution of outcomes in a large number of repetitions.

The **probability** of any outcome of a random phenomenon is the proportion of times the outcome would occur in a very long series of repetitions.

Not all coins are fair. In fact, most real coins have bumps and imperfections that make the probability of heads a little different from 0.5. The probability might be 0.499999 or 0.500002. For our study of probability in this chapter, we will assume that we know the actual values of probabilities. Thus, we assume things like fair coins, even though we know that real coins are not exactly fair. We do this to learn what kinds of outcomes we are likely to see when we make such assumptions. When we study statistical inference in later chapters, we look at the situation from the opposite point of view: given that we have observed certain outcomes, what can we say about the probabilities that generated these outcomes?

## USE YOUR KNOWLEDGE

### 4.1 Use Table B.

We can use the random digits in Table B in the back of the book to simulate tossing a fair coin. Start at line 121 and read the numbers from left to right. If the number is 0, 2, 4, 6, or 8, you will say that the coin toss resulted in a head; if the number is a 1, 3, 5, 7, or 9, the outcome is tails. Use the first 20 random digits on line 121 to simulate 20 tosses of a fair coin. What is the actual proportion of heads in your simulated sample? Explain why you did not get exactly 10 heads.

Probability describes what happens in very many trials, and we must actually observe many trials to pin down a probability. In the case of tossing a coin, some diligent people have in fact made thousands of tosses.

## Example

### 4.3 Many tosses of a coin.

The French naturalist Count Buffon (1707–1788) tossed a coin 4040 times. Result: 2048 heads, or proportion  $2048/4040 = 0.5069$  for heads.

Around 1900, the English statistician Karl Pearson heroically tossed a coin 24,000 times. Result: 12,012 heads, a proportion of 0.5005.

While imprisoned by the Germans during World War II, the South African statistician John Kerrich tossed a coin 10,000 times. Result: 5067 heads, proportion of heads 0.5067.

## Thinking about randomness

That some things are random is an observed fact about the world. The outcome of a coin toss, the time between emissions of particles by a radioactive source, and the sexes of the next litter of lab rats are all random. So is the outcome of a random sample or a randomized experiment. Probability theory is the branch of mathematics that describes random behavior. Of course, we can never observe a probability exactly. We could always continue tossing the coin, for example. Mathematical probability is an idealization based on imagining what would happen in an indefinitely long series of trials.

The best way to understand randomness is to observe random behavior—not only the long-run regularity but the unpredictable results of short runs. You can do this with physical devices such as coins and dice, but software simulations of random behavior allow faster exploration. As you explore randomness, remember:

- You must have a long series of **independent** trials. That is, the outcome of one trial must not influence the outcome of any other. Imagine a crooked gambling house where the operator of a roulette wheel can stop it where she chooses—she can prevent the proportion of “red” from settling down to a fixed number. These trials are not independent.

### independence

- The idea of probability is empirical. Simulations start with given probabilities and imitate random behavior, but we can estimate a real-world probability only by actually observing many trials.
- Nonetheless, simulations are very useful because we need long runs of trials. In situations such as coin tossing, the proportion of an outcome often requires several hundred trials to settle down to the probability of that outcome. The kinds of physical random devices suggested in the exercises are too slow to make performing so many trials practical. Short runs give only rough estimates of a probability.

## The uses of probability

Probability theory originated in the study of games of chance. Tossing dice, dealing shuffled cards, and spinning a roulette wheel are examples of deliberate randomization. In that respect, they are similar to random sampling. Although games of chance are ancient, they were not studied by mathematicians until the sixteenth and seventeenth centuries.

It is only a mild simplification to say that probability as a branch of mathematics arose when seventeenth-century French gamblers asked the mathematicians Blaise Pascal and Pierre de Fermat for help. Gambling is still with us, in casinos and state lotteries. We will make use of games of chance as simple examples that illustrate the principles of probability.

Careful measurements in astronomy and surveying led to further advances in probability in the eighteenth and nineteenth centuries because the results of repeated measurements are random and can be described by distributions much like those arising from random sampling. Similar distributions appear in data on human life span (mortality tables) and in data on lengths or weights in a population of skulls, leaves, or cockroaches.<sup>1</sup>

Now, we employ the mathematics of probability to describe the flow of traffic through a highway system, the Internet, or a computer processor; the genetic makeup of individuals or populations; the energy states of subatomic particles; the spread of epidemics or tweets; and the rate of return on risky investments. Although we are interested in probability because of its usefulness in statistics, the mathematics of chance is important in many fields of study.

### SECTION 4.1 Summary

A **random phenomenon** has outcomes that we cannot predict but that nonetheless have a regular distribution in very many repetitions.

The **probability** of an event is the proportion of times the event occurs in many repeated trials of a random phenomenon.

Trials are **independent** if the outcome of one trial does not influence the outcome of any other trial.

### SECTION 4.1 Exercises

*For Exercise 4.1, see page 234.*

## 4.2 Are these phenomena random?

Identify each of the following phenomena as random or not. Give reasons for your answers.

- The outside temperature in your town at noon on Groundhog Day, February 2.
- The first digit in your student identification number.

(c) You draw an ace from a well-shuffled deck of 52 cards.

### 4.3 Interpret the probabilities.

Refer to the previous exercise. In each case, interpret the term “probability” for the phenomena that are random. For those that are not random, explain why the term “probability” does not apply.

### 4.4 Are the trials independent?

For each of the following situations, identify the trials as independent or not. Explain your answers.

- (a) You record the outside temperature in your town at noon on Groundhog Day, February 2, each year for the next 5 years.
- (b) The number of tweets that you receive on the next 10 Mondays.
- (c) Your grades in the five courses that you are taking this semester.

### 4.5 Winning at craps.

The game of craps starts with a “come-out” roll, in which the shooter rolls a pair of dice. If the total of the “spots” on the up-faces is 7 or 11, the shooter wins immediately (there are ways that the shooter can win on later rolls if other numbers are rolled on the come-out roll). Roll a pair of dice 25 times and estimate the probability that the shooter wins immediately on the come-out roll. For a pair of perfectly made dice, the probability is 0.2222.

### 4.6 Is music playing on the radio?

Turn on your favorite music radio station 8 times at least 10 minutes apart. Each time record whether or not music is playing. Calculate the number of times music is playing divided by 8. This number is an estimate of the probability that music is playing when you turn on this station. It is also an estimate of the proportion of time that music is playing on this station.

### 4.7 Wait 5 seconds between each observation.

Refer to the previous exercise. Explain why you would not want to wait only 5 seconds between each time you turn the radio station on.



### 4.8 Use the *Probability* applet.

The idea of probability is that the *proportion* of heads in many tosses of a balanced coin eventually gets close to 0.5. But does the actual *count* of heads get close to one-half the number of tosses? Let’s find out. Set the “Probability of Heads” in the *Probability* applet to 0.5 and the number of tosses to 50. You can extend the number of tosses by clicking “Toss” again to get 50 more. Don’t click “Reset” during this exercise.

- (a) After 50 tosses, what is the proportion of heads? What is the count of heads? What is the difference between the count of heads and 25 (one-half the number of tosses)?
- (b) Keep going to 150 tosses. Again record the proportion and count of heads and the difference between

the count and 75 (half the number of tosses).

(c) Keep going. Stop at 300 tosses and again at 600 tosses to record the same facts. Although it may take a long time, the laws of probability say that the proportion of heads will always get close to 0.5 and also that the difference between the count of heads and half the number of tosses will always grow without limit.



#### **4.9 A question about dice.**

Here is a question that a French gambler asked the mathematicians Fermat and Pascal at the very beginning of probability theory: what is the probability of getting at least one 6 in rolling four dice? The *Law of Large Numbers* applet allows you to roll several dice and watch the outcomes. (Ignore the title of the applet for now.) Because simulation—just like real random phenomena—often takes very many trials to estimate a probability accurately, let's simplify the question: is this probability clearly greater than 0.5, clearly less than 0.5, or quite close to 0.5? Use the applet to roll four dice until you can confidently answer this question. You will have to set "Rolls" to 1 so that you have time to look at the four up-faces. Keep clicking "Roll dice" to roll again and again. How many times did you roll four dice? What percent of your rolls produced at least one 6?

## 4.2 Probability Models

When you complete this section, you will be able to

- Describe a sample space from a description of a random phenomenon.
- Apply the five probability rules.
- Identify random phenomena that have equally likely outcomes and distinguish them from those that do not.

The idea of probability as a proportion of outcomes in very many repeated trials guides our intuition but is hard to express in mathematical form. A description of a random phenomenon in the language of mathematics is called a **probability model**. To see how to proceed, think first about a very simple random phenomenon, tossing a coin once. When we toss a coin, we cannot know the outcome in advance. What do we know? We are willing to say that the outcome will be either heads or tails. Because the coin appears to be balanced, we believe that each of these outcomes has probability  $1/2$ . This description of coin tossing has two parts:

### probability model

- A list of possible outcomes
- A probability for each outcome

This two-part description is the starting point for a probability model. We will begin by describing the outcomes of a random phenomenon and then learn how to assign probabilities to the outcomes.

### Sample spaces

A probability model first tells us what outcomes are possible.

#### SAMPLE SPACE

The **sample space**  $S$  of a random phenomenon is the set of all possible outcomes.

The name “sample space” is natural in random sampling, where each possible outcome is a sample and the sample space contains all possible samples. To specify

$S$ , we must state what constitutes an individual outcome and then state which outcomes can occur. We often have some freedom in defining the sample space, so the choice of  $S$  is a matter of convenience as well as correctness. The idea of a sample space, and the freedom we may have in specifying it, are best illustrated by examples.

### **Example**

#### **4.4 Sample space for tossing a coin.**

Toss a coin. There are only two possible outcomes, and the sample space is

$$S = \{\text{heads, tails}\}$$

or, more briefly,  $S = \{H, T\}$ .

### **Example**

#### **4.5 Sample space for random digits.**

Let your pencil point fall blindly into Table B of random digits. Record the value of the digit it lands on. The possible outcomes are

$$S = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

### **Example**

#### **4.6 Sample space for tossing a coin four times.**

Toss a coin four times and record the results. That's a bit vague. To be exact, record the results of each of the four tosses in order. A typical outcome is then

HTTH. Counting shows that there are 16 possible outcomes. The sample space  $S$  is the set of all 16 strings of four H's and T's.

Suppose that our only interest is the number of heads in four tosses. Now we can be exact in a simpler fashion. The random phenomenon is to toss a coin four times and count the number of heads. The sample space contains only five outcomes:

$$S = \{0, 1, 2, 3, 4\}$$

This example illustrates the importance of carefully specifying what constitutes an individual outcome.

Although these examples seem remote from the practice of statistics, the connection is surprisingly close. Suppose that in conducting an opinion poll you select four people at random from a large population and ask each if he or she favors reducing federal spending on low-interest student loans. The answers are “Yes” or “No.” The possible outcomes—the sample space—are exactly as in Example 4.4 if we replace heads by “Yes” and tails by “No.” Similarly, the possible outcomes of an SRS of 1500 people are the same in principle as the possible outcomes of tossing a coin 1500 times. One of the great advantages of mathematics is that the essential features of quite different phenomena can be described by the same mathematical model.

## USE YOUR KNOWLEDGE

### 4.10 What color are your eyes?

A student is asked what color eyes he or she has. Set up an appropriate sample space for this setting. Note that there is not a single correct answer to this exercise, so give reasons for your choice.

The sample spaces described above correspond to categorical variables where we can list all the possible values. Other sample spaces correspond to quantitative variables. Here is an example.

## Example

## 4.7 Using software.

Most statistical software has a function that will generate a random number between 0 and 1. The sample space is

$$S = \{\text{all numbers between 0 and 1}\}$$

This  $S$  is a mathematical idealization. Any specific random number generator produces numbers with some limited number of decimal places so that, strictly speaking, not all numbers between 0 and 1 are possible outcomes. For example, Minitab generates random numbers like 0.736891, with six decimal places. The entire interval from 0 to 1 is easier to think about. It also has the advantage of being a suitable sample space for different software systems that produce random numbers with different numbers of digits.

## USE YOUR KNOWLEDGE

### 4.11 How many hours do you text?



You record the number of hours per week that a randomly selected student spends texting. What is the sample space?

A sample space  $S$  lists the possible outcomes of a random phenomenon. To complete a mathematical description of the random phenomenon, we must also give the probabilities with which these outcomes occur.

The true long-term proportion of any outcome—say, “exactly 2 heads in four tosses of a coin”—can be found only empirically, and then only approximately. How then can we describe probability mathematically? Rather than immediately attempting to give “correct” probabilities, let’s confront the easier task of laying down rules that any assignment of probabilities must satisfy. We need to assign

probabilities not only to single outcomes but also to sets of outcomes.

## EVENT

An **event** is an outcome or a set of outcomes of a random phenomenon. That is, an event is a subset of the sample space.

### Example

#### 4.8 Exactly 2 heads in four tosses.

Take the sample space  $S$  for four tosses of a coin to be the 16 possible outcomes in the form HTHH. Then “exactly 2 heads” is an event. Call this event  $A$ . The event  $A$  expressed as a set of outcomes is

$$A = \{TTHH, THTH, THHT, HTTH, HTHT, HHTT\}$$

In a probability model, events have probabilities. What properties must any assignment of probabilities to events have? Here are some basic facts about any probability model. These facts follow from the idea of probability as “the long-run proportion of repetitions on which an event occurs.”

1. **Any probability is a number between 0 and 1.** Any proportion is a number between 0 and 1, so any probability is also a number between 0 and 1. An event with probability 0 never occurs, and an event with probability 1 occurs on every trial. An event with probability 0.5 occurs in half the trials in the long run.
2. **All possible outcomes together must have probability 1.** Because every trial will produce an outcome, the sum of the probabilities for all possible outcomes must be exactly 1.
3. **If two events have no outcomes in common, the probability that one or the other occurs is the sum of their individual probabilities.** If one event occurs in 40% of all trials, a different event occurs in 25% of all trials, and the two can never occur together, then one or the other occurs on 65% of all trials because  $40\% + 25\% = 65\%$ .
4. **The probability that an event does not occur is 1 minus the probability that the event does occur.** If an event occurs in (say) 70% of all trials, it fails to occur in the other 30%. The probability that an event occurs and the probability

that it does not occur always add to 100%, or 1.

## Probability rules

Formal probability uses mathematical notation to state Facts 1 to 4 more concisely. We use capital letters near the beginning of the alphabet to denote events. If  $A$  is any event, we write its probability as  $P(A)$ . Here are our probability facts in formal language. As you apply these rules, remember that they are just another form of intuitively true facts about long-run proportions.

### PROBABILITY RULES

**Rule 1.** The probability  $P(A)$  of any event  $A$  satisfies  $0 \leq P(A) \leq 1$ .

**Rule 2.** If  $S$  is the sample space in a probability model, then  $P(S) = 1$ .

**Rule 3.** Two events  $A$  and  $B$  are **disjoint** if they have no outcomes in common and so can never occur together. If  $A$  and  $B$  are disjoint,

$$P(A \text{ or } B) = P(A) + P(B)$$

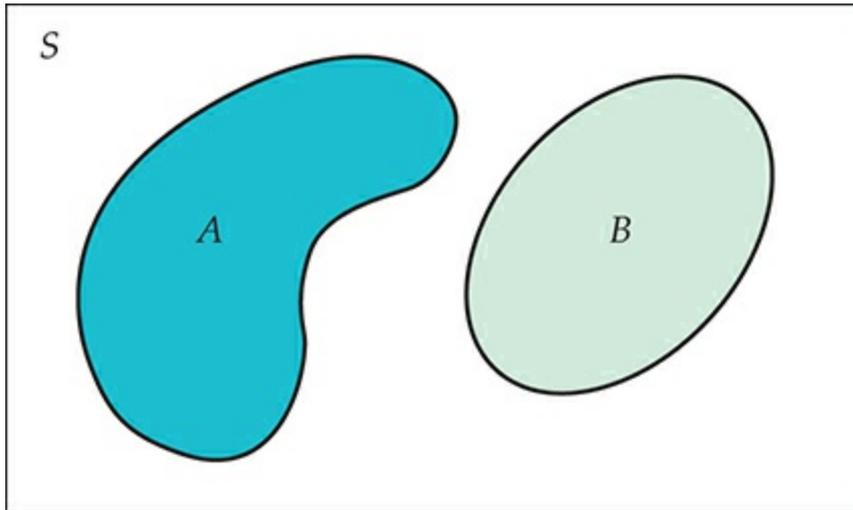
This is the **addition rule for disjoint events**.

**Rule 4.** The **complement** of any event  $A$  is the event that  $A$  does not occur, written as  $A^c$ . The **complement rule** states that

$$P(A^c) = 1 - P(A)$$

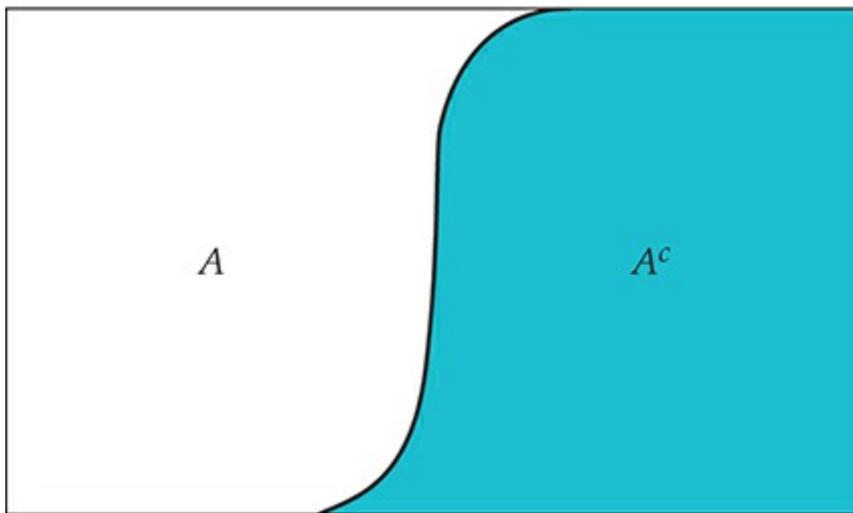
You may find it helpful to draw a picture to remind yourself of the meaning of complements and disjoint events. A picture like Figure 4.2 that shows the sample space  $S$  as a rectangular area and events as areas within  $S$  is called a **Venn diagram**. The events  $A$  and  $B$  in Figure 4.2 are disjoint because they do not overlap. As Figure 4.3 shows, the complement  $A^c$  contains exactly the outcomes that are not in  $A$ .

### Venn diagram



**FIGURE 4.2**

Venn diagram showing disjoint events  $A$  and  $B$ . Disjoint events have no common outcomes.



**FIGURE 4.3**

Venn diagram showing the complement  $A^c$  of an event  $A$ . The complement consists of all outcomes that are not in  $A$ .

## Example

### 4.9 Favorite vehicle colors.



What is your favorite color for a vehicle? Our preferences can be related to our personality, our moods, or particular objects. Here is a probability model for color preferences.<sup>2</sup>

Color	White	Black	Silver	Gray
Probability	0.23	0.18	0.16	0.13
Color	Red	Blue	Brown	Other
Probability	0.10	0.09	0.05	0.06

Each probability is between 0 and 1. The probabilities add to 1 because these outcomes together make up the sample space  $S$ . Our probability model corresponds to selecting a person at random and asking what is their favorite color.

Let's use the probability Rules 3 and 4 to find some probabilities for favorite vehicle colors.

## Example

### 4.10 Black or silver?

What is the probability that a person's favorite vehicle color is black or silver? If the favorite is black, it cannot be silver, so these two events are disjoint. Using Rule 3, we find

$$\begin{aligned}
 P(\text{black or silver}) &= P(\text{black}) + P(\text{silver}) \\
 &= 0.18 + 0.16 = 0.34
 \end{aligned}$$

There is a 34% chance that a randomly selected person will choose black or silver as their favorite color. Suppose that we want to find the probability that the

favorite color is not blue.

## Example

### 4.11 Use the complement rule.

To solve this problem, we could use Rule 3 and add the probabilities for white, black, silver, gray, red, brown and other. However, it is easier to use the probability that we have for blue and Rule 4. The event that the favorite is not blue is the complement of the event that the favorite is blue. Using our notation for events, we have

$$\begin{aligned}P(\text{not blue}) &= 1 - P(\text{blue}) \\ &= 1 - 0.09 = 0.91\end{aligned}$$

We see that 91% of people have a favorite vehicle color that is not blue.

## USE YOUR KNOWLEDGE

### 4.12 Red or brown.

Find the probability that the favorite color is red or brown.

### 4.13 White, black, silver, gray, or red.

Find the probability that the favorite color is white, black, silver, gray, or red using Rule 4. Explain why this calculation is easier than finding the answer using Rule 3.

## Assigning probabilities: finite number of outcomes

The individual outcomes of a random phenomenon are always disjoint. So the addition rule provides a way to assign probabilities to events with more than one outcome: start with probabilities for individual outcomes and add to get

probabilities for events. This idea works well when there are only a finite (fixed and limited) number of outcomes.

## PROBABILITIES IN A FINITE SAMPLE SPACE

Assign a probability to each individual outcome. These probabilities must be numbers between 0 and 1 and must have sum 1.

The probability of any event is the sum of the probabilities of the outcomes making up the event.

### Example

#### 4.12 Benford's law.

Faked numbers in tax returns, payment records, invoices, expense account claims, and many other settings often display patterns that aren't present in legitimate records. Some patterns, such as too many round numbers, are obvious and easily avoided by a clever crook. Others are more subtle. It is a striking fact that the first digits of numbers in legitimate records often follow a distribution known as **Benford's law**. Here it is (note that a first digit can't be 0):<sup>3</sup>

#### Benford's law

First digit	1	2	3	4	5	6	7	8	9
Probability	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Benford's law usually applies to the first digits of the sizes of similar quantities, such as invoices, expense account claims, and county populations. Investigators can detect fraud by comparing the first digits in records, such as invoices paid by a business, with these probabilities.

### Example

### 4.13 Find some probabilities for Benford's law.

Consider the events

$$A = \{\text{first digit is 5}\}$$

$$B = \{\text{first digit is 3 or less}\}$$

From the table of probabilities in Example 4.12,

$$P(A) = P(5) = 0.079$$

$$\begin{aligned} P(B) &= P(1) + P(2) + P(3) \\ &= 0.301 + 0.176 + 0.125 = 0.602 \end{aligned}$$

Note that  $P(B)$  is not the same as the probability that a first digit is strictly less than 3. The probability  $P(3)$  that a first digit is 3 is included in “3 or less” but not in “less than 3.”

## USE YOUR KNOWLEDGE

### 4.14 Benford's law.

Using the probabilities for Benford's law, find the probability that a first digit is anything other than 4.

### 4.15 Use the addition rule.

Use the addition rule with the probabilities for the events  $A$  and  $B$  from Example 4.13 to find the probability that a first digit is either 5 or 3 or less.

Be careful to apply the addition rule only to disjoint events.

## Example

#### 4.14 Find more probabilities for Benford's law.

Check that the probability of the event  $C$  that a first digit is even is

$$P(C) = P(2) + P(4) + P(6) + P(8) = 0.391$$

The probability

$$P(B \text{ or } C) = P(1) + P(2) + P(3) + P(4) + P(6) + P(8) = 0.817$$

is *not* the sum of  $P(B)$  and  $P(C)$ , because events  $B$  and  $C$  are not disjoint. Outcomes 2 is common to both events.

#### Assigning probabilities: equally likely outcomes

Assigning correct probabilities to individual outcomes often requires long observation of the random phenomenon. In some circumstances, however, we are willing to assume that individual outcomes are equally likely because of some balance in the phenomenon. Ordinary coins have a physical balance that should make heads and tails equally likely, for example, and the table of random digits comes from a deliberate randomization.

#### Example

#### 4.15 First digits that are equally likely.

You might think that first digits are distributed “at random” among the digits 1 to 9 in business records. The 9 possible outcomes would then be equally likely. The sample space for a single digit is

$$S = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

Because the total probability must be 1, the probability of each of the 9 outcomes must be  $1/9$ . That is, the assignment of probabilities to outcomes is

First digit	1	2	3	4	5	6	7	8	9
Probability	$1/9$	$1/9$	$1/9$	$1/9$	$1/9$	$1/9$	$1/9$	$1/9$	$1/9$

The probability of the event  $B$  that a randomly chosen first digit is 3 or less is

$$\begin{aligned} P(B) &= P(1) + P(2) + P(3) \\ &= \frac{1}{9} + \frac{1}{9} + \frac{1}{9} = \frac{3}{9} = 0.333 \end{aligned}$$

Compare this with the Benford's law probability in Example 4.13. A crook who fakes data by using "random" digits will end up with too few first digits that are 3 or less.

In Example 4.15 all outcomes have the same probability. Because there are 9 equally likely outcomes, each must have probability  $1/9$ . Because exactly 3 of the 9 equally likely outcomes are 3 or less, the probability of this event is  $3/9$ . In the special situation where all outcomes are equally likely, we have a simple rule for assigning probabilities to events.

## EQUALLY LIKELY OUTCOMES

If a random phenomenon has  $k$  possible outcomes, all equally likely, then each individual outcome has probability  $1/k$ . The probability of any event  $A$  is

$$P(A) = \frac{\text{count of outcomes in } A}{\text{count of outcomes in } S}$$

$$P(A) = \frac{\text{count of outcomes in } A}{k}$$

Most random phenomena do not have equally likely outcomes, so the general rule for finite sample spaces (page 282) is more important than the special rule for equally likely outcomes.

## USE YOUR KNOWLEDGE

### 4.16 Possible outcomes for rolling a die.

A die has six sides with 1 to 6 spots on the sides. Give the probability distribution for the six possible outcomes that can result when a perfect die is rolled.

## Independence and the multiplication rule

Rule 3, the addition rule for disjoint events, describes the probability that *one or the other* of two events  $A$  and  $B$  will occur in the special situation when  $A$  and  $B$  cannot occur together because they are disjoint. Our final rule describes the probability that *both* events  $A$  and  $B$  occur, again only in a special situation. More general rules appear in Section 4.5, but in our study of statistics we will need only

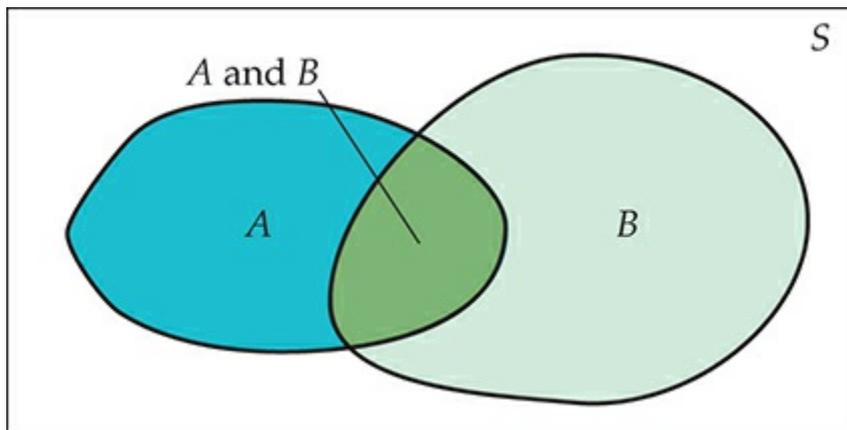
the rules that apply to special situations.

Suppose that you toss a fair coin twice. You are counting heads, so two events of interest are

$$A = \{\text{first toss is a head}\}$$

$$B = \{\text{second toss is a head}\}$$

The events  $A$  and  $B$  are not disjoint. They occur together whenever both tosses give heads. We want to compute the probability of the event  $\{A \text{ and } B\}$  that *both* tosses are heads. The Venn diagram in Figure 4.4 illustrates the event  $\{A \text{ and } B\}$  as the overlapping area that is common to both  $A$  and  $B$ .



**FIGURE 4.4**

Venn diagram showing the event  $\{A \text{ and } B\}$ . This event consists of outcomes common to  $A$  and  $B$ .

The coin tossing of Buffon, Pearson, and Kerrich described in Example 4.3 makes us willing to assign probability  $1/2$  to a head when we toss a coin. So

$$P(A) = 0.5$$

$$P(B) = 0.5$$

What is  $P(A \text{ and } B)$ ? Our common sense says that it is  $1/4$ . The first toss will give a head half the time and the second toss will give a head half the time, so both tosses will give heads on  $1/2 \times 1/2 = 1/4$  of all trials in the long run. This reasoning assumes that the second toss still has probability  $1/2$  of a head after the first has given a head. This is true—we can verify it by tossing a coin twice many times and observing the proportion of heads on the second toss after the first toss has produced a head. We say that the events “head on the first toss” and “head on the second toss” are *independent*. Here is our final probability rule.

## MULTIPLICATION RULE FOR INDEPENDENT EVENTS

**Rule 5.** Two events  $A$  and  $B$  are **independent** if knowing that one occurs does not change the probability that the other occurs. If  $A$  and  $B$  are independent,

$$P(A \text{ and } B) = P(A)P(B)$$

This is the **multiplication rule for independent events**.

Our definition of independence is rather informal. We will make this informal idea precise in Section 4.5. In practice, though, we rarely need a precise definition of independence, because independence is usually *assumed* as part of a probability model when we want to describe random phenomena that seem to be physically unrelated to each other. Here is an example of independence.

## Example

### 4.16 Coins do not have memory.

Because a coin has no memory, we assume that successive coin tosses are independent. For a fair coin, this means that the outcome of the first toss does not influence the outcome of any other toss.

## USE YOUR KNOWLEDGE

### 4.17 A head and then a tail in two tosses.

What is the probability of obtaining a head and then a tail on two tosses of a fair coin?

Here is an example of a situation where there are dependent events.

## Example

#### 4.17 Dependent events in cards.

The colors of successive cards dealt from the same deck are not independent. A standard 52-card deck contains 26 red and 26 black cards. For the first card dealt from a shuffled deck, the probability of a red card is  $26/52 = 0.50$  because the 52 possible cards are equally likely. Once we see that the first card is red, we know that there are only 25 reds among the remaining 51 cards. The probability that the second card is red is therefore only  $25/51 = 0.49$ . Knowing the outcome of the first deal changes the probabilities for the second.

### USE YOUR KNOWLEDGE

#### 4.18 The probability of a second ace.

A deck of 52 cards contains 4 aces, so the probability that a card drawn from this deck is an ace is  $4/52$ . If we know that the first card drawn is an ace, what is the probability that the second card drawn is also an ace? Using the idea of independence, explain why this probability is not  $4/52$ .

Here is another example of a situation where events are dependent.

### Example

#### 4.18 Taking a test twice.

If you take an IQ test or other mental test twice in succession, the two test scores are not independent. The learning that occurs on the first attempt influences your second attempt. If you learn a lot, then your second test score might be a lot higher than your first test score.

When independence is part of a probability model, the multiplication rule applies. Here is an example.

## Example

### 4.19 Mendel's peas.



Gregor Mendel used garden peas in some of the experiments that revealed that inheritance operates randomly. The seed color of Mendel's peas can be either green or yellow. Two parent plants are "crossed" (one pollinates the other) to produce seeds.

Each parent plant carries two genes for seed color, and each of these genes has probability  $1/2$  of being passed to a seed. The two genes that the seed receives, one from each parent, determine its color. The parents contribute their genes independently of each other.

Suppose that both parents carry the  $G$  and the  $Y$  genes. The seed will be green if both parents contribute a  $G$  gene; otherwise, it will be yellow. If  $M$  is the event that the male contributes a  $G$  gene and  $F$  is the event that the female contributes a  $G$  gene, then the probability of a green seed is

$$\begin{aligned} P(M \text{ and } F) &= P(M)P(F) \\ &= (0.5)(0.5) = 0.25 \end{aligned}$$

In the long run,  $1/4$  of all seeds produced by crossing these plants will be green.



*The multiplication rule applies only to independent events; you cannot use it if events are not independent.* Here is a distressing example of misuse of the multiplication rule.

## Example

### 4.20 Sudden infant death syndrome.

Sudden infant death syndrome (SIDS) causes babies to die suddenly (often in their cribs) with no explanation. Deaths from SIDS have been greatly reduced by placing babies on their backs, but as yet no cause is known.

When more than one SIDS death occurs in a family, the parents are sometimes accused. One “expert witness” popular with prosecutors in England told juries that there is only a 1 in 73 million chance that two children in the same family could have died from SIDS. Here’s his calculation: the rate of SIDS in a nonsmoking middle-class family is 1 in 8500. So the probability of two deaths is

$$18500 \times 18500 = 172,250,000$$

Several women were convicted of murder on this basis, without any direct evidence that they harmed their children.

As the Royal Statistical Society said, this reasoning is nonsense. It assumes that SIDS deaths in the same family are independent events. The cause of SIDS is unknown: “There may well be unknown genetic or environmental factors that predispose families to SIDS, so that a second case within the family becomes much more likely.”<sup>4</sup> The British government decided to review the cases of 258 parents convicted of murdering their babies.



The multiplication rule  $P(A \text{ and } B) = P(A) P(B)$  holds if  $A$  and  $B$  are independent but not otherwise. The addition rule  $P(A \text{ or } B) = P(A) + P(B)$  holds if  $A$  and  $B$  are disjoint but not otherwise. Resist the temptation to use these simple formulas when the circumstances that justify them are not present. *You must also be certain not to confuse disjointness and independence. Disjoint events cannot be independent.* If  $A$  and  $B$  are disjoint, then the fact that  $A$  occurs tells us that  $B$  cannot occur—look again at Figure 4.2 (page 240). Unlike disjointness or complements, independence cannot be pictured by a Venn diagram, because it involves the probabilities of the events rather than just the outcomes that make up the events.

## Applying the probability rules

If two events  $A$  and  $B$  are independent, then their complements  $A^c$  and  $B^c$  are also independent and  $A^c$  is independent of  $B$ . Suppose, for example, that 75% of all registered voters in a suburban district are Republicans. If an opinion poll interviews two voters chosen independently, the probability that the first is a Republican and the second is not a Republican is  $(0.75)(0.25) = 0.1875$ .

The multiplication rule also extends to collections of more than two events, provided that all are independent. Independence of events  $A$ ,  $B$ , and  $C$  means that no information about any one or any two can change the probability of the remaining events. The formal definition is a bit messy. Fortunately, independence is usually assumed in setting up a probability model. We can then use the multiplication rule freely.

By combining the rules we have learned, we can compute probabilities for rather complex events. Here is an example.

## Example

### 4.21 HIV testing.

Many people who come to clinics to be tested for HIV, the virus that causes AIDS, don't come back to learn the test results. Clinics now use "rapid HIV tests" that give a result in a few minutes. The false-positive rate for a diagnostic test is the probability that a person with no disease will have a positive test result. For the rapid HIV tests, the Food and Drug Administration (FDA) has established 2% as the maximum false-positive rate allowed for a rapid HIV test.<sup>5</sup> If a clinic uses a test that meets the FDA standard and tests 50 people who are free of HIV antibodies, what is the probability that at least 1 false-positive will occur?

It is reasonable to assume as part of the probability model that the test results for different individuals are independent. The probability that the test is positive for a single person is 0.02, so the probability of a negative result is  $1 - 0.02 = 0.98$  by the complement rule. The probability of at least 1 false-positive among the 50 people tested is therefore

$$\begin{aligned} P(\text{at least 1 positive}) &= 1 - P(\text{no positives}) \\ &= 1 - P(50 \text{ negatives}) \\ &= 1 - 0.98^{50} \\ &= 1 - 0.3642 = 0.6358 \end{aligned}$$

There is approximately a 64% chance that at least 1 of the 50 people will test positive for HIV even though none of them has the virus.

Concern about excessive numbers of false-positives led the New York City Department of Health and Mental Hygiene to suspend the use of one particular rapid HIV test.<sup>6</sup>

## SECTION 4.2 Summary

A **probability model** for a random phenomenon consists of a sample space  $S$  and an assignment of probabilities  $P$ .

The **sample space**  $S$  is the set of all possible outcomes of the random phenomenon. Sets of outcomes are called **events**.  $P$  assigns a number  $P(A)$  to an event  $A$  as its probability.

The **complement**  $A^c$  of an event  $A$  consists of exactly the outcomes that are not in  $A$ . Events  $A$  and  $B$  are **disjoint** if they have no outcomes in common. Events  $A$  and  $B$  are **independent** if knowing that one event occurs does not change the probability we would assign to the other event.

Any assignment of probability must obey the rules that state the basic properties of probability:

**Rule 1.**  $0 \leq P(A) \leq 1$  for any event  $A$

**Rule 2.**  $P(S) = 1$ .

**Rule 3. Addition rule:** If events  $A$  and  $B$  are **disjoint**, then  $P(A \text{ or } B) = P(A) + P(B)$ .

**Rule 4. Complement rule:** For any event  $A$ ,  $P(A^c) = 1 - P(A)$ .

**Rule 5. Multiplication rule:** If events  $A$  and  $B$  are **independent**, then  $P(A \text{ and } B) = P(A)P(B)$

## SECTION 4.2 Exercises

*For Exercise 4.10, see page 238; for Exercise 4.11, see page 239; for Exercises 4.12 and 4.13, see page 241; for Exercises 4.14 and 4.15, see page 243; for Exercise 4.16, see page 244; for Exercise 4.17, see page 246; and for Exercise 4.18, see page 246.*

### 4.19 What is the sample space?

For each of the following questions, define a sample space for the associated random phenomenon. Explain your answers. Be sure to specify units if that is appropriate.

- Will it rain tomorrow?
- How many times do you tweet in a typical day?
- What is the average age of your Facebook friends?
- What are the majors for students at your college?

## 4.20 Probability rules.

For each of the following situations, state the probability rule or rules that you would use and apply it or them. Write a sentence explaining how the situation illustrates the use of the probability rules.

- (a) The probability of event  $A$  is 0.224. What is the probability that event  $A$  does not occur?
- (b) A coin is tossed three times. The probability of zero heads is  $1/8$  and the probability of zero tails is  $1/8$ . What is the probability that all three tosses result in the same outcome?
- (c) Refer to part (b). What is the probability that there is at least one head and at least one tail?
- (d) The probability of event  $A$  is 0.5 and the probability of event  $B$  is 0.6. Events  $A$  and  $B$  are disjoint. Can this happen?
- (e) Event  $A$  is very rare. Its probability is  $-0.01$ . Can this happen?

## 4.21 Equally likely events.

For each of the following situations, explain why you think that the events are equally likely or not. Explain your answers.

- (a) The outcome of the next tennis match for Victoria Azarenka is either a win or a loss. (You might want to check the Internet for information about this tennis player.)
- (b) You draw a king or a two from a shuffled deck of 52 cards.
- (c) You are observing turns at an intersection. You classify each turn as a right turn or a left turn.
- (d) For college basketball games, you record the times that the home team wins and the times that the home team loses.

## 4.22 The multiplication rule for independent events.

The probability that a randomly selected person prefers the vehicle color white is 0.23. Can you apply the multiplication rule for independent events in the situations described in parts (a) and (b)? If your answer is Yes, apply the rule.

- (a) Two people are chosen at random from the population. What is the probability that both prefer white?
- (b) Two people who are sisters are chosen. What is the probability that both prefer white?
- (c) Write a short summary about the multiplication rule for independent events using your answers to parts (a) and (b) to illustrate the basic idea.

## 4.23 What's wrong?

In each of the following scenarios, there is something wrong. Describe what is wrong and give a reason for your answer.

- (a) If two events are disjoint, we can multiply their probabilities to determine the probability that they will both occur.
- (b) If the probability of  $A$  is 0.2 and the probability of  $B$  is 0.5, the probability of both  $A$  and  $B$  happening is 1.1.

(c) If the probability of  $A$  is 0.35, then the probability of the complement of  $A$  is  $-0.35$ .

#### 4.24 What's wrong?

In each of the following scenarios, there is something wrong. Describe what is wrong and give a reason for your answer.

(a) If the sample space consists of two outcomes, then each outcome has probability 0.5.

(b) If we select a digit at random, then the probability of selecting a 2 is 0.2.

(c) If the probability of  $A$  is 0.2, the probability of  $B$  is 0.3, and the probability of  $A$  and  $B$  is 0.5, then  $A$  and  $B$  are independent.

#### 4.25 Evaluating web page designs.

You are a web page designer and you set up a page with five different links. A user of the page can click on one of the links or he or she can leave that page. Describe the sample space for the outcome of someone visiting your web page.

#### 4.26 Record the length of time spent on the page.

Refer to the previous exercise. You also decide to measure the length of time a visitor spends on your page. Give the sample space for this measure.

#### 4.27 Ringtones.

What are the popular ringtones? The website **funtonia.com** updates its list of top ringtones frequently. Here are probabilities for the top 10 ringtones recently listed by the site:<sup>7</sup>

Ringtone	Probability	Ringtone	Probability
No Worries	0.182	Gangnam Style	0.086
Adorn	0.153	Try	0.081
Girl on Fire	0.134	Better Dig Two	0.062
The Only Way I Know	0.096	Thinkin Bout You	0.062
Wanted	0.086	Diamonds	0.058

(a) What is the probability that a randomly selected ringtone from this list is either Wanted or Gangnam Style?

(b) What is the probability that a randomly selected ringtone from this list is not Wanted and not Gangnam Style? Be sure to show how you computed your answer.

#### 4.28 More ringtones.

Refer to the previous exercise.

(a) If two ringtones are selected independently, what is the probability that both are Girl on Fire?

(b) Describe in words the complement of the event described in part (a) of this exercise. Find the

probability of this event.

#### 4.29 Distribution of blood types.

All human blood can be “ABO-typed” as one of O, A, B, or AB, but the distribution of the types varies a bit among groups of people. Here is the distribution of blood types for a randomly chosen person in the United States:<sup>8</sup>

Blood type	A	B	AB	O
U.S. probability	0.42	0.11	?	0.44

(a) What is the probability of type AB blood in the United States?

(b) Maria has type B blood. She can safely receive blood transfusions from people with blood types O and B. What is the probability that a randomly chosen person from the United States can donate blood to Maria?

#### 4.30 Blood types in Ireland.

The distribution of blood types in Ireland differs from the U.S. distribution given in the previous exercise:

Blood type	A	B	AB	O
Ireland probability	0.35	0.10	0.03	0.52

Choose a person from the United States and a person from Ireland at random, independently of each other. What is the probability that both have type O blood? What is the probability that both have the same blood type?

#### 4.31 Are the probabilities legitimate?

In each of the following situations, state whether or not the given assignment of probabilities to individual outcomes is legitimate, that is, satisfies the rules of probability. If not, give specific reasons for your answer.

(a) Choose a college student at random and record gender and enrollment status:  $P(\text{female full-time}) = 0.44$ ,  $P(\text{female part-time}) = 0.56$ ,  $P(\text{male full-time}) = 0.46$ ,  $P(\text{male part-time}) = 0.54$ .

(b) Deal a card from a shuffled deck:  $P(\text{clubs}) = 16/52$ ,  $P(\text{diamonds}) = 12/52$ ,  $P(\text{hearts}) = 12/52$ ,  $P(\text{spades}) = 12/52$ .

(c) Roll a die and record the count of spots on the up-face:  $P(1) = 1/3$ ,  $P(2) = 0$ ,  $P(3) = 1/6$ ,  $P(4) = 1/3$ ,  $P(5) = 1/6$ ,  $P(6) = 0$ ,

#### 4.32 French and English in Canada.

Canada has two official languages, English and French. Choose a Canadian at random and ask, “What is your mother tongue?” Here is the distribution of responses, combining many separate languages from the broad Asian/Pacific region:<sup>9</sup>

Language	English	French	Asian/Pacific	Other
Probability	0.59	?	0.07	0.11

- (a) What probability should replace “?” in the distribution?
- (b) What is the probability that a Canadian’s mother tongue is not English? Explain how you computed your answer.

### 4.33 Education levels of young adults.

Choose a young adult (age 25 to 34 years) at random. The probability is 0.12 that the person chosen did not complete high school, 0.31 that the person has a high school diploma but no further education, and 0.29 that the person has at least a bachelor’s degree.

- (a) What must be the probability that a randomly chosen young adult has some education beyond high school but does not have a bachelor’s degree?
- (b) What is the probability that a randomly chosen young adult has at least a high school education?



### 4.34 Loaded dice.

There are many ways to produce crooked dice. To *load* a die so that 6 comes up too often and 1 (which is opposite 6) comes up too seldom, add a bit of lead to the filling of the spot on the 1 face. Because the spot is solid plastic, this works even with transparent dice. If a die is loaded so that 6 comes up with probability 0.21 and the probabilities of the 2, 3, 4, and 5 faces are not affected, what is the assignment of probabilities to the six faces?

### 4.35 Rh blood types.

Human blood is typed as O, A, B, or AB and also as Rh-positive or Rh-negative. ABO type and Rh-factor type are independent because they are governed by different genes. In the American population, 84% of people are Rh-positive. Use the information about ABO type in Exercise 4.29 to give the probability distribution of blood type (ABO and Rh) for a randomly chosen American.

### 4.36 Roulette.

A roulette wheel has 38 slots, numbered 0, 00, and 1 to 36. The slots 0 and 00 are colored green, 18 of the others are red, and 18 are black. The dealer spins the wheel and at the same time rolls a small ball along the wheel in the opposite direction. The wheel is carefully balanced so that the ball is equally likely to land in any slot when the wheel slows. Gamblers can bet on various combinations of numbers and colors.

- (a) What is the probability that the ball will land in any one slot?
- (b) If you bet on “red,” you win if the ball lands in a red slot. What is the probability of winning?
- (c) The slot numbers are laid out on a board on which gamblers place their bets. One column of numbers on the board contains all multiples of 3, that is, 3, 6, 9, . . . , 36. You place a “column bet” that wins if any of these numbers comes up. What is your probability of winning?

### 4.37 Winning the lottery.

A state lottery’s Pick 3 game asks players to choose a three-digit number, 000 to 999. The state chooses the winning three-digit number at random, so that each number has probability  $1/1000$ . You win if the winning number contains the digits in your number, in any order.

- (a) Your number is 491. What is your probability of winning?
- (b) Your number is 222. What is your probability of winning?

### 4.38 PINs.

The personal identification numbers (PINs) for automatic teller machines usually consist of four digits. You notice that most of your PINs have at least one 0, and you wonder if the issuers use lots of 0s to make the numbers easy to remember. Suppose that PINs are assigned at random, so that all four-digit numbers are equally likely.

- (a) How many possible PINs are there?
- (b) What is the probability that a PIN assigned at random has at least one 0?

### 4.39 Universal blood donors.

People with type O-negative blood are universal donors. That is, any patient can receive a transfusion of O-negative blood. Only 7% of the American population have O-negative blood. If 10 people appear at random to give blood, what is the probability that at least 1 of them is a universal donor?



### 4.40 Axioms of probability.

Show that any assignment of probabilities to events that obeys Rules 2 and 3 on page 239 automatically obeys the complement rule (Rule 4). This implies that a mathematical treatment of probability can start from just Rules 1, 2, and 3. These rules are sometimes called *axioms* of probability.



### 4.41 Independence of complements.

Show that if events  $A$  and  $B$  obey the multiplication rule,  $P(A \text{ and } B) = P(A)P(B)$ , then  $A$  and the complement  $B^c$  of  $B$  also obey the multiplication rule,  $P(A \text{ and } B^c) = P(A)P(B^c)$ . That is, if events  $A$  and  $B$  are independent, then  $A$  and  $B^c$  are also independent. (*Hint*: Start by drawing a Venn diagram and noticing that the events “ $A$  and  $B$ ” and “ $A$  and  $B^c$ ” are disjoint.)

### *Mendelian inheritance.*

*Some traits of plants and animals depend on inheritance of a single gene. This is called Mendelian inheritance, after Gregor Mendel (1822–1884). Exercises 4.42 to 4.45 are based on the following information about Mendelian inheritance of blood type.*

*Each of us has an ABO blood type, which describes whether two characteristics called  $A$  and  $B$  are present. Every human being has two blood type alleles (gene forms), one inherited from our mother and one from our father. Each of these alleles can be  $A$ ,  $B$ , or  $O$ . Which two we inherit determines our blood type. Here is a table that shows what our blood type is for each combination of two alleles:*

<b>Alleles inherited</b>	<b>Blood type</b>
A and A	A
A and B	AB
A and O	A
B and B	B
B and O	B
O and O	O

*We inherit each of a parent's two alleles with probability 0.5. We inherit independently from our mother and father.*

#### **4.42 Blood types of children.**

Hannah and Jacob both have alleles A and B.

- What blood types can their children have?
- What is the probability that their next child has each of these blood types?

#### **4.43 Parents with alleles B and O.**

Nancy and David both have alleles B and O.

- What blood types can their children have?
- What is the probability that their next child has each of these blood types?

#### **4.44 Two children.**

Jennifer has alleles A and O. José has alleles A and B. They have two children. What is the probability that both children have blood type A? What is the probability that both children have the same blood type?

#### **4.45 Three children.**

Jasmine has alleles A and O. Joshua has alleles B and O.

- What is the probability that a child of these parents has blood type O?
- If Jasmine and Joshua have three children, what is the probability that all three have blood type O? What is the probability that the first child has blood type O and the next two do not?

## 4.3 Random Variables

When you complete this section, you will be able to

- Describe the probability distribution of a discrete random variable.
- Use a probability histogram to provide a graphical description of the probability distribution of a discrete random variable.
- Use the distribution of a discrete random variable to calculate probabilities of events.
- Find probabilities of events for the uniform distribution.

Sample spaces need not consist of numbers. When we toss a coin four times, we can record the outcome as a string of heads and tails, such as HTTH. In statistics, however, we are most often interested in numerical outcomes such as the count of heads in the four tosses. It is convenient to use a shorthand notation: Let  $X$  be the number of heads. If our outcome is HTTH, then  $X = 2$ . If the next outcome is TTTH, the value of  $X$  changes to  $X = 1$ . The possible values of  $X$  are 0, 1, 2, 3, and 4. Tossing a coin four times will give  $X$  one of these possible values. Tossing four more times will give  $X$  another and probably different value. We call  $X$  a *random variable* because its values vary when the coin tossing is repeated.

### RANDOM VARIABLE

A **random variable** is a variable whose value is a numerical outcome of a random process.

In our coin-tossing example above, the process is the tossing of a coin four times. The random variable is the number of heads in the four tosses.

We usually denote random variables by capital letters near the end of the alphabet, such as  $X$  or  $Y$ . Of course, the random variables of greatest interest to us are outcomes such as the mean  $\bar{x}$  of a random sample, for which we will keep the familiar notation.<sup>10</sup> As we progress from general rules of probability toward statistical inference, we will concentrate on random variables.

When a random variable  $X$  describes a random process, the sample space  $S$  just lists the possible values of the random variable. We usually do not mention  $S$  separately. There remains the second part of any probability model, the assignment of probabilities to events. There are two main ways of assigning probabilities to the values of a random variable. The two types of probability models that result will

dominate our application of probability to statistical inference.

## Discrete random variables

We have learned several rules of probability, but only one method of assigning probabilities: state the probabilities of the individual outcomes and assign probabilities to events by summing over the outcomes. The outcome probabilities must be between 0 and 1 and have sum 1. When the outcomes are numerical, they are values of a random variable. We will now attach a name to random variables having probability assigned in this way.<sup>11</sup>

### DISCRETE RANDOM VARIABLE

A **discrete random variable**  $X$  has possible values that can be given in an ordered list. The **probability distribution** of  $X$  lists the values and their probabilities:

Value of $X$	$x_1$	$x_2$	$x_3$	...
Probability	$p_1$	$p_2$	$p_3$	...

The probabilities  $p_i$  must satisfy two requirements:

1. Every probability  $p_i$  is a number between 0 and 1.
2.  $p_1 + p_2 + \dots = 1$ .

Find the probability of any event by adding the probabilities  $p_i$  of the particular values  $x_i$  that make up the event.

In most of the situations that we will study, the number of possible values is a finite number,  $k$ . Think about the number of heads in four tosses of a coin. There are  $k = 5$  possible values: 0, 1, 2, 3, and 4.

However, there are settings in which the number of possible values is infinite. Think about tossing a fair coin until you get a head.

### Example

#### 4.22 Grade distributions.

A liberal arts college posts the grade distributions for its courses. In a recent semester, students in one section of English 130 received 31% A's, 40% B's, 20% C's, 4% D's, and 5% F's. Choose an English 130 student at random. To "choose at random" means to give every student the same chance to be chosen. The student's grade on a five-point scale (with A = 4) is a random variable  $X$

The value of  $X$  changes when we repeatedly choose students at random, but it is always one of 0, 1, 2, 3, or 4. Here is the distribution of  $X$

Value of $X$	0	1	2	3	4
Probability	0.05	0.04	0.20	0.40	0.31

The probability that the student got a B or better is the sum of the probabilities of an A and a B. In the language of random variables,

$$\begin{aligned}
 P(X \geq 3) &= P(X = 3) + P(X = 4) \\
 &= 0.40 + 0.31 = 0.71
 \end{aligned}$$

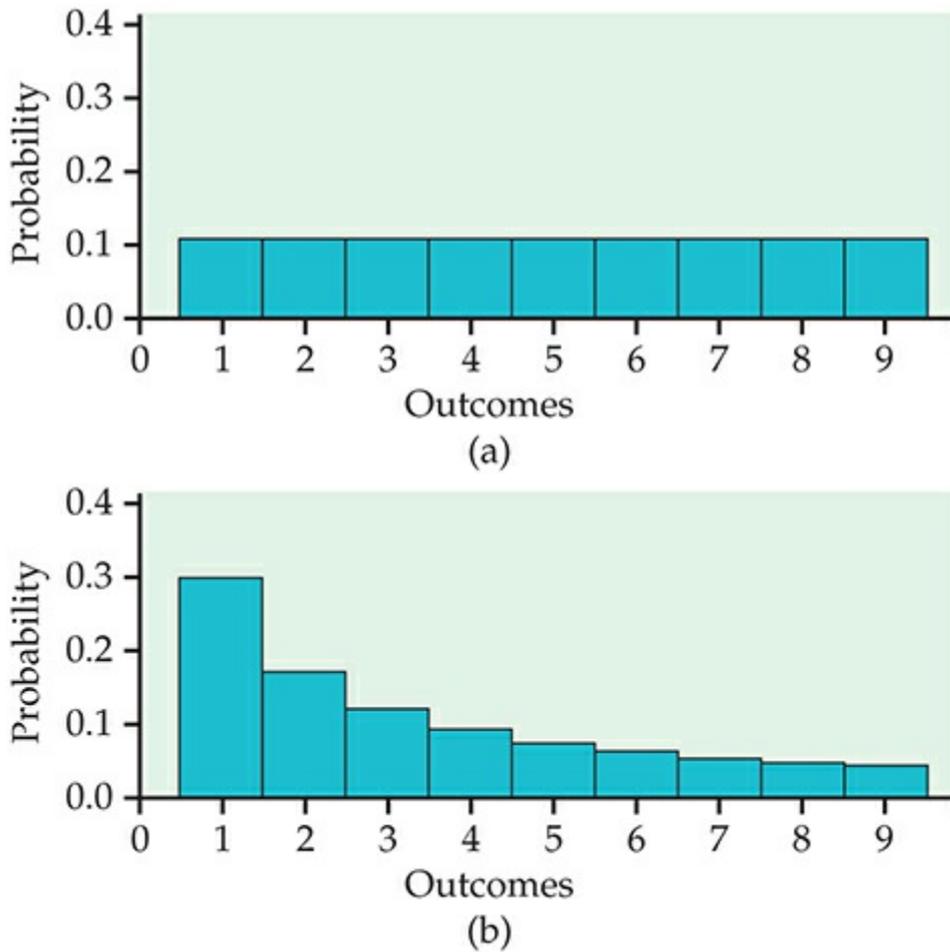
## USE YOUR KNOWLEDGE

### 4.46 Will the course satisfy the requirement?

Refer to Example 4.22. Suppose that a grade of D or F in English 130 will not count as satisfying a requirement for a major in linguistics. What is the probability that a randomly selected student will not satisfy this requirement?

We can use histograms to show probability distributions as well as distributions of data. Figure 4.5 displays **probability histograms** that compare the probability model for equally likely random digits (Example 4.15) with the model given by Benford's law (Example 4.12). The height of each bar shows the probability of the outcome at its base. Because the heights are probabilities, they add to 1. As usual, all the bars in a histogram have the same width. So the areas also display the assignment of probability to outcomes. Think of these histograms as idealized pictures of the results of very many trials. The histograms make it easy to quickly compare the two distributions.

probability histogram



**FIGURE 4.5** Probability histograms for (a) equally likely random digits 1 to 9 and (b) Benford's law. The height of each bar shows the probability assigned to a single outcome.

## Example

### 4.23 Number of heads in four tosses of a coin.

What is the probability distribution of the discrete random variable  $X$  that counts the number of heads in four tosses of a coin? We can derive this distribution if we make two reasonable assumptions:

- The coin is balanced, so it is fair and each toss is equally likely to give H or T.
- The coin has no memory, so tosses are independent.

The outcome of four tosses is a sequence of heads and tails such as HTTH.

There are 16 possible outcomes in all. Figure 4.6 lists these outcomes along with the value of  $X$  for each outcome. The multiplication rule for independent events tells us that, for example,

$$P(\text{HTTH})=12 \times 12 \times 12 \times 12=116$$

Each of the 16 possible outcomes similarly has probability  $1/16$ . That is, these outcomes are equally likely.

The number of heads  $X$  has possible values 0, 1, 2, 3, and 4. These values are *not* equally likely. As Figure 4.6 shows, there is only one way that  $X = 0$  can occur: namely, when the outcome is TTTT. So

$$P(X=0)=1/16=0.0625$$

The event  $\{X = 2\}$  can occur in six different ways, so that

$$\begin{aligned} P(X=2) &= \text{count of ways } X=2 \text{ can occur} / 16 \\ &= 6/16=0.375 \end{aligned}$$

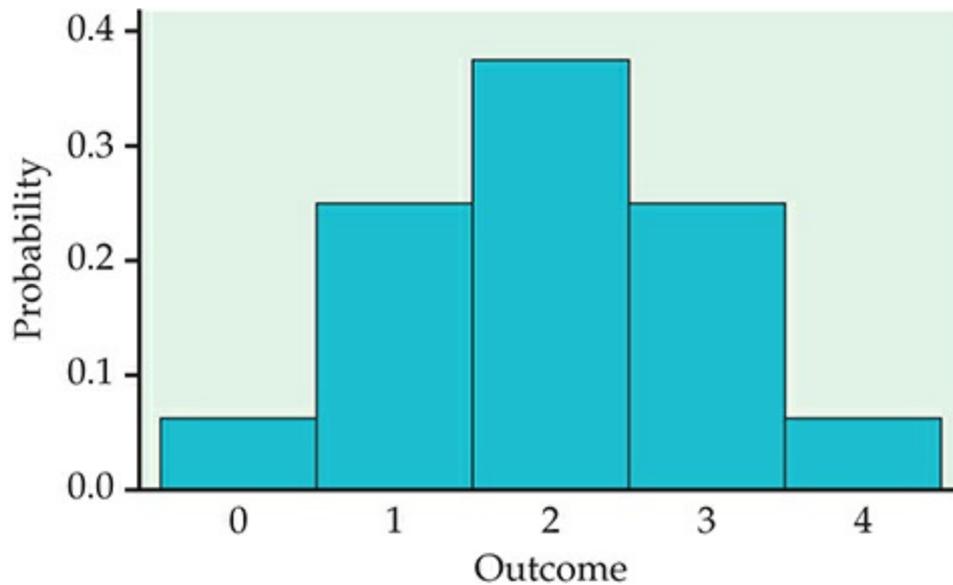
We can find the probability of each value of  $X$  from Figure 4.6 in the same way. Here is the result:

Value of $X$	0	1	2	3	4
Probability	0.0625	0.25	0.375	0.25	0.0625

			HTTH		
			HTHT		
	H T T T		T H T H	H H H T	
	T H T T		H H T T	H H T H	
	T T H T		T H H T	H T H H	
T T T T	T T T H		T T H H	T H H H	H H H H
$X = 0$	$X = 1$		$X = 2$	$X = 3$	$X = 4$

**FIGURE 4.6**

Possible outcomes in four tosses of a coin, for Example 4.23. The outcomes are arranged by the values of the random variable  $X$ , the number of heads.



**FIGURE 4.7**

Probability histogram for the number of heads in four tosses of a coin.

Figure 4.7 is a probability histogram for the distribution in Example 4.23. The probability distribution is exactly symmetric. The probabilities (bar heights) are idealizations of the proportions after very many tosses of four coins. The actual distribution of proportions observed would be nearly symmetric but is unlikely to be exactly symmetric.

## Example

### 4.24 Probability of at least two heads.

Any event involving the number of heads observed can be expressed in terms of  $X$ , and its probability can be found from the distribution of  $X$ . For example, the probability of tossing at least two heads is

$$P(X \geq 2) = 0.375 + 0.25 + 0.0625 = 0.6875$$

The probability of at least one head is most simply found by use of the complement rule:

$$\begin{aligned} P(X \geq 1) &= 1 - P(X = 0) \\ &= 1 - 0.0625 = 0.9375 \end{aligned}$$

Recall that tossing a coin  $n$  times is similar to choosing an SRS of size  $n$  from a

large population and asking a Yes or No question. We will extend the results of Example 4.23 when we return to sampling distributions in the next chapter.

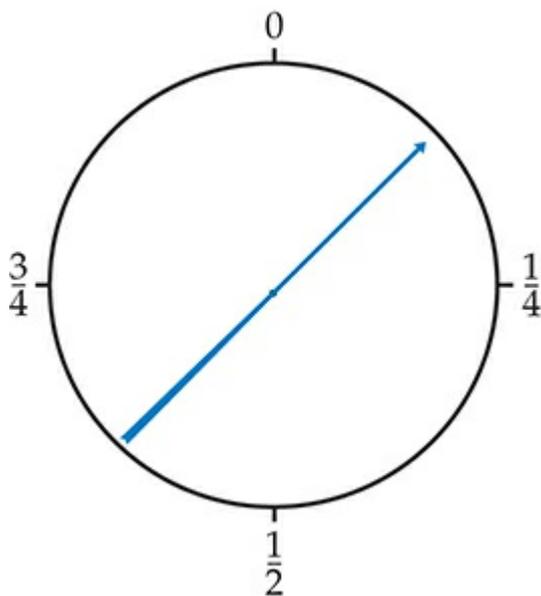
## USE YOUR KNOWLEDGE

### 4.47 Two tosses of a fair coin.

Find the probability distribution for the number of heads that appear in two tosses of a fair coin.

## Continuous random variables

When we use the table of random digits to select a digit between 0 and 9, the result is a discrete random variable. The probability model assigns probability  $1/10$  to each of the 10 possible outcomes. Suppose that we want to choose a number at random between 0 and 1, allowing *any* number between 0 and 1 as the outcome. Software random number generators will do this.



**FIGURE 4.8**

A spinner that generates a random number between 0 and 1.

You can visualize such a random number by thinking of a spinner (Figure 4.8) that turns freely on its axis and slowly comes to a stop. The pointer can come to rest anywhere on a circle that is marked from 0 to 1. The sample space is now an entire interval of numbers:

$$S = \{\text{all numbers } x \text{ such that } 0 \leq x \leq 1\}$$

How can we assign probabilities to events such as  $\{0.3 \leq x \leq 0.7\}$ ? As in the case of selecting a random digit, we would like all possible outcomes to be equally likely. But we cannot assign probabilities to each individual value of  $x$  and then sum, because there are infinitely many possible values. Instead, we use a new way of assigning probabilities directly to events—as *areas under a density curve*. Any density curve has area exactly 1 underneath it, corresponding to total probability 1.

## Example

### 4.25 Uniform random numbers.

The random number generator will spread its output uniformly across the entire interval from 0 to 1 as we allow it to generate a long sequence of numbers. The results of many trials are represented by the density curve of a **uniform distribution**.

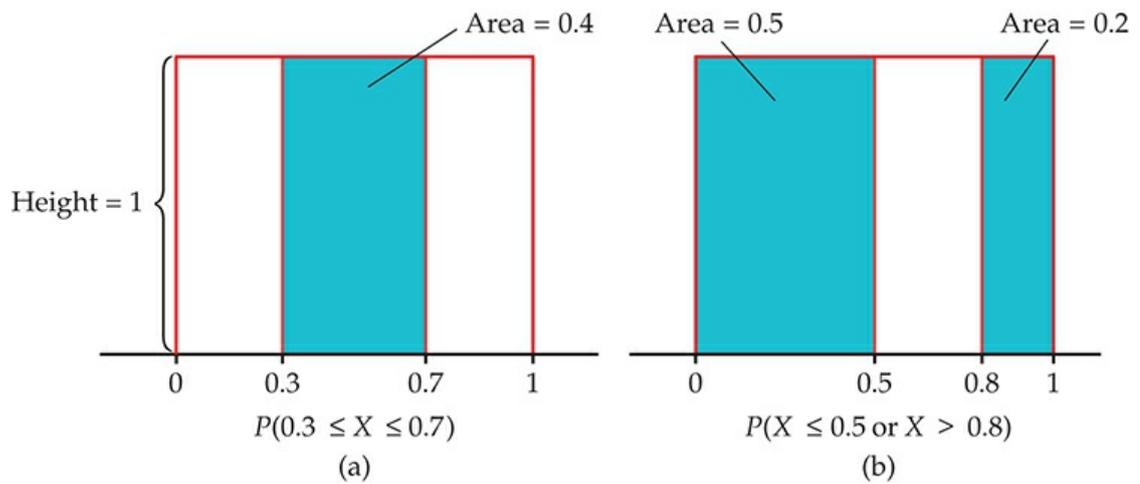
#### uniform distribution

This density curve appears in red in Figure 4.9. It has height 1 over the interval from 0 to 1, and height 0 everywhere else. The area under the density curve is 1: the area of a square with base 1 and height 1. The probability of any event is the area under the density curve and above the event in question.

As Figure 4.9(a) illustrates, the probability that the random number generator produces a number  $X$  between 0.3 and 0.7 is

$$P(0.3 \leq X \leq 0.7) = 0.4$$

because the area under the density curve and above the interval from 0.3 to 0.7 is 0.4. The height of the density curve is 1, and the area of a rectangle is the product of height and length, so the probability of any interval of outcomes is just the length of the interval.



**FIGURE 4.9**

Assigning probabilities for generating a random number between 0 and 1, for Example 4.25. The probability of any interval of numbers is the area above the interval and under the density curve.

Similarly,

$$P(X \leq 0.5) = 0.5$$

$$P(X > 0.8) = 0.2$$

$$P(X \leq 0.5 \text{ or } X > 0.8) = 0.7$$

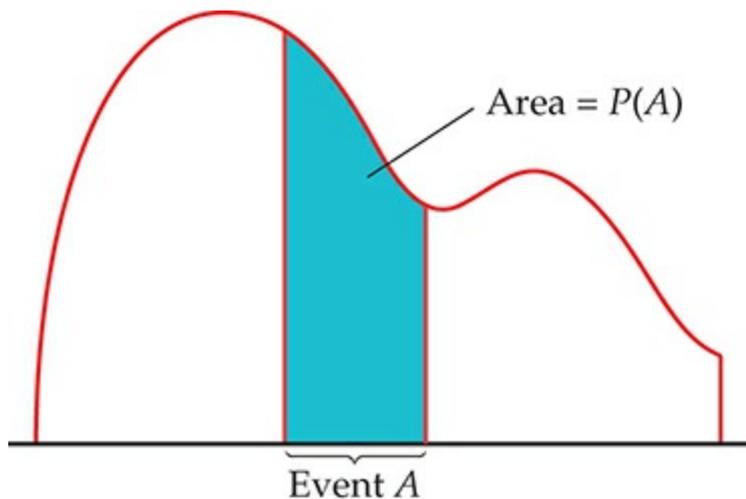
Notice that the last event consists of two nonoverlapping intervals, so the total area above the event is found by adding two areas, as illustrated by Figure 4.9(b). This assignment of probabilities obeys all of our rules for probability.

## USE YOUR KNOWLEDGE

### 4.48 Find the probability.

For the uniform distribution described in Example 4.25, find the probability that  $X$  is between 0.2 and 0.7.

Probability as area under a density curve is a second important way of assigning probabilities to events. Figure 4.10 illustrates this idea in general form. We call  $X$  in Example 4.25 a *continuous random variable* because its values are not isolated numbers but an entire interval of numbers.



**FIGURE 4.10**

The probability distribution of a continuous random variable assigns probabilities as areas under a density curve. The total area under any density curve is 1.

## CONTINUOUS RANDOM VARIABLE

A **continuous random variable**  $X$  takes all values in an interval of numbers. The **probability distribution** of  $X$  is described by a density curve. The probability of any event is the area under the density curve and above the values of  $X$  that make up the event.

The probability model for a continuous random variable assigns probabilities to intervals of outcomes rather than to individual outcomes. In fact, **all continuous probability distributions assign probability 0 to every individual outcome**. Only intervals of values have positive probability. To see that this is true, consider a specific outcome such as  $P(X = 0.8)$  in the context of Example 4.25. The probability of any interval is the same as its length. The point 0.8 has no length, so its probability is 0.

Although this fact may seem odd, it makes intuitive, as well as mathematical, sense. The random number generator produces a number between 0.79 and 0.81 with probability 0.02. An outcome between 0.799 and 0.801 has probability 0.002. A result between 0.799999 and 0.800001 has probability 0.000002. You see that as we approach 0.8 the probability gets closer to 0.



To be consistent, the probability of an outcome *exactly* equal to 0.8 must be 0. Because there is no probability exactly at  $X = 0.8$ , the two events  $\{X > 0.8\}$  and  $\{X \geq 0.8\}$  have the same probability. *We can ignore the distinction between  $>$  and  $\geq$*

when finding probabilities for continuous (but not discrete) random variables.

## Normal distributions as probability distributions

The density curves that are most familiar to us are the Normal curves. Because any density curve describes an assignment of probabilities, *Normal distributions are probability distributions*. Recall that  $N(\mu, \sigma)$  is our shorthand for the Normal distribution having mean  $\mu$  and standard deviation  $\sigma$ . In the language of random variables, if  $X$  has the  $N(\mu, \sigma)$  distribution, then the standardized variable

$$Z = \frac{X - \mu}{\sigma}$$

is a standard Normal random variable having the distribution  $N(0, 1)$

### Example

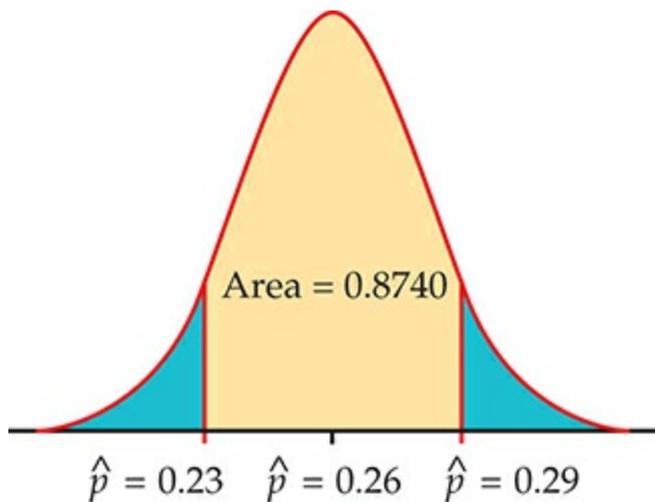
#### 4.26 Texting while driving.



parameter, statistic, p. 206

Texting while driving can be dangerous, but young people want to remain connected. Suppose that 26% of teen drivers text while driving. If we take a sample of 500 teen drivers, what percent would we expect to say that they text while driving?<sup>12</sup>

The proportion  $p = 0.26$  is a *parameter* that describes the population of teen drivers. The proportion  $\hat{p}$  of the sample who say that they text while driving is a *statistic* used to estimate  $p$ . The statistic  $\hat{p}$  is a random variable because repeating the SRS would give a different sample of 500 teen drivers and a different value of  $\hat{p}$ .



**FIGURE 4.11**

Probability in Example 4.26 as area under a Normal density curve.

The statistic  $\hat{p}$  has approximately the  $N(0.26, 0.0196)$  distribution. The mean 0.26 of this distribution is the same as the population parameter because  $\hat{p}$  is an unbiased estimate of  $p$ . The standard deviation is controlled mainly by the size of the sample.

**← LOOK BACK**

Normal distribution calculations, p. 63

What is the probability that the survey result differs from the truth about the population by no more than 3 percentage points? We can use what we learned about Normal distribution calculations to answer this question. Because  $p = 0.26$ , the survey misses by no more than 3 percentage points if the sample proportion is between 0.23 and 0.29.

Figure 4.11 shows this probability as an area under a Normal density curve. You can find it by software or by standardizing and using Table A. From Table A,

$$\begin{aligned} P(0.23 \leq \hat{p} \leq 0.29) &= P(0.23 - 0.26 / 0.0196 \leq \hat{p} - 0.26 / 0.0196 \leq 0.29 - 0.26 / 0.0196) \\ &= P(-1.53 \leq Z \leq 1.53) \\ &= 0.9370 - 0.0630 = 0.8740 \end{aligned}$$

About 87% of the time, the sample  $\hat{p}$  will be within 3 percentage points of the parameter  $p$ .

We began this chapter with a general discussion of the idea of probability and the properties of probability models. Two very useful specific types of probability models are distributions of discrete and continuous random variables. In our study of statistics we will employ only these two types of probability models.

## SECTION 4.3 Summary

A **random variable** is a variable taking numerical values determined by the outcome of a random phenomenon. The **probability distribution** of a random variable  $X$  tells us what the possible values of  $X$  are and how probabilities are assigned to those values.

A random variable  $X$  and its distribution can be **discrete** or **continuous**.

A **discrete random variable** has possible values that can be given in an ordered list. The probability distribution assigns each of these values a probability between 0 and 1 such that the sum of all the probabilities is exactly 1. The probability of any event is the sum of the probabilities of all the values that make up the event.

A **continuous random variable** takes all values in some interval of numbers. A **density curve** describes the probability distribution of a continuous random variable. The probability of any event is the area under the curve and above the values that make up the event.

**Normal distributions** are one type of continuous probability distribution.

You can picture a probability distribution by drawing a **probability histogram** in the discrete case or by graphing the density curve in the continuous case.

## SECTION 4.3 Exercises

*For Exercise 4.46, see page 254; for Exercise 4.47, see page 256; and for Exercise 4.48, see page 258.*

### 4.49 How many courses?

At a small liberal arts college, students can register for one to six courses. Let  $X$  be the number of courses taken in the fall by a randomly selected student from this college. In a typical fall semester, 5% take one course, 5% take two courses, 13% take three courses, 26% take four courses, 36% take five courses, and 15% take six courses. Let  $X$  be the number of courses taken in the fall by a randomly selected student from this college. Describe the probability distribution of this random variable.

### 4.50 Make a graphical display.

Refer to the previous exercise. Use a probability histogram to provide a graphical description of the distribution of  $X$ .

### 4.51 Find some probabilities.

Refer to Exercise 4.49.

- Find the probability that a randomly selected student takes three or fewer courses.
- Find the probability that a randomly selected student takes four or five courses.
- Find the probability that a randomly selected student takes eight courses.

### 4.52 Use the uniform distribution.

Suppose that a random variable  $X$  follows the uniform distribution described in Example 4.25 (page 257). For each of the following events, find the probability and illustrate your calculations with a sketch of the density curve similar to the ones in Figure 4.9 (page 258).

- (a) The probability that  $X$  is less than 0.1.
- (b) The probability that  $X$  is greater than or equal to 0.8.
- (c) The probability that  $X$  is less than 0.7 and greater than 0.5.
- (d) The probability that  $X$  is 0.5.

### 4.53 What's wrong?

In each of the following scenarios, there is something wrong. Describe what is wrong and give a reason for your answer.

- (a) The probabilities for a discrete statistic always add to 1.
- (b) A continuous random variable can take any value between 0 and 1.
- (c) Normal distributions are discrete random variables.

### 4.54 Use of Twitter.

Suppose that the population proportion of Internet users who say that they use Twitter or another service to post updates about themselves or to see updates about others is 19%.<sup>13</sup> Think about selecting random samples from a population in which 19% are Twitter users.

- (a) Describe the sample space for selecting a single person.
- (b) If you select three people, describe the sample space.
- (c) Using the results of (b), define the sample space for the random variable that expresses the number of Twitter users in the sample of size 3.
- (d) What information is contained in the sample space for part (b) that is not contained in the sample space for part (c)? Do you think this information is important? Explain your answer.

### 4.55 Use of Twitter.

Find the probabilities for parts (a), (b), and (c) of the previous exercise.

### 4.56 Households and families in government data.

In government data, a household consists of all occupants of a dwelling unit, while a family consists of two or more persons who live together and are related by blood or marriage. So all families form households, but some households are not families. Here are the distributions of household size and of family size in the United States:

Number of persons	1	2	3	4	5	6	7
Household probability	0.27	0.33	0.16	0.14	0.06	0.03	0.01
Family Probability	0	0.44	0.22	0.20	0.09	0.03	0.02

Make probability histograms for these two discrete distributions, using the same scales. What are the most important differences between the sizes of households and families?

### 4.57 Discrete or continuous?

In each of the following situations decide whether the random variable is discrete or continuous and give a reason for your answer.

- (a) Your web page has five different links, and a user can click on one of the links or can leave the page. You record the length of time that a user spends on the web page before clicking one of the links or leaving the page.
- (b) The number of hits on your web page.
- (c) The yearly income of a visitor to your web page.

### 4.58 Texas hold 'em.

The game of Texas hold 'em starts with each player receiving two cards. Here is the probability distribution for the number of aces in two-card hands:

Number of aces	0	1	2
Probability	0.8507	0.1448	0.0045

- (a) Verify that this assignment of probabilities satisfies the requirement that the sum of the probabilities for a discrete distribution must be 1.
- (b) Make a probability histogram for this distribution.
- (c) What is the probability that a hand contains at least one ace? Show two different ways to calculate this probability.

### 4.59 Tossing two dice.

Some games of chance rely on tossing two dice. Each die has six faces, marked with 1, 2, . . . , 6 spots called pips. The dice used in casinos are carefully balanced so that each face is equally likely to come up. When two dice are tossed, each of the 36 possible pairs of faces is equally likely to come up. The outcome of interest to a gambler is the sum of the pips on the two up-faces. Call this random variable  $X$ .

- (a) Write down all 36 possible pairs of up-faces.
- (b) If all pairs have the same probability, what must be the probability of each pair?
- (c) Write the value of  $X$  next to each pair of up-faces and use this information with the result of (b) to give the probability distribution of  $X$ . Draw a probability histogram to display the distribution.
- (d) One bet available in the game called craps wins if a 7 or an 11 comes up on the next roll of two dice. What is the probability of rolling a 7 or an 11 on the next roll?
- (e) Several bets in craps lose if a 7 is rolled. If any outcome other than 7 occurs, these bets either win or continue to the next roll. What is the probability that anything other than a 7 is rolled?



### 4.60 Nonstandard dice.

Nonstandard dice can produce interesting distributions of outcomes. You have two balanced, six-sided dice. One is a standard die, with faces having 1, 2, 3, 4, 5, and 6 spots. The other die has three faces with 0 spots and three faces with 6 spots. Find the probability distribution for the total number of spots  $Y$  on the up-faces when you roll these two dice.

#### 4.61 Spell-checking software.

Spell-checking software catches “nonword errors,” which are strings of letters that are not words, as when “the” is typed as “eth.” When undergraduates are asked to write a 250-word essay (without spell-checking), the number  $X$  of nonword errors has the following distribution:

Value of $X$	0	1	2	3	4
Probability	0.1	0.3	0.3	0.2	0.1

- Sketch the probability distribution for this random variable.
- Write the event “at least one nonword error” in terms of  $X$ . What is the probability of this event?
- Describe the event  $X \leq 2$  in words. What is its probability? What is the probability that  $X < 2$ ?

#### 4.62 Find the probabilities.

Let the random variable  $X$  be a random number with the uniform density curve in Figure 4.9 (page 258). Find the following probabilities:

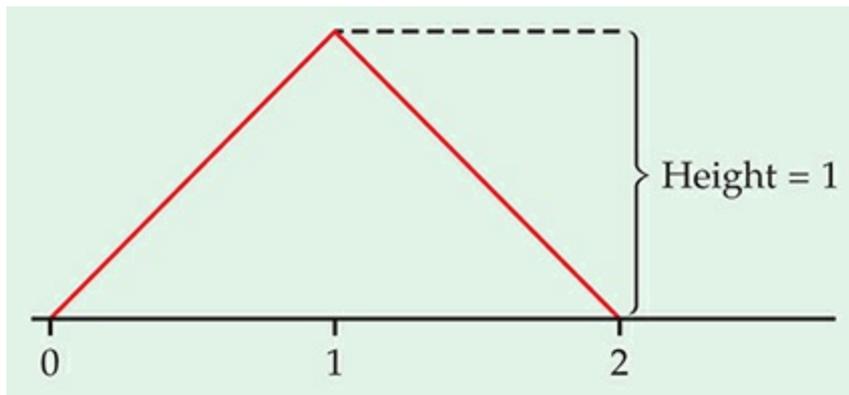
- $P(X \geq 0.30)$
- $P(X = 0.30)$
- $P(0.30 < X < 1.30)$
- $P(0.20 \leq X \leq 0.25 \text{ or } 0.7 \leq X \leq 0.92)$
- $X$  is not in the interval 0.4 to 0.7

#### 4.63 Uniform numbers between 0 and 2.

Many random number generators allow users to specify the range of the random numbers to be produced. Suppose that you specify that the range is to be all numbers between 0 and 2. Call the random number generated  $Y$ . Then the density curve of the random variable  $Y$  has constant height between 0 and 2, and height 0 elsewhere.

- What is the height of the density curve between 0 and 2? Draw a graph of the density curve.
- Use your graph from (a) and the fact that probability is area under the curve to find  $P(Y \leq 1.6)$ .
- Find  $P(0.5 < Y < 1.7)$ .
- Find  $P(Y \geq 0.95)$ .

#### 4.64 The sum of two uniform random numbers.



**FIGURE 4.12**

The density curve for the sum  $Y$  of two random numbers, for Exercise 4.64.

Generate *two* random numbers between 0 and 1 and take  $Y$  to be their sum. Then  $Y$  is a continuous random variable that can take any value between 0 and 2. The density curve of  $Y$  is the triangle shown in Figure 4.12.

- Verify by geometry that the area under this curve is 1.
- What is the probability that  $Y$  is less than 1? (Sketch the density curve, shade the area that represents the probability, then find that area. Do this for (c) also.)
- What is the probability that  $Y$  is greater than 0.6?

#### 4.65 How many close friends?

How many close friends do you have? Suppose that the number of close friends adults claim to have varies from person to person with mean  $\mu = 9$  and standard deviation  $\sigma = 2.4$ . An opinion poll asks this question of an SRS of 1100 adults. We will see in the next chapter that in this situation the sample mean response  $\bar{x}$  has approximately the Normal distribution with mean 9 and standard deviation 0.0724. What is  $P(8 \leq \bar{x} \leq 10)$  the probability that the statistic  $\bar{x}$  estimates the parameter  $\mu$  to within  $\pm 1$ ?

#### 4.66 Normal approximation for a sample proportion.

A sample survey contacted an SRS of 700 registered voters in Oregon shortly after an election and asked respondents whether they had voted. Voter records show that 56% of registered voters had actually voted. We will see in the next chapter that in this situation the proportion  $\hat{p}$  of the sample who voted has approximately the Normal distribution with mean  $\mu = 0.56$  and standard deviation  $\sigma = 0.019$ .

- If the respondents answer truthfully, what is  $P(0.52 \leq \hat{p} \leq 0.60)$  This is the probability that the statistic  $\hat{p}$  estimates the parameter 0.56 within plus or minus 0.04.
- In fact, 72% of the respondents said they had voted ( $\hat{p} = 0.72$ ). If respondents answer truthfully, what is  $P(\hat{p} \geq 0.72)$ ? This probability is so small that it is good evidence that some people who did not vote claimed that they did vote.

## 4.4 Means and Variances of Random Variables

When you complete this section, you will be able to

- Use a probability distribution to find the mean of a discrete random variable.
- Apply the law of large numbers to describe the behavior of the sample mean as the sample size increases.
- Find means using the rules for means of linear transformations, sums, and differences.
- Use a probability distribution to find the variance and the standard deviation of a discrete random variable.
- Find variances and standard deviations using the rules for variances and standard deviations for linear transformations.
- Find variances and standard deviations using the rules for variances and standard deviations for sums of and differences between two random variables, for uncorrelated and for correlated random variables.

The probability histograms and density curves that picture the probability distributions of random variables resemble our earlier pictures of distributions of data. In describing data, we moved from graphs to numerical measures such as means and standard deviations. Now we will make the same move to expand our descriptions of the distributions of random variables. We can speak of the mean winnings in a game of chance or the standard deviation of the randomly varying number of calls a travel agency receives in an hour. In this section we will learn more about how to compute these descriptive measures and about the laws they obey.

### The mean of a random variable

In Chapter 1 (page 31), we learned that the mean  $\bar{x}$  is the average of the observations in a *sample*. Recall that a random variable  $X$  is a numerical outcome of a random process. Think about repeating the random process many times and recording the resulting values of the random variable. You can think of the value of a random variable as the average of a very large sample where the relative frequencies of the values are the same as their probabilities.

If we think of the random process as corresponding to the population, then the mean of the random variable is a parameter of this population. Here is an example.

## Example

### 4.27 The Tri-State Pick 3 lottery.

Most states and Canadian provinces have government-sponsored lotteries. Here is a simple lottery wager, from the Tri-State Pick 3 game that New Hampshire shares with Maine and Vermont. You choose a three-digit number, 000 to 999. The state chooses a three-digit winning number at random and pays you \$500 if your number is chosen.

Because there are 1000 three-digit numbers, you have probability  $1/1000$  of winning. Taking  $X$  to be the amount your ticket pays you, the probability distribution of  $X$  is

Payoff $X$	\$0	\$500
Probability	0.999	0.001

The random process consists of drawing a three-digit number. The population consists of the numbers 000 to 999. Each of these possible outcomes is equally likely in this example. In the setting of sampling in Chapter 3 (page 194), we can view the random process as selecting an SRS of size 1 from the population. The random variable  $X$  is 1 if the selected number is equal to the one that you chose and is 0 if it is not.

What is your average payoff from many tickets? The ordinary average of the two possible outcomes \$0 and \$500 is \$250, but that makes no sense as the average because \$500 is much less likely than \$0. In the long run you receive \$500 once in every 1000 tickets and \$0 on the remaining 999 of 1000 tickets. The long-run average payoff is

$$\$500 \frac{1}{1000} + \$0 \frac{999}{1000} = \$0.50$$

or 50 cents. That number is the mean of the random variable  $X$ . (Tickets cost \$1, so in the long run the state keeps half the money you wager.)

If you play Tri-State Pick 3 several times, we would as usual call the mean of the actual amounts you win  $\bar{x}$ . The mean in Example 4.27 is a different quantity—it is the long-run average winnings you expect if you play a very large number of times.

## USE YOUR KNOWLEDGE

#### 4.67 Find the mean of the probability distribution.

You toss a fair coin. If the outcome is heads, you win \$5.00; if the outcome is tails, you win nothing. Let  $X$  be the amount that you win in a single toss of a coin. Find the probability distribution of this random variable and its mean.

Just as probabilities are an idealized description of long-run proportions, the mean of a probability distribution describes the long-run average outcome. We can't call this mean  $\bar{x}$ , so we need a different symbol. The common symbol for the **mean of a probability distribution** is  $\mu$  the Greek letter mu. We used  $\mu$  in Chapter 1 for the mean of a Normal distribution, so this is not a new notation. We will often be interested in several random variables, each having a different probability distribution with a different mean.

#### mean $\mu$

To remind ourselves that we are talking about the mean of  $X$  we often write  $\mu_X$  rather than simply  $\mu$ . In Example 4.27,  $\mu_X = \$0.50$ . Notice that, as often happens, the mean is not a possible value of  $X$ . You will often find the mean of a random variable  $X$  called the **expected value** of  $X$ . This term can be misleading, for we don't necessarily expect one observation on  $X$  to be close to its expected value.

#### expected value

The mean of any discrete random variable is found just as in Example 4.27. It is an average of the possible outcomes, but a weighted average in which each outcome is weighted by its probability. Because the probabilities add to 1, we have total weight 1 to distribute among the outcomes. An outcome that occurs half the time has probability one-half and gets one-half the weight in calculating the mean. Here is the general definition.

### MEAN OF A DISCRETE RANDOM VARIABLE

Suppose that  $X$  is a **discrete random variable** whose distribution is

Value of $X$	$x_1$	$x_2$	$x_3$	...	$x_k$
Probability	$p_1$	$p_2$	$p_3$	...	$p_k$

To find the **mean** of  $X$  multiply each possible value by its probability, then add all the products:

$$\mu_X = x_1 p_1 + x_2 p_2 + \dots + x_k p_k$$

$$= \sum x_i p_i$$

## Example

### 4.28 The mean of equally likely first digits.

If first digits in a set of data all have the same probability, the probability distribution of the first digit  $X$  is then

First digit $X$	1	2	3	4	5	6	7	8	9
Probability	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9

The mean of this distribution is

$$\begin{aligned} \mu_X &= 1 \times \frac{1}{9} + 2 \times \frac{1}{9} + 3 \times \frac{1}{9} + 4 \times \frac{1}{9} + 5 \times \frac{1}{9} \\ &\quad + 6 \times \frac{1}{9} + 7 \times \frac{1}{9} + 8 \times \frac{1}{9} + 9 \times \frac{1}{9} \\ &= 45 \times \frac{1}{9} = 5 \end{aligned}$$

Suppose that the random digits in Example 4.28 had a different probability distribution. In Example 4.12 (page 242) we described Benford's law as a probability distribution that describes first digits of numbers in many real situations. Let's calculate the mean for Benford's law.

## Example

### 4.29 The mean of first digits that follow Benford's law.

Here is the distribution of the first digit for data that follow Benford's law. We use the letter  $V$  for this random variable to distinguish it from the one that we studied in Example 4.28. The distribution of  $V$  is

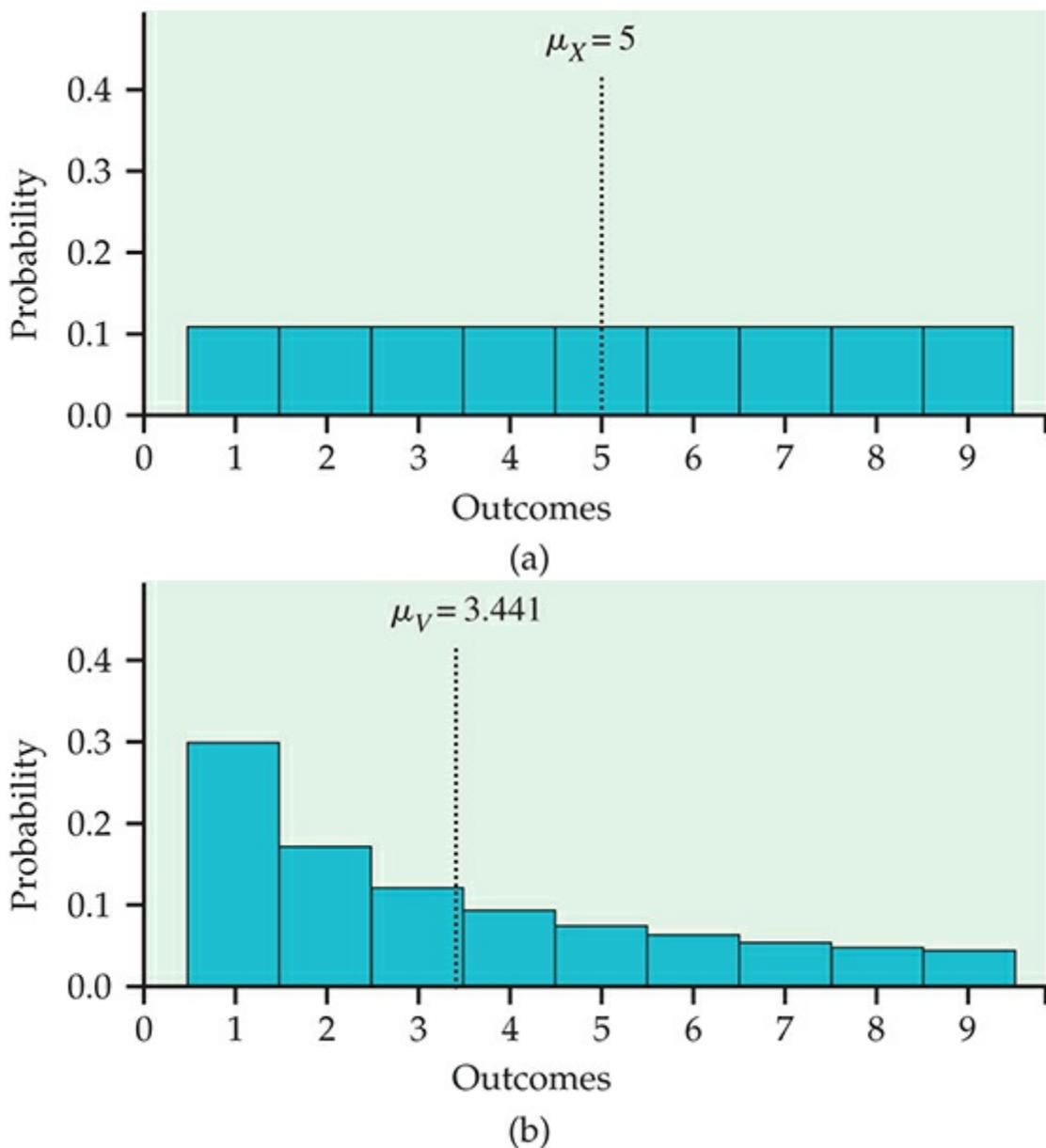
First digit $V$	1	2	3	4	5	6	7	8	9
Probability	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

The mean of  $V$  is

$$\begin{aligned}\mu_V &= (1)(0.301) + (2)(0.176) + (3)(0.125) + (4)(0.097) + (5)(0.079) + (6) \\ &\quad (0.067) + (7)(0.058) + (8)(0.051) + (9)(0.046) \\ &= 3.441\end{aligned}$$

The mean reflects the greater probability of smaller first digits under Benford's law than when first digits 1 to 9 are equally likely.

Figure 4.13 locates the means of  $X$  and  $V$  on the two probability histograms. Because the discrete uniform distribution of Figure 4.13(a) is symmetric, the mean lies at the center of symmetry. We can't locate the mean of the right-skewed distribution of Figure 4.13(b) by eye—calculation is needed.



**FIGURE 4.13**

Locating the mean of a discrete random variable on the probability histogram for (a) digits between 1 and 9 chosen at random; (b) digits between 1 and 9 chosen from records that obey Benford's law.

What about continuous random variables? The probability distribution of a continuous random variable  $X$  is described by a density curve. Chapter 1 (page 56) showed how to find the mean of the distribution: it is the point at which the area under the density curve would balance if it were made out of solid material. The mean lies at the center of symmetric density curves such as the Normal curves. Exact calculation of the mean of a distribution with a skewed density curve requires advanced mathematics.<sup>14</sup> The idea that the mean is the balance point of the distribution applies to discrete random variables as well, but in the discrete case we have a formula that gives us this point.

## Statistical estimation and the law of large numbers

We would like to estimate the mean height  $\mu$  of the population of all American women between the ages of 18 and 24 years. This  $\mu$  is the mean  $\mu_X$  of the random variable  $X$  obtained by choosing a young woman at random and measuring her height. To estimate  $\mu$  we choose an SRS of young women and use the sample mean  $\bar{x}$  to estimate the unknown population mean  $\mu$ . In the language of Section 3.4 (page 205),  $\mu$  is a *parameter* and  $\bar{x}$  is a *statistic*.



**sampling distributions**, p. 208

Statistics obtained from probability samples are random variables because their values vary in repeated sampling. The sampling distributions of statistics are just the probability distributions of these random variables.

It seems reasonable to use  $\bar{x}$  to estimate  $\mu$ . An SRS should fairly represent the population, so the mean  $\bar{x}$  of the sample should be somewhere near the mean  $\mu$  of the population. Of course, we don't expect  $\bar{x}$  to be exactly equal to  $\mu$  and we realize that if we choose another SRS, the luck of the draw will probably produce a different  $\bar{x}$ .

If  $\bar{x}$  is rarely exactly right and varies from sample to sample, why is it nonetheless a reasonable estimate of the population mean  $\mu$ ? We gave one answer in Section 3.4:  $\bar{x}$  is unbiased and we can control its variability by choosing the sample size. Here is another answer: if we keep on adding observations to our random sample, the statistic  $\bar{x}$  is *guaranteed* to get as close as we wish to the parameter  $\mu$  and then stay that close. We have the comfort of knowing that if we can afford to keep on measuring more women, eventually we will estimate the mean height of all young women very accurately. This remarkable fact is called the *law of large numbers*. It is remarkable because it holds for *any* population, not just for some special class such as Normal distributions.

## LAW OF LARGE NUMBERS

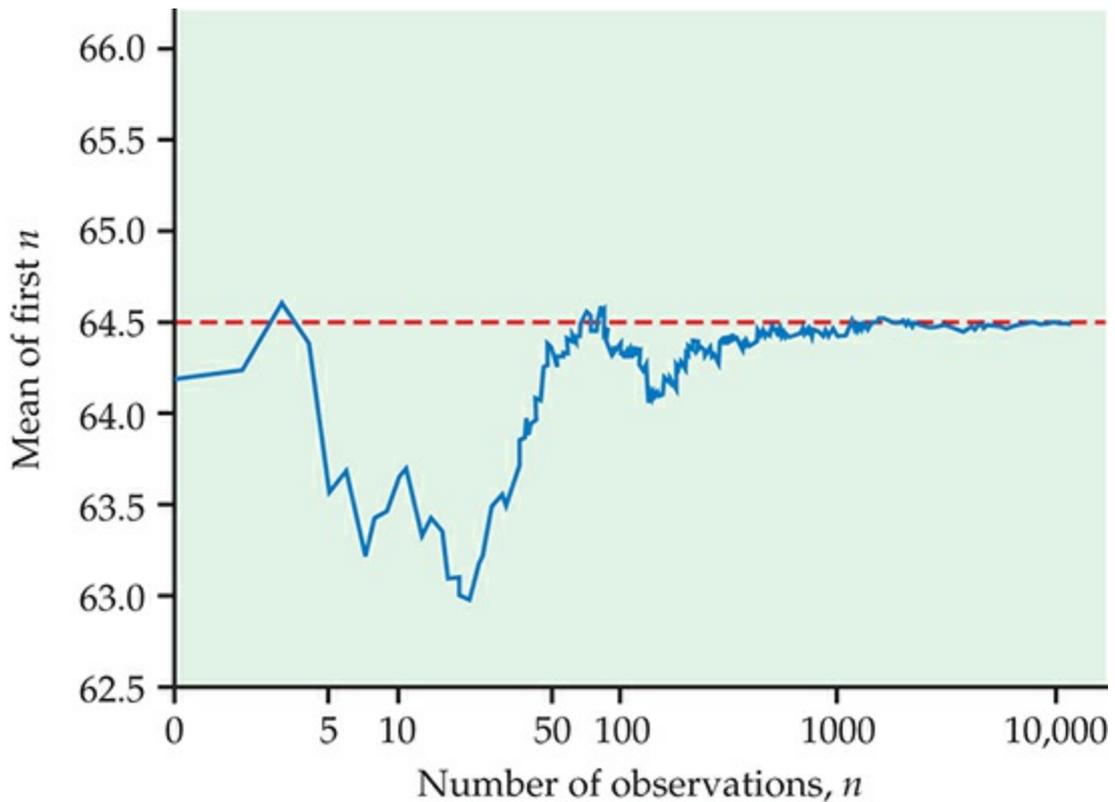
Draw independent observations at random from any population with finite mean  $\mu$ . Decide how accurately you would like to estimate  $\mu$ . As the number of observations drawn increases, the mean  $\bar{x}$  of the observed values eventually approaches the mean  $\mu$  of the population as closely as you specified and then stays that close.

The behavior of  $\bar{x}$  is similar to the idea of probability. In the long run, the *proportion* of outcomes taking any value gets close to the *probability* of that value, and the *average outcome* gets close to the *distribution mean*. Figure 4.1 (page 232) shows how proportions approach probability in one example. Here is an example of how sample means approach the distribution mean.

### Example

#### 4.30 Heights of young women.

The distribution of the heights of all young women is close to the Normal distribution with mean 64.5 inches and standard deviation 2.5 inches. Suppose that  $\mu = 64.5$  were exactly true.



**FIGURE 4.14**

The law of large numbers in action. As we take more observations, the sample mean always approaches the mean of the population.

Figure 4.14 shows the behavior of the mean height  $\bar{x}$  of  $n$  women chosen at random from a population whose heights follow the  $N(64.5, 2.5)$  distribution. The graph plots the values of  $\bar{x}$  as we add women to our sample. The first woman drawn had height 64.21 inches, so the line starts there. The second had height 64.35 inches, so for  $n = 2$  the mean is

$$\bar{x} = \frac{64.21 + 64.35}{2} = 64.28$$

This is the second point on the line in the graph.

At first, the graph shows that the mean of the sample changes as we take more observations. Eventually, however, the mean of the observations gets close to the population mean  $\mu = 64.5$  and settles down at that value. The law of large numbers says that this *always* happens.

## USE YOUR KNOWLEDGE

**4.68 Use the *Law of Large Numbers* applet.**



The *Law of Large Numbers* applet animates a graph like Figure 4.14 for rolling dice. Use it to better understand the law of large numbers by making a similar graph.

The mean  $\mu$  of a random variable is the average value of the variable in two senses. By its definition,  $\mu$  is the average of the possible values, weighted by their probability of occurring. The law of large numbers says that  $\mu$  is also the long-run average of many independent observations on the variable. The law of large numbers can be proved mathematically starting from the basic laws of probability.

### Thinking about the law of large numbers

The law of large numbers says broadly that the average results of many independent observations are stable and predictable. The gamblers in a casino may win or lose, but the casino will win in the long run because the law of large numbers says what the average outcome of many thousands of bets will be. An insurance company deciding how much to charge for life insurance and a fast-food restaurant deciding how many beef patties to prepare also rely on the fact that averaging over many individuals produces a stable result. It is worth the effort to think a bit more closely about so important a fact.

### The “law of small numbers”

Both the rules of probability and the law of large numbers describe the regular behavior of chance phenomena *in the long run*. Psychologists have discovered that our intuitive understanding of randomness is quite different from the true laws of chance.<sup>15</sup> For example, most people believe in an incorrect “law of small numbers.” That is, we expect even short sequences of random events to show the kind of average behavior that in fact appears only in the long run.

Some teachers of statistics begin a course by asking students to toss a coin 50 times and bring the sequence of heads and tails to the next class. The teacher then announces which students just wrote down a random-looking sequence rather than actually tossing a coin. The faked tosses don’t have enough “runs” of consecutive heads or consecutive tails. Runs of the same outcome don’t look random to us but are in fact common. For example, the probability of a run of three or more consecutive heads or tails in just 10 tosses is greater than 0.8.<sup>16</sup> The runs of consecutive heads or consecutive tails that appear in real coin tossing (and that are predicted by the mathematics of probability) seem surprising to us. Because we don’t expect to see long runs, we may conclude that the coin tosses are not

independent or that some influence is disturbing the random behavior of the coin.

## Example

### 4.31 The “hot hand” in basketball.

Belief in the law of small numbers influences behavior. If a basketball player makes several consecutive shots, both the fans and her teammates believe that she has a “hot hand” and is more likely to make the next shot. This is doubtful.

Careful study suggests that runs of baskets made or missed are no more frequent in basketball than would be expected if each shot were independent of the player’s previous shots. Baskets made or missed are just like heads and tails in tossing a coin. (Of course, some players make 30% of their shots in the long run and others make 50%, so a coin-toss model for basketball must allow coins with different probabilities of a head.) Our perception of hot or cold streaks simply shows that we don’t perceive random behavior very well.<sup>17</sup>



*Our intuition doesn’t do a good job of distinguishing random behavior from systematic influences. This is also true when we look at data. We need statistical inference to supplement exploratory analysis of data because probability calculations can help verify that what we see in the data is more than a random pattern.*

### How large is a large number?

The law of large numbers says that the actual mean outcome of many trials gets close to the distribution mean  $\mu$  as more trials are made. It doesn’t say how many trials are needed to guarantee a mean outcome close to  $\mu$ . That depends on the *variability* of the random outcomes. The more variable the outcomes, the more trials are needed to ensure that the mean outcome  $\bar{x}$  is close to the distribution mean  $\mu$ . Casinos understand this: the outcomes of games of chance are variable enough to hold the interest of gamblers. Only the casino plays often enough to rely on the law of large numbers. Gamblers get entertainment; the casino has a business.

---

## BEYOND THE BASICS

### More laws of large numbers

The law of large numbers is one of the central facts about probability. It helps us understand the mean  $\mu$  of a random variable. It explains why gambling casinos and insurance companies make money. It assures us that statistical estimation will be accurate if we can afford enough observations. The basic law of large numbers applies to independent observations that all have the same distribution. Mathematicians have extended the law to many more general settings. Here are two of these.

**Is there a winning system for gambling?** Serious gamblers often follow a system of betting in which the amount bet on each play depends on the outcome of previous plays. You might, for example, double your bet on each spin of the roulette wheel until you win—or, of course, until your fortune is exhausted. Such a system tries to take advantage of the fact that you have a memory even though the roulette wheel does not. Can you beat the odds with a system based on the outcomes of past plays? No. Mathematicians have established a stronger version of the law of large numbers that says that, if you do not have an infinite fortune to gamble with, your long-run average winnings  $\mu$  remain the same as long as successive trials of the game (such as spins of the roulette wheel) are independent.

**What if observations are not independent?** You are in charge of a process that manufactures video screens for computer monitors. Your equipment measures the tension on the metal mesh that lies behind each screen and is critical to its image quality. You want to estimate the mean tension  $\mu$  for the process by the average  $\bar{x}$  of the measurements. Alas, the tension measurements are not independent. If the tension on one screen is a bit too high, the tension on the next is more likely to also be high. Many real-world processes are like this—the process stays stable in the long run, but two observations made close together are likely to both be above or both be below the long-run mean. Again the mathematicians come to the rescue: as long as the dependence dies out fast enough as we take measurements farther and farther apart in time, the law of large numbers still holds.

---

### Rules for means

You are studying flaws in the painted finish of refrigerators made by your firm. Dimples and paint sags are two kinds of surface flaw. Not all refrigerators have the same number of dimples: many have none, some have one, some two, and so on.

You ask for the average number of imperfections on a refrigerator. The inspectors report finding an average of 0.7 dimples and 1.4 sags per refrigerator. How many total imperfections of both kinds (on the average) are there on a refrigerator? That's easy: if the average number of dimples is 0.7 and the average number of sags is 1.4, then counting both gives an average of  $0.7 + 1.4 = 2.1$  flaws.

In more formal language, the number of dimples on a refrigerator is a random variable  $X$  that varies as we inspect one refrigerator after another. We know only that the mean number of dimples is  $\mu_X = 0.7$ . The number of paint sags is a second random variable  $Y$  having mean  $\mu_Y = 1.4$ . (As usual, the subscripts keep straight which variable we are talking about.) The total number of both dimples and sags is another random variable, the sum  $X + Y$ . Its mean  $\mu_{X+Y}$  is the average number of dimples and sags together. It is just the sum of the individual means  $\mu_X$  and  $\mu_Y$ . That's an important rule for how means of random variables behave.

Here's another rule. The crickets living in a field have mean length 1.2 inches. What is the mean in centimeters? There are 2.54 centimeters in an inch, so the length of a cricket in centimeters is 2.54 times its length in inches. If we multiply every observation by 2.54, we also multiply their average by 2.54. The mean in centimeters must be  $2.54 \times 1.2$ , or about 3.05 centimeters. More formally, the length in inches of a cricket chosen at random from the field is a random variable  $X$  with mean  $\mu_X$ . The length in centimeters is  $2.54X$ , and this new random variable has mean  $2.54\mu_X$ .

The point of these examples is that means behave like averages. Here are the rules we need.

## RULES FOR MEANS OF LINEAR TRANSFORMATIONS, SUMS, AND DIFFERENCES

**Rule 1.** If  $X$  is a random variable and  $a$  and  $b$  are fixed numbers, then

$$\mu_{a + bX} = a + b\mu_X$$

**Rule 2.** If  $X$  and  $Y$  are random variables, then

$$\mu_{X + Y} = \mu_X + \mu_Y$$

**Rule 3.** If  $X$  and  $Y$  are random variables, then

$$\mu_{X - Y} = \mu_X - \mu_Y$$



linear transformation, p. 45

Note that  $a + bX$  is a linear transformation of the random variable  $X$ .

## Example

### 4.32 How many courses?



In Exercise 4.49 (page 261) you described the probability distribution of the number of courses taken in the fall by students at a small liberal arts college. Here is the distribution:

Courses in the fall	1	2	3	4	5	6
Probability	0.05	0.05	0.13	0.26	0.36	0.15

For the spring semester, the distribution is a little different.

Courses in the spring	1	2	3	4	5	6
Probability	0.06	0.08	0.15	0.25	0.34	0.12

For a randomly selected student, let  $X$  be the number of courses taken in the fall semester, and let  $Y$  be the number of courses taken in the spring semester. The means of these random variables are

$$\begin{aligned}\mu_X &= (1)(0.05) + (2)(0.05) + (3)(0.13) + (4)(0.26) + (5)(0.36) + (6)(0.15) \\ &= 4.28\end{aligned}$$

$$\begin{aligned}\mu_Y &= (1)(0.06) + (2)(0.08) + (3)(0.15) + (4)(0.25) + (5)(0.34) + (6)(0.12) \\ &= 4.09\end{aligned}$$

The mean course load for the fall is 4.28 courses and the mean course load for the spring is 4.09 courses. We assume that these distributions apply to students

who earned credit for courses taken in the fall and the spring semesters. The mean of the total number of courses taken for the academic year is  $X + Y$ . Using Rule 2, we calculate the mean of the total number of courses:

$$\begin{aligned}\mu_Z &= \mu_X + \mu_Y \\ &= 4.28 + 4.09 = 8.37\end{aligned}$$

Note that it is not possible for a student to take 8.37 courses in an academic year. This number is the mean of the probability distribution.

## Example

### 4.33 What about credit hours?

In the previous exercise, we examined the number of courses taken in the fall and in the spring at a small liberal arts college. Suppose that we were interested in the total number of credit hours earned for the academic year. We assume that for each course taken at this college, three credit hours are earned. Let  $T$  be the mean of the distribution of the total number of credit hours earned for the academic year. What is the mean of the distribution of  $T$ ? To find the answer, we can use Rule 1 with  $a = 0$  and  $b = 3$ . Here is the calculation:

$$\begin{aligned}\mu_T &= \mu_a + b\mu_Z \\ &= a + b\mu_Z \\ &= 0 + (3)(8.37) = 25.11\end{aligned}$$

The mean of the distribution of the total number of credit hours earned is 25.11.

## USE YOUR KNOWLEDGE

### 4.69 Find $\mu_Y$ .

The random variable  $X$  has mean  $\mu_X = 8$ . If  $Y = 12 + 7X$ , what is  $\mu_Y$ ?

#### 4.70 Find $\mu_W$ .

The random variable  $U$  has mean  $\mu_U = 22$ , and the random variable  $V$  has mean  $\mu_V = 22$ . If  $W + 0.5U + 0.5V$ , find  $\mu_W$ .

### The variance of a random variable

The mean is a measure of the center of a distribution. A basic numerical description requires in addition a measure of the spread or variability of the distribution. The variance and the standard deviation are the measures of spread that accompany the choice of the mean to measure center. Just as for the mean, we need a distinct symbol to distinguish the variance of a random variable from the variance  $s^2$  of a data set. We write the variance of a random variable  $X$  as  $\sigma_X^2$ . Once again the subscript reminds us which variable we have in mind. The definition of the variance  $\sigma_X^2$  of a random variable is similar to the definition of the sample variance  $s^2$  given in Chapter 1. That is, the variance is an average value of the squared deviation  $(X - \mu_X)^2$  of the variable  $X$  from its mean  $\mu_X$ . As for the mean, the average we use is a weighted average in which each outcome is weighted by its probability in order to take account of outcomes that are not equally likely. Calculating this weighted average is straightforward for discrete random variables but requires advanced mathematics in the continuous case. Here is the definition.

#### VARIANCE OF A DISCRETE RANDOM VARIABLE

Suppose that  $X$  is a **discrete random variable** whose distribution is

Value of $X$	$x_1$	$x_2$	$x_3$	...	$x_k$
Probability	$p_1$	$p_2$	$p_3$	...	$p_k$

and that  $\mu_X$  is the mean of  $X$ . The **variance** of  $X$  is

$$\begin{aligned}\sigma_X^2 &= (x_1 - \mu_X)^2 p_1 + (x_2 - \mu_X)^2 p_2 + \dots + (x_k - \mu_X)^2 p_k \\ &= \sum (x_i - \mu_X)^2 p_i\end{aligned}$$

The **standard deviation**  $\sigma_X$  of  $X$  is the square root of the variance.

## Example

### 4.34 Find the mean and the variance.

In Example 4.32 we saw that the distribution of the number  $X$  of fall courses taken by students at a small liberal arts college is

Courses in the fall	1	2	3	4	5	6
Probability	0.05	0.05	0.13	0.26	0.36	0.15

We can find the mean and variance of  $X$  by arranging the calculation in the form of a table. Both  $\mu_X$  and  $\sigma_X^2$  are sums of columns in this table.

$x_i$	$p_i$	$x_i p_i$	$(x_i - \mu_x)^2 p_i$
1	0.05	0.05	$(1 - 4.28)^2(0.05) = 0.53792$
2	0.05	0.10	$(2 - 4.28)^2(0.05) = 0.25992$
3	0.13	0.39	$(3 - 4.28)^2(0.13) = 0.21299$
4	0.26	1.04	$(4 - 4.28)^2(0.26) = 0.02038$
5	0.36	1.80	$(5 - 4.28)^2(0.36) = 0.18662$
6	0.15	0.90	$(6 - 4.28)^2(0.15) = 0.44376$
		$\mu_X = 4.28$	$\sigma_X^2 = 1.662$

We see that  $\sigma_X^2 = 1.662$ . The standard deviation of  $X$  is  $\sigma_X = 1.662 = 1.289$ . The standard deviation is a measure of the variability of the number of fall courses taken by the students at the small liberal arts college. As in the case of distributions for data, the standard deviation of a probability distribution is easiest to understand for Normal distributions.

## USE YOUR KNOWLEDGE

### 4.71 Find the variance and the standard deviation.

The random variable  $X$  has the following probability distribution:

Value of $X$	\$0	3
Probability	0.4	0.6

Find the variance  $\sigma_X^2$  and the standard deviation  $\sigma_X$  for this random variable.

## Rules for variances and standard deviations



What are the facts for variances that parallel Rules 1, 2, and 3 for means? *The mean of a sum of random variables is always the sum of their means, but this addition rule is true for variances only in special situations.* To understand why, take  $X$  to be the percent of a family's after-tax income that is spent, and take  $Y$  to be the percent that is saved. When  $X$  increases,  $Y$  decreases by the same amount. Though  $X$  and  $Y$  may vary widely from year to year, their sum  $X + Y$  is always 100% and does not vary at all. It is the association between the variables  $X$  and  $Y$  that prevents their variances from adding.

If random variables are independent, this kind of association between their values is ruled out and their variances do add. Two random variables  $X$  and  $Y$  are **independent** if knowing that any event involving  $X$  alone did or did not occur tells us nothing about the occurrence of any event involving  $Y$  alone.

### independence

Probability models often assume independence when the random variables describe outcomes that appear unrelated to each other. You should ask in each instance whether the assumption of independence seems reasonable.

When random variables are not independent, the variance of their sum depends on the **correlation** between them as well as on their individual variances. In Chapter 2, we met the correlation  $r$  between two observed variables measured on the same individuals. We defined (page 104) the correlation  $r$  as an average of the products of the standardized  $x$  and  $y$  observations. The correlation between two random variables is defined in the same way, once again using a weighted average with probabilities as weights. We won't give the details—it is enough to know that the correlation between two random variables has the same basic properties as the correlation  $r$  calculated from data. We use  $\rho$ , the Greek letter rho, for the correlation between two random variables. The correlation  $\rho$  is a number between  $-1$  and  $1$  that measures the direction and strength of the linear relationship between two variables. **The correlation between two independent random variables is zero.**

### correlation

Returning to family finances, if  $X$  is the percent of a family's after-tax income that is spent and  $Y$  is the percent that is saved, then  $Y = 100 - X$ . This is a perfect linear relationship with a negative slope, so the correlation between  $X$  and  $Y$  is  $\rho = -1$ . With the correlation at hand, we can state the rules for manipulating variances.

## RULES FOR VARIANCES AND STANDARD DEVIATIONS OF LINEAR TRANSFORMATIONS, SUMS, AND DIFFERENCES

**Rule 1.** If  $X$  is a random variable and  $a$  and  $b$  are fixed numbers, then

$$\sigma_{a+bX}^2 = b^2\sigma_X^2$$

**Rule 2.** If  $X$  and  $Y$  are independent random variables, then

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$$

$$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2$$

This is the **addition rule for variances of independent random variables**.

**Rule 3.** If  $X$  and  $Y$  have correlation  $\rho$ , then

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y$$

$$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y$$

This is the **general addition rule for variances of random variables**.

To find the standard deviation, take the square root of the variance.



*Because a variance is the average of squared deviations from the mean, multiplying  $X$  by a constant  $b$  multiplies  $\sigma_X^2$  by the square of the constant. Adding a constant  $a$  to a random variable changes its mean but does not change its variability. The variance of  $X + a$  is therefore the same as the variance of  $X$ . Because the square of  $-1$  is  $1$ , the addition rule says that the variance of a difference between independent random variables is the *sum* of the variances. For independent random variables, the difference  $X - Y$  is more variable than either  $X$  or  $Y$  alone because variations in both  $X$  and  $Y$  contribute to variation in their difference.*



As with data, we prefer the standard deviation to the variance as a measure of the variability of a random variable. *Rule 2 for variances implies that standard deviations of independent random variables do not add. To combine standard deviations, use the rules for variances.* For example, the standard deviations of  $2X$  and  $-2X$  are both equal to  $2\sigma_X$ , because this is the square root of the variance

$4\sigma_X^2$ .

## Example

### 4.35 Payoff in the Tri-State Pick 3 lottery.

The payoff  $X$  of a \$1 ticket in the Tri-State Pick 3 game is \$500 with probability  $1/1000$  and 0 the rest of the time. Here is the combined calculation of mean and variance:

$x_i$	$p_i$	$x_i p_i$	$(x_i - \mu_X)^2 p_i$
0	0.999	0	$(0 - 0.05)^2(0.999) = 0.24975$
500	0.001	0.5	$(500 - 0.05)^2(0.001) = 249.50025$
		$\mu_X = 0.5$	$\sigma_X^2 = 249.75$

The mean payoff is 50 cents. The standard deviation is  $\sigma_X = 249.75 = \$15.80$ . It is usual for games of chance to have large standard deviations because large variability makes gambling exciting.

If you buy a Pick 3 ticket, your winnings are  $W = X - 1$  because the dollar you paid for the ticket must be subtracted from the payoff. Let's find the mean and variance for this random variable.

## Example

### 4.36 Winnings in the Tri-State Pick 3 lottery.

By the rules for means, the mean amount you win is

$$\mu_W = \mu_X - 1 = -\$0.50$$

That is, you lose an average of 50 cents on a ticket. The rules for variances remind us that the variance and standard deviation of the winnings  $W = X - 1$  are the same as those of  $X$ . Subtracting a fixed number changes the mean but not the variance.

Suppose now that you buy a \$1 ticket on each of two different days. The payoffs  $X$  and  $Y$  on the two tickets are independent because separate drawings are held each day. Your total payoff is  $X + Y$ . Let's find the mean and standard deviation for this payoff.

## Example

### 4.37 Two tickets.

The mean for the payoff for the two tickets is

$$\mu_{X+Y} = \mu_X + \mu_Y = \$0.50 + \$0.50 = \$1.00$$

Because  $X$  and  $Y$  are independent, the variance of  $X + Y$  is

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 = 249.75 + 249.75 = 499.5$$

The standard deviation of the total payoff is

$$\sigma_{X+Y} = \sqrt{499.5} = \$22.35$$

This is not the same as the sum of the individual standard deviations, which is  $\$15.80 + \$15.80 = \$31.60$ . Variances of independent random variables add; standard deviations do not.

When we add random variables that are correlated, we need to use the correlation for the calculation of the variance, but not for the calculation of the mean. Here is an example.

## Example

### 4.38 Utility bills.

Consider a household where the monthly bill for natural gas averages \$125 with a standard deviation of \$75, while the monthly bill for electricity averages \$174 with a standard deviation of \$41. The correlation between the two bills is  $-0.55$ .

Let's compute the mean and standard deviation of the sum of the natural-gas bill and the electricity bill. We let  $X$  stand for the natural-gas bill and  $Y$  stand for the electricity bill. Then the total is  $X + Y$ . Using the rules for means, we have

$$\mu_{X+Y} = \mu_X + \mu_Y = 125 + 174 = 299$$

To find the standard deviation we first find the variance and then take the square root to determine the standard deviation. From the general addition rule for variances of random variables,

$$\begin{aligned}\sigma_{X+Y}^2 &= \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y \\ &= (75)^2 + (41)^2 + (2)(-0.55)(75)(41) \\ &= 3923.5\end{aligned}$$

Therefore, the standard deviation is

$$\sigma_{X+Y} = \sqrt{3923.5} = 63$$

The total of the natural-gas bill and the electricity bill has mean \$299 and standard deviation \$63.

The negative correlation in Example 4.38 is due to the fact that, in this household, natural gas is used for heating and electricity is used for air-conditioning. So, when it is warm, the electricity charges are high and the natural-gas charges are low. When it is cool, the reverse is true. This causes the standard deviation of the sum to be less than it would be if the two bills were uncorrelated (see Exercise 4.83, on page 281).

There are situations where we need to combine several of our rules to find means and standard deviations. Here is an example.

## Example

### 4.39 Calcium intake.

To get enough calcium for optimal bone health, tablets containing calcium are often recommended to supplement the calcium in the diet. One study designed to evaluate the effectiveness of a supplement followed a group of young people for seven years. Each subject was assigned to take either a tablet containing 1000 milligrams of calcium per day (mg/d) or a placebo tablet that

was identical except that it had no calcium.<sup>18</sup> A major problem with studies like this one is compliance: subjects do not always take the treatments assigned to them.

In this study, the compliance rate declined to about 47% toward the end of the seven-year period. The standard deviation of compliance was 22%. Calcium from the diet averaged 850 mg/d with a standard deviation of 330 mg/d. The correlation between compliance and dietary intake was 0.68. Let's find the mean and standard deviation for the total calcium intake. We let  $S$  stand for the intake from the supplement and  $D$  stand for the intake from the diet.

We start with the intake from the supplement. Since the compliance is 47% and the amount in each tablet is 1000 mg, the mean for  $S$  is

$$\mu_S = 1000(0.47) = 470$$

Since the standard deviation of the compliance is 22%, the variance of  $S$  is

$$\sigma_S^2 = 1000^2(0.22)^2 = 48,400$$

The standard deviation is

$$\sigma_S = \sqrt{48,400} = 220$$

Be sure to verify which rules for means and variances are used in these calculations.

We can now find the mean and standard deviation for the total intake. The mean is

$$\mu_{S+D} = \mu_S + \mu_D = 470 + 850 = 1320$$

and the variance is

$$\sigma_{S+D}^2 = \sigma_S^2 + \sigma_D^2 + 2\rho\sigma_S\sigma_D = (220)^2 + (330)^2 + 2(0.68)(220)(330) = 256,036$$

and the standard deviation is

$$\sigma_{S+D} = \sqrt{256,036} = 506$$

The mean of the total calcium intake is 1320 mg/d and the standard deviation is 506 mg/d.

The correlation in this example illustrates an unfortunate fact about compliance and having an adequate diet. Some of the subjects in this study have diets that provide an adequate amount of calcium while others do not. The positive correlation between compliance and dietary intake tells us that those who have relatively high dietary intakes are more likely to take the assigned supplements. On the other hand, those subjects with relatively low dietary intakes, the ones who need the supplement the most, are less likely to take the assigned supplements.

## Section 4.4 Summary

The probability distribution of a random variable  $X$ , like a distribution of data, has a **mean**  $\mu_X$  and a **standard deviation**  $\sigma_X$ .

The **law of large numbers** says that the average of the values of  $X$  observed in many trials must approach  $\mu$ .

The **mean**  $\mu$  is the balance point of the probability histogram or density curve. If  $X$  is **discrete** with possible values  $x_i$  having probabilities  $p_i$ , the mean is the average of the values of  $X$ , each weighted by its probability:

$$\mu_X = x_1 p_1 + x_2 p_2 + \dots + x_k p_k$$

The **variance**  $\sigma_X^2$  is the average squared deviation of the values of the variable from their mean. For a discrete random variable,

$$\sigma_X^2 = (x_1 - \mu_X)^2 p_1 + (x_2 - \mu_X)^2 p_2 + \dots + (x_k - \mu_X)^2 p_k$$

The **standard deviation**  $\rho_X$  is the square root of the variance. The standard deviation measures the variability of the distribution about the mean. It is easiest to interpret for Normal distributions.

The **mean and variance of a continuous random variable** can be computed from the density curve, but to do so requires more advanced mathematics.

The means and variances of random variables obey the following rules. If  $a$  and  $b$  are fixed numbers, then

$$\mu_{a + bX} = a + b\mu_X$$

$$\sigma_{a + bX}^2 = b^2 \sigma_X^2$$

If  $X$  and  $Y$  are any two random variables having correlation  $\rho$  then

$$\mu_{X + Y} = \mu_X + \mu_Y$$

$$\mu_{X - Y} = \mu_X - \mu_Y$$

$$\sigma_{X + Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y$$

$$\sigma_{X - Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y$$

If  $X$  and  $Y$  are **independent**, then  $\rho = 0$ . In this case,

$$\sigma_{X + Y}^2 = \sigma_X^2 + \sigma_Y^2$$

$$\sigma_{X - Y}^2 = \sigma_X^2 + \sigma_Y^2$$

To find the standard deviation, take the square root of the variance.

## SECTION 4.4 Exercises

For Exercise 4.67, see page 265; for Exercise 4.68, see page 269; for Exercises 4.69 and 4.70, see page

273; and for Exercise 4.71, see page 274.

#### 4.72 Find the mean of the random variable.

A random variable  $X$  has the following distribution.

$X$	-1	0	1	2
Probability	0.3	0.2	0.2	0.3

Find the mean for this random variable. Show your work.

#### 4.73 Explain what happens when the sample size gets large.

Consider the following scenarios: (1) You take a sample of two observations on a random variable and compute the sample mean, (2) you take a sample of 100 observations on the same random variable and compute the sample mean, (3) you take a sample of 1000 observations on the same random variable and compute the sample mean. Explain in simple language how close you expect the sample mean to be to the mean of the random variable as you move from Scenario 1 to Scenario 2 to Scenario 3.

#### 4.74 Find some means.

Suppose that  $X$  is a random variable with mean 20 and standard deviation 5. Also suppose that  $Y$  is a random variable with mean 40 and standard deviation 10. Find the mean of the random variable  $Z$  for each of the following cases. Be sure to show your work.

(a)  $Z = 2 + 10X$ .

(b)  $Z = 10X - 2$ .

(c)  $Z = X + Y$ .

(d)  $Z = X - Y$ .

(e)  $Z = -3X - 2Y$ .

#### 4.75 Find the variance and the standard deviation.

A random variable  $X$  has the following distribution.

$X$	-1	0	1	2
Probability	0.3	0.2	0.2	0.3

Find the variance and the standard deviation for this random variable. Show your work.

#### 4.76 Find some variances and standard deviations.

Suppose that  $X$  is a random variable with mean 20 and standard deviation 5. Also suppose that  $Y$  is a random variable with mean 40 and standard deviation 10. Find the variance and standard deviation of the random variable  $Z$  for each of the following cases. Be sure to show your work.

- (a)  $Z = 2 + 10X$ .
- (b)  $Z = 10X - 2$ .
- (c)  $Z = X + Y$ .
- (d)  $Z = X - Y$ .
- (e)  $Z = -3X - 2Y$ .

#### 4.77 What happens if the correlation is not zero?

Suppose that  $X$  is a random variable with mean 20 and standard deviation 5. Also suppose that  $Y$  is a random variable with mean 40 and standard deviation 10. Assume that the correlation between  $X$  and  $Y$  is 0.5. Find the mean of the random variable  $Z$  for each of the following cases. Be sure to show your work.

- (a)  $Z = 2 + 10X$ .
- (b)  $Z = 10X - 2$ .
- (c)  $Z = X + Y$ .
- (d)  $Z = X - Y$ .
- (e)  $Z = -3X - 2Y$ .

#### 4.78 What's wrong?

In each of the following scenarios, there is something wrong. Describe what is wrong and give a reason for your answer.

- (a) If you toss a fair coin three times and get heads all three times, then the probability of getting a tail on the next toss is much greater than one-half.
- (b) If you multiply a random variable by 10, then the mean is multiplied by 10 and the variance is multiplied by 10.
- (c) When finding the mean of the sum of two random variables, you need to know the correlation between them.

#### 4.79 Servings of fruits and vegetables.

The following table gives the distribution of the number of servings of fruits and vegetables consumed per day in a population.

Number of servings $X$	0	1	2	3	4	5
Probability	0.3	0.1	0.1	0.2	0.2	0.1

Find the mean for this random variable.

#### 4.80 Mean of the distribution for the number of aces.

In Exercise 4.58 (page 262) you examined the probability distribution for the number of aces when you are

dealt two cards in the game of Texas hold 'em. Let  $X$  represent the number of aces in a randomly selected deal of two cards in this game. Here is the probability distribution for the random variable  $X$ :

Value of $X$	0	1	2
Probability	0.8507	0.1448	0.0045

Find  $\mu_X$ , the mean of the probability distribution of  $X$ .

#### 4.81 Standard deviation of the number of aces.

Refer to Exercise 4.80. Find the standard deviation of the number of aces.

#### 4.82 Standard deviation for fruits and vegetables.

Refer to Exercise 4.79. Find the variance and the standard deviation for the distribution of the number of servings of fruits and vegetables.

#### 4.83 Suppose that the correlation is zero.

Refer to Example 4.38 (page 277).

- Recompute the standard deviation for the total of the natural-gas bill and the electricity bill assuming that the correlation is zero.
- Is this standard deviation larger or smaller than the standard deviation computed in Example 4.38? Explain why.

#### 4.84 Find the mean of the sum.

Figure 4.12 (page 263) displays the density curve of the sum  $Y = X_1 + X_2$  of two independent random numbers, each uniformly distributed between 0 and 1.

- The mean of a continuous random variable is the balance point of its density curve. Use this fact to find the mean of  $Y$  from Figure 4.12.
- Use the same fact to find the means of  $X_1$  and  $X_2$ . (They have the density curve pictured in Figure 4.9, page 258.) Verify that the mean of  $Y$  is the sum of the mean of  $X_1$  and the mean of  $X_2$ .

#### 4.85 Calcium supplements and calcium in the diet.

Refer to Example 4.39 (page 278). Suppose that people who have high intakes of calcium in their diets are more compliant than those who have low intakes. What effect would this have on the calculation of the standard deviation for the total calcium intake? Explain your answer.



#### 4.86 Toss a four-sided die twice.

Role-playing games like Dungeons & Dragons use many different types of dice. Suppose that a four-sided die has faces marked 1, 2, 3, and 4. The intelligence of a character is determined by rolling this die twice and adding 1 to the sum of the spots. The faces are equally likely and the two rolls are independent. What

is the average (mean) intelligence for such characters? How spread out are their intelligences, as measured by the standard deviation of the distribution?

#### 4.87 Means and variances of sums.

The rules for means and variances allow you to find the mean and variance of a sum of random variables without first finding the distribution of the sum, which is usually much harder to do.

- (a) A single toss of a balanced coin has either 0 or 1 head, each with probability  $1/2$ . What are the mean and standard deviation of the number of heads?
- (b) Toss a coin four times. Use the rules for means and variances to find the mean and standard deviation of the total number of heads.
- (c) Example 4.23 (page 255) finds the distribution of the number of heads in four tosses. Find the mean and standard deviation from this distribution. Your results in parts (b) and (c) should agree.

#### 4.88 What happens when the correlation is 1?

We know that variances add if the random variables involved are uncorrelated ( $\rho = 0$ ), but not otherwise. The opposite extreme is perfect positive correlation ( $\rho = 1$ ). Show by using the general addition rule for variances that in this case the standard deviations add. That is,  $\sigma_{X+Y} = \sigma_X + \sigma_Y$  if  $\rho_{XY} = 1$ .

#### 4.89 Will you assume independence?

In which of the following games of chance would you be willing to assume independence of  $X$  and  $Y$  in making a probability model? Explain your answer in each case.

- (a) In blackjack, you are dealt two cards and examine the total points  $X$  on the cards (face cards count 10 points). You can choose to be dealt another card and compete based on the total points  $Y$  on all three cards.
- (b) In craps, the betting is based on successive rolls of two dice.  $X$  is the sum of the faces on the first roll, and  $Y$  the sum of the faces on the next roll.

#### 4.90 Transform the distribution of heights from centimeters to inches.

A report of the National Center for Health Statistics says that the heights of 20-year-old men have mean 176.8 centimeters (cm) and standard deviation 7.2 cm. There are 2.54 centimeters in an inch. What are the mean and standard deviation in inches?

#### ***Insurance.***

*The business of selling insurance is based on probability and the law of large numbers. Consumers buy insurance because we all face risks that are unlikely but carry high cost. Think of a fire destroying your home. So we form a group to share the risk: we all pay a small amount, and the insurance policy pays a large amount to those few of us whose homes burn down. The insurance company sells many policies, so it can rely on the law of large numbers. Exercises 4.91 to 4.94 explore aspects of insurance.*

### 4.91 Fire insurance.

An insurance company looks at the records for millions of homeowners and sees that the mean loss from fire in a year is  $\mu = \$300$  per person. (Most of us have no loss, but a few lose their homes. The \$300 is the average loss.) The company plans to sell fire insurance for \$300 plus enough to cover its costs and profit. Explain clearly why it would be stupid to sell only 10 policies. Then explain why selling thousands of such policies is a safe business.

### 4.92 Mean and standard deviation for 10 and for 12 policies.

In fact, the insurance company sees that in the entire population of homeowners, the mean loss from fire is  $\mu = \$300$  and the standard deviation of the loss is  $\sigma = \$400$ . What are the mean and standard deviation of the average loss for 10 policies? (Losses on separate policies are independent.) What are the mean and standard deviation of the average loss for 12 policies?

### 4.93 Life insurance.

Assume that a 25-year-old man has these probabilities of dying during the next five years:

Age at death	25	26	27	28	29
Probability	0.00039	0.00044	0.00051	0.00057	0.00060

(a) What is the probability that the man does not die in the next five years?

(b) An online insurance site offers a term insurance policy that will pay \$100,000 if a 25-year-old man dies within the next five years. The cost is \$175 per year. So the insurance company will take in \$875 from this policy if the man does not die within five years. If he does die, the company must pay \$100,000. Its loss depends on how many premiums the man paid, as follows:

Age at death	25	26	27	28	29
Probability	\$99,825	\$99,650	\$99,475	\$99,300	\$99,125

What is the insurance company's mean cash intake from such policies?

### 4.94 Risk for one versus thousands of life insurance policies.

It would be quite risky for you to insure the life of a 25-year-old friend under the terms of Exercise 4.93. There is a high probability that your friend would live and you would gain \$875 in premiums. But if he were to die, you would lose almost \$100,000. Explain carefully why selling insurance is not risky for an insurance company that insures many thousands of 25-year-old men.

## 4.5 General Probability Rules

When you complete this section, you will be able to

- Apply the five rules of probability.
- Apply the general addition rule for unions of two or more events.
- Find conditional probabilities.
- Apply the multiplication rule.
- Use a tree diagram to find probabilities.
- Use Bayes's rule to find probabilities.
- Determine whether or not two events that both have positive probability are independent.

Our study of probability has concentrated on random variables and their distributions. Now we return to the laws that govern any assignment of probabilities. The purpose of learning more laws of probability is to be able to give probability models for more complex random phenomena. We have already met and used five rules.

### PROBABILITY RULES

**Rule 1.**  $0 \leq P(A) \leq 1$  for any event  $A$

**Rule 2.**  $P(S) = 1$

**Rule 3. Addition rule:** If  $A$  and  $B$  are **disjoint** events, then

$$P(A \text{ or } B) = P(A) + P(B)$$

**Rule 4. Complement rule:** For any event  $A$

$$P(A^c) = 1 - P(A)$$

**Rule 5. Multiplication rule:** If  $A$  and  $B$  are **independent** events, then

$$P(A \text{ and } B) = P(A) P(B)$$

### General addition rules

Probability has the property that if  $A$  and  $B$  are disjoint events, then  $P(A \text{ or } B) = P(A) + P(B)$ . What if there are more than two events, or if the events are not disjoint? These circumstances are covered by more general addition rules for probability.

## UNION

The **union** of any collection of events is the event that at least one of the collection occurs.

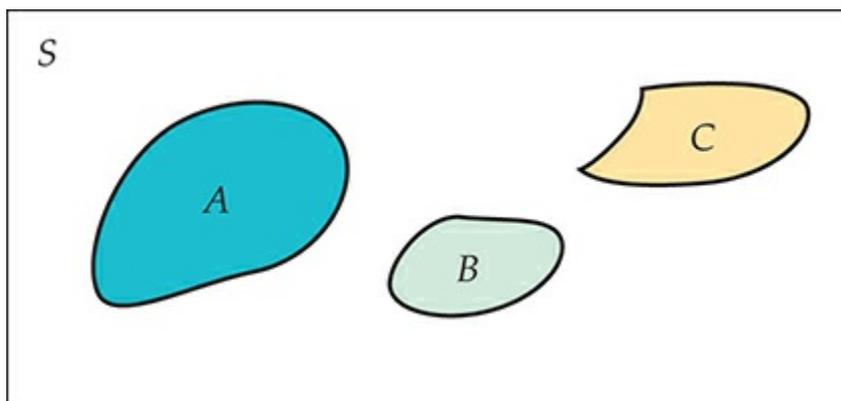
For two events  $A$  and  $B$  the union is the event  $\{A \text{ or } B\}$  that  $A$  or  $B$  or both occur. From the addition rule for two disjoint events we can obtain rules for more general unions. Suppose first that we have several events—say  $A$ ,  $B$  and  $C$ —that are disjoint in pairs. That is, no two can occur simultaneously. The Venn diagram in Figure 4.15 illustrates three disjoint events. The addition rule for two disjoint events extends to the following law.

## ADDITION RULE FOR DISJOINT EVENTS

If events  $A$ ,  $B$  and  $C$  are disjoint in the sense that no two have any outcomes in common, then

$$P(\text{one or more of } A, B, C) = P(A) + P(B) + P(C)$$

This rule extends to any number of disjoint events.



**FIGURE 4.15**

The addition rule for disjoint events:  $P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C)$  when events  $A$ ,  $B$ , and  $C$  are disjoint.

## Example

### 4.40 Probabilities as areas.

Generate a random number  $X$  between 0 and 1. What is the probability that the first digit after the decimal point will be odd? The random number  $X$  is a continuous random variable whose density curve has constant height 1 between 0 and 1 and is 0 elsewhere. The event that the first digit of  $X$  is odd is the union of five disjoint events. These events are

$$0.10 \leq X < 0.20$$

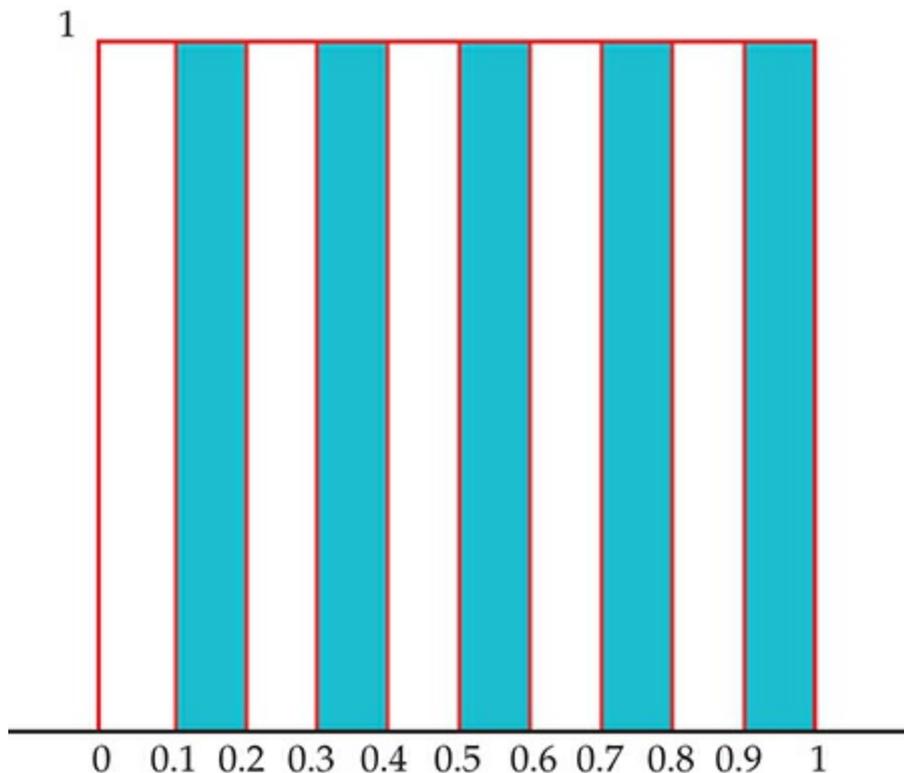
$$0.30 \leq X < 0.40$$

$$0.50 \leq X < 0.60$$

$$0.70 \leq X < 0.80$$

$$0.90 \leq X < 1.00$$

Figure 4.16 illustrates the probabilities of these events as areas under the density curve. Each area is 0.1. The union of the five therefore has probability equal to the sum, or 0.5. As we should expect, a random number is equally likely to begin with an odd or an even digit.



### FIGURE 4.16

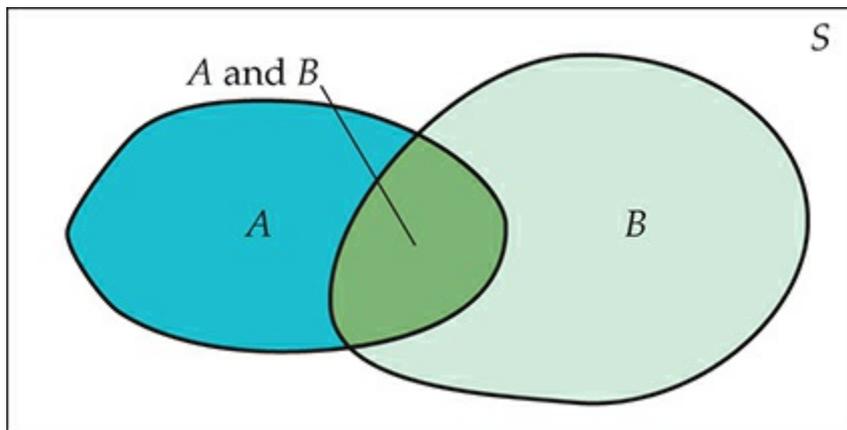
The probability that the first digit after the decimal point of a random number is odd is the sum of the probabilities of the 5 disjoint events shown. See Example 4.40.

## USE YOUR KNOWLEDGE

### 4.95 Probability that you roll a 2 or a 4 or a 5.

If you roll a die, the probability of each of the six possible outcomes (1, 2, 3, 4, 5, 6) is  $1/6$ . What is the probability that you roll a 2 or a 4 or a 5?

If events  $A$  and  $B$  are not disjoint, they can occur simultaneously. The probability of their union is then *less* than the sum of their probabilities. As Figure 4.17 suggests, the outcomes common to both are counted twice when we add probabilities, so we must subtract this probability once. Here is the addition rule for the union of any two events, disjoint or not.



### FIGURE 4.17

The union of two events that are not disjoint. The general addition rule says that  $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ .

## GENERAL ADDITION RULE FOR UNIONS OF TWO EVENTS

For any two events  $A$  and  $B$ ,

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

If  $A$  and  $B$  are disjoint, the event  $\{A \text{ or } B\}$  that both occur has no outcomes in it.

This *empty event* is the complement of the sample space  $S$  and must have probability 0. So the general addition rule includes Rule 3, the addition rule for disjoint events.

## Example

### 4.41 Adequate sleep and exercise.



Suppose that 40% of adults get enough sleep and 46% exercise regularly. What is the probability that an adult gets enough sleep or exercises regularly? To find this probability, we also need to know the percent who get enough sleep and exercise. Let's assume that 24% do both.

We will use the notation of the general addition rule for unions of two events. Let  $A$  be the event that an adult gets enough sleep and let  $B$  be the event that a person exercises regularly. We are given that  $P(A) = 0.40$ ,  $P(B) = 0.46$ , and  $P(A \text{ and } B) = 0.24$ . Therefore,

$$\begin{aligned}P(A \text{ or } B) &= P(A) + P(B) - P(A \text{ and } B) \\&= 0.40 + 0.46 - 0.24 \\&= 0.62\end{aligned}$$

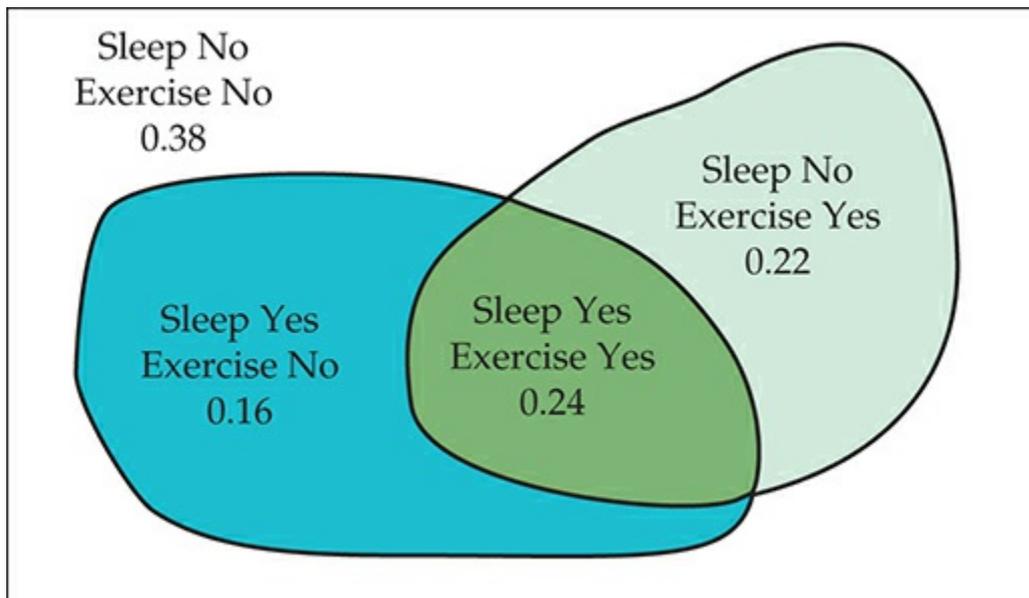
The probability that an adult gets enough sleep or exercises regularly is 0.62, or 62%.

## USE YOUR KNOWLEDGE

#### 4.96 Probability that your roll is odd or greater than 4.

If you roll a die, the probability of each of the six possible outcomes (1, 2, 3, 4, 5, 6) is  $1/6$ . What is the probability that your roll is odd or greater than 4?

Venn diagrams are a great help in finding probabilities for unions because you can just think of adding and subtracting areas. Figure 4.18 shows some events and their probabilities for Example 4.41. What is the probability that an adult gets adequate sleep and does not exercise?



**FIGURE 4.18**

Venn diagram and probabilities for Example 4.41.

The Venn diagram shows that this is the probability that an adult gets adequate sleep minus the probability that an adult gets adequate sleep and exercises regularly,  $0.40 - 0.24 = 0.16$ . Similarly, the probability that an adult does not get adequate sleep and exercises regularly is  $0.46 - 0.24 = 0.22$ . The four probabilities that appear in the figure add to 1 because they refer to four disjoint events whose union is the entire sample space.

### Conditional probability

The probability we assign to an event can change if we know that some other event has occurred. This idea is the key to many applications of probability.

## Example

### 4.42 Probability of being dealt an ace.

Slim is a professional poker player. He stares at the dealer, who prepares to deal. What is the probability that the card dealt to Slim is an ace? There are 52 cards in the deck. Because the deck was carefully shuffled, the next card dealt is equally likely to be any of the cards that Slim has not seen. Four of the 52 cards are aces. So

$$P(\text{ace}) = \frac{4}{52} = \frac{1}{13}$$

This calculation assumes that Slim knows nothing about any cards already dealt. Suppose now that he is looking at 4 cards already in his hand, and that one of them is an ace. He knows nothing about the other 48 cards except that exactly 3 aces are among them. Slim's probability of being dealt an ace *given what he knows* is now

$$P(\text{ace} \mid 1 \text{ ace in 4 visible cards}) = \frac{3}{48} = \frac{1}{16}$$

Knowing that there is 1 ace among the 4 cards Slim can see changes the probability that the next card dealt is an ace.

The new notation  $P(A \mid B)$  is a **conditional probability**. That is, it gives the probability of one event (the next card dealt is an ace) under the condition that we know another event (exactly 1 of the 4 visible cards is an ace). You can read the bar  $\mid$  as “given the information that.”

### conditional probability

## MULTIPLICATION RULE

The probability that both of two events  $A$  and  $B$  happen together can be found by

$$P(A \text{ and } B) = P(A)P(B \mid A)$$

Here  $P(B \mid A)$  is the conditional probability that  $B$  occurs, given the information that  $A$  occurs.

## USE YOUR KNOWLEDGE

#### 4.97 The probability of another ace.

Refer to Example 4.42. Suppose that two of the four cards in Slim's hand are aces. What is the probability that the next card dealt to him is an ace?

### Example

#### 4.43 Downloading music from the Internet.

The multiplication rule is just common sense made formal. For example, suppose that 29% of Internet users download music files, and 67% of downloaders say they don't care if the music is copyrighted. So the percent of Internet users who download music (event  $A$  and don't care about copyright (event  $B$  is 67% of the 29% who download, or

$$(0.67)(0.29) = 0.1943 = 19.43\%$$

The multiplication rule expresses this as

$$\begin{aligned} P(A \text{ and } B) &= P(A) \times P(B | A) \\ &= (0.29)(0.67) = 0.1943 \end{aligned}$$

Here is another example that uses conditional probability.

### Example

#### 4.44 Probability of a favorable draw.

Slim is still at the poker table. At the moment, he wants very much to draw two diamonds in a row. As he sits at the table looking at his hand and at the upturned cards on the table, Slim sees 11 cards. Of these, 4 are diamonds. The full deck contains 13 diamonds among its 52 cards, so 9 of the 41 unseen cards are diamonds. To find Slim's probability of drawing two diamonds, first calculate

$$P(\text{first card diamond})=9/41$$

$$P(\text{second card diamond} \mid \text{first card diamond})=8/40$$

Slim finds both probabilities by counting cards. The probability that the first card drawn is a diamond is  $9/41$  because 9 of the 41 unseen cards are diamonds. If the first card is a diamond, that leaves 8 diamonds among the 40 remaining cards. So the *conditional* probability of another diamond is  $8/40$ . The multiplication rule now says that

$$P(\text{both cards diamonds})=9/41 \times 8/40=0.044$$

Slim will need luck to draw his diamonds.

## USE YOUR KNOWLEDGE

### 4.98 The probability that the next two cards are diamonds.

In the setting of Example 4.42, suppose that Slim sees 23 cards and the only diamonds are the 3 in his hand. What is the probability that the next 2 cards dealt to Slim will be diamonds? This outcome would give him 5 cards from the same suit, a hand that is called a flush.

If  $P(A)$  and  $P(A \text{ and } B)$  are given, we can rearrange the multiplication rule to produce a *definition* of the conditional probability  $P(B \mid A)$  in terms of unconditional probabilities.

## DEFINITION OF CONDITIONAL PROBABILITY

When  $P(A) > 0$ , the **conditional probability** of  $B$  given  $A$  is

$$P(B \mid A)=P(A \text{ and } B)/P(A)$$



*Be sure to keep in mind the distinct roles in  $P(B \mid A)$  of the event  $B$  whose probability we are computing and the event  $A$  that represents the information we*

are given. The conditional probability  $P(B | A)$  makes no sense if the event  $A$  can never occur, so we require that  $P(A) > 0$  whenever we talk about  $P(B | A)$ .

## Example

### 4.45 College students.

Here is the distribution of U.S. college students classified by age and full-time or part-time status:

Age (years)	Full-time	Part-time
15 to 19	0.21	0.02
20 to 24	0.32	0.07
25 to 29	0.10	0.10
30 and over	0.05	0.13

Let's compute the probability that a student is aged 15 to 19, given that the student is full-time. We know that the probability that a student is full-time *and* aged 15 to 19 is 0.21 from the table of probabilities. But what we want here is a conditional probability, given that a student is full-time. Rather than asking about age among all students, we restrict our attention to the subpopulation of students who are full-time. Let

$A$  = the student is between 15 and 19 years of age

$B$  = the student is a full-time student

Our formula is

$$P(A | B) = \frac{P(A \text{ and } B)}{P(B)}$$

We read  $P(A \text{ and } B) = 0.21$  from the table as we mentioned previously. What about  $P(B)$ ? This is the probability that a student is full-time. Notice that there are four groups of students in our table that fit this description. To find the probability needed, we add the entries:

$$P(B) = 0.21 + 0.32 + 0.10 + 0.05 = 0.68$$

We are now ready to complete the calculation of the conditional probability:

$$\begin{aligned} P(A | B) &= \frac{P(A \text{ and } B)}{P(B)} \\ &= \frac{0.21}{0.68} \end{aligned}$$

$$= 0.31$$

The probability that a student is 15 to 19 years of age, given that the student is full-time, is 0.31.

Here is another way to give the information in the last sentence of this example: 31% of full-time college students are 15 to 19 years old. Which way do you prefer?

## USE YOUR KNOWLEDGE

### 4.99 What rule did we use?

In Example 4.45, we calculated  $P(B)$ . What rule did we use for this calculation? Explain why this rule applies in this setting.

### 4.100 Find the conditional probability.

Refer to Example 4.45. What is the probability that a student is part-time, given that the student is 15 to 19 years old? Explain in your own words the difference between this calculation and the one that we did in Example 4.45.

## General multiplication rules

The definition of conditional probability reminds us that in principle all probabilities, including conditional probabilities, can be found from the assignment of probabilities to events that describe random phenomena. More often, however, conditional probabilities are part of the information given to us in a probability model, and the multiplication rule is used to compute  $P(A \text{ and } B)$ . This rule extends to more than two events.

The union of a collection of events is the event that *any* of them occur. Here is the corresponding term for the event that *all* of them occur.

## INTERSECTION

The **intersection** of any collection of events is the event that *all* the events occur.

To extend the multiplication rule to the probability that all of several events occur, the key is to condition each event on the occurrence of *all* the preceding events. For example, the intersection of three events  $A$ ,  $B$  and  $C$  has probability

$$P(A \text{ and } B \text{ and } C) = P(A)P(B | A)P(C | A \text{ and } B)$$

## Example

### 4.46 High school athletes and professional careers.

Only 5% of male high school basketball, baseball, and football players go on to play at the college level. Of these, only 1.7% enter major league professional sports. About 40% of the athletes who compete in college and then reach the pros have a career of more than three years. Define these events:

$$A = \{\text{competes in college}\}$$

$$B = \{\text{competes professionally}\}$$

$$C = \{\text{pro career longer than 3 years}\}$$

What is the probability that a high school athlete competes in college and then goes on to have a pro career of more than three years? We know that

$$P(A) = 0.05$$

$$P(B | A) = 0.017$$

$$P(C | A \text{ and } B) = 0.4$$

The probability we want is therefore

$$\begin{aligned} P(A \text{ and } B \text{ and } C) &= P(A)P(B | A)P(C | A \text{ and } B) \\ &= 0.05 \times 0.017 \times 0.4 = 0.00034 \end{aligned}$$

Only about 3 of every 10,000 high school athletes can expect to compete in college and have a professional career of more than three years. High school students would be wise to concentrate on studies rather than on unrealistic hopes of fortune from pro sports.

## Tree diagrams

Probability problems often require us to combine several of the basic rules into a

more elaborate calculation. Here is an example that illustrates how to solve problems that have several stages.

## Example

### 4.47 Online chat rooms.

Online chat rooms are dominated by the young. Teens are the biggest users. If we look only at adult Internet users (aged 18 and over), 47% of the 18 to 29 age group chat, as do 21% of the 30 to 49 age group and just 7% of those 50 and over. To learn what percent of all Internet users participate in chat, we also need the age breakdown of users. Here it is: 29% of adult Internet users are 18 to 29 years old (event  $A_1$ ), another 47% are 30 to 49 (event  $A_2$ ), and the remaining 24% are 50 and over (event  $A_3$ ).

What is the probability that a randomly chosen adult user of the Internet participates in chat rooms (event  $C$ )? To find out, use the **tree diagram** in Figure 4.19 to organize your thinking. Each segment in the tree is one stage of the problem. Each complete branch shows a path through the two stages. The probability written on each segment is the conditional probability of an Internet user following that segment, given that he or she has reached the node from which it branches.

#### tree diagram

Starting at the left, an Internet user falls into one of the three age groups. The probabilities of these groups

$$P(A_1) = 0.29 \quad P(A_2) = 0.47 \quad P(A_3) = 0.24$$

mark the leftmost branches in the tree. Conditional on being 18 to 29 years old, the probability of participating in chat is  $P(C | A_1) = 0.47$ . So the conditional probability of *not* participating is

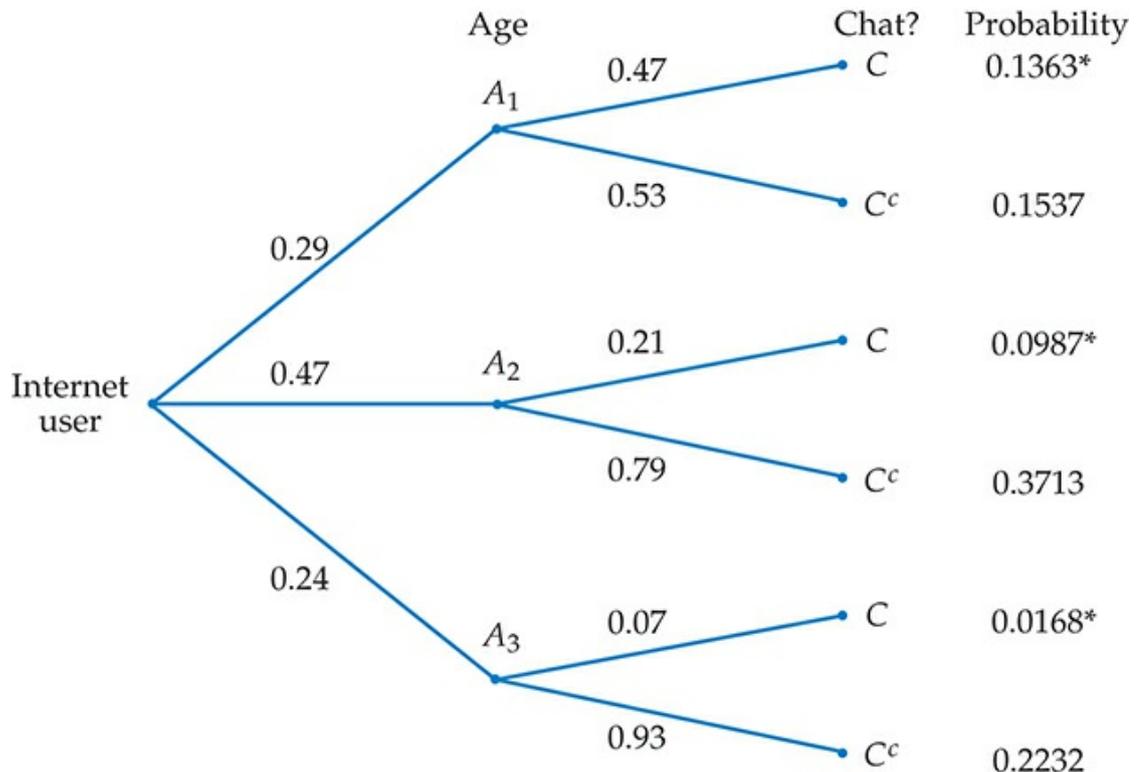
$$P(C^c | A_1) = 1 - 0.47 = 0.53$$

These conditional probabilities mark the paths branching out from the  $A_1$  node in Figure 4.19. The other two age group nodes similarly lead to two branches marked with the conditional probabilities of chatting or not. The probabilities on the branches from any node add to 1 because they cover all possibilities, given that this node was reached.

There are three disjoint paths to  $C$  one for each age group. By the addition rule,  $P(C)$  is the sum of their probabilities. The probability of reaching  $C$  through the 18 to 29 age group is

$$P(C \text{ and } A_1) = P(A_1)P(C | A_1)$$

$$= 0.29 \times 0.47 = 0.1363$$



**FIGURE 4.19**

Tree diagram for Example 4.47. The probability  $P(C)$  is the sum of the probabilities of the three branches marked with asterisks (\*).

Follow the paths to  $C$  through the other two age groups. The probabilities of these paths are

$$P(C \text{ and } A_2) = P(A_2)P(C | A_2) = (0.47)(0.21) = 0.0987$$

$$P(C \text{ and } A_3) = P(A_3)P(C | A_3) = (0.24)(0.07) = 0.0168$$

The final result is

$$P(C) = 0.1363 + 0.0987 + 0.0168 = 0.2518$$

About 25% of all adult Internet users take part in chat rooms.

It takes longer to explain a tree diagram than it does to use it. Once you have understood a problem well enough to draw the tree, the rest is easy. Tree diagrams combine the addition and multiplication rules. The multiplication rule says that the

probability of reaching the end of any complete branch is the product of the probabilities written on its segments. The probability of any outcome, such as the event  $C$  that an adult Internet user takes part in chat rooms, is then found by adding the probabilities of all branches that are part of that event.

## USE YOUR KNOWLEDGE

### 4.101 Draw a tree diagram.

Refer to Slim's chances of a flush in Exercise 4.98 (page 288). Draw a tree diagram to describe the outcomes for the two cards that he will be dealt. At the first stage, his draw can be a diamond or a nondiamond. At the second stage, he has the same possible outcomes but the probabilities are different.

## Bayes's rule

There is another kind of probability question that we might ask in the context of thinking about online chat. What percent of adult chat room participants are aged 18 to 29?

## Example

### 4.48 Conditional versus unconditional probabilities.

In the notation of Example 4.47 this is the conditional probability  $P(A_1 | C)$ . Start from the definition of conditional probability and then apply the results of Example 4.46:

$$\begin{aligned}P(A_1|C) &= P(A_1 \text{ and } C)P(C) \\ &= 0.13630.2518 = 0.5413\end{aligned}$$

Over half of adult chat room participants are between 18 and 29 years old. Compare this conditional probability with the original information (unconditional) that 29% of adult Internet users are between 18 and 29 years

old. Knowing that a person chats increases the probability that he or she is young.

We know the probabilities  $P(A_1)$ ,  $P(A_2)$ , and  $P(A_3)$  that give the age distribution of adult Internet users. We also know the conditional probabilities  $P(C | A_1)$ ,  $P(C | A_2)$ , and  $P(C | A_3)$ , that a person from each age group chats. Example 4.47 shows how to use this information to calculate  $P(C)$ . The method can be summarized in a single expression that adds the probabilities of the three paths to  $C$  in the tree diagram:

$$P(C) = P(A_1) P(C | A_1) + P(A_2) P(C | A_2) + P(A_3) P(C | A_3)$$

In Example 4.48 we calculated the “reverse” conditional probability  $P(A_1 | C)$ . The denominator 0.2518 in that example came from the previous expression. Put in this general notation, we have another probability law.

## BAYES’S RULE

Suppose that  $A_1, A_2, \dots, A_k$  are disjoint events whose probabilities are not 0 and add to exactly 1. That is, any outcome is in exactly one of these events. Then if  $C$  is any other event whose probability is not 0 or 1,

$$P(A_i | C) = \frac{P(C | A_i)P(A_i)}{P(C | A_1)P(A_1) + P(C | A_2)P(A_2) + \dots + P(A_k)P(C | A_k)}$$

The numerator in Bayes’s rule is always one of the terms in the sum that makes up the denominator. The rule is named after Thomas Bayes, who wrestled with arguing from outcomes like  $C$  back to the  $A_i$  in a book published in 1763. It is far better to think your way through problems like Examples 4.47 and 4.48 than to memorize these formal expressions.

## Independence again

The conditional probability  $P(B | A)$  is generally not equal to the unconditional probability  $P(B)$ . That is because the occurrence of event  $A$  generally gives us some additional information about whether or not event  $B$  occurs. If knowing that  $A$  occurs gives no additional information about  $B$ , then  $A$  and  $B$  are independent events. The formal definition of independence is expressed in terms of conditional probability.

## INDEPENDENT EVENTS

Two events  $A$  and  $B$  that both have positive probability are **independent** if

$$P(B | A) = P(B)$$

This definition makes precise the informal description of independence given in Section 4.2. We now see that the multiplication rule for independent events,  $P(A \text{ and } B) = P(A)P(B)$ , is a special case of the general multiplication rule,  $P(A \text{ and } B) = P(A)P(B | A)$ , just as the addition rule for disjoint events is a special case of the general addition rule.

### Section 4.5 Summary

The **complement**  $A^c$  of an event  $A$  contains all outcomes that are not in  $A$ . The **union**  $\{A \text{ or } B\}$  of events  $A$  and  $B$  contains all outcomes in  $A$ , in  $B$ , and in both  $A$  and  $B$ . The **intersection**  $\{A \text{ and } B\}$  contains all outcomes that are in both  $A$  and  $B$  but not outcomes in  $A$  alone or  $B$  alone.

The **conditional probability**  $P(B | A)$  of an event  $B$ , given an event  $A$ , is defined by

$$P(B | A) = \frac{P(A \text{ and } B)}{P(A)}$$

when  $P(A) > 0$ . In practice, conditional probabilities are most often found from directly available information.

The essential general rules of elementary probability are

**Legitimate values:**  $0 \leq P(A) \leq 1$  for any event  $A$

**Total probability 1:**  $P(S) = 1$

**Complement rule:**  $P(A^c) = 1 - P(A)$

**Addition rule:**  $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

**Multiplication rule:**  $P(A \text{ and } B) = P(A)P(B | A)$

If  $A$  and  $B$  are **disjoint**, then  $P(A \text{ and } B) = 0$ . The general addition rule for unions then becomes the special addition rule,  $P(A \text{ or } B) = P(A) + P(B)$ .

$A$  and  $B$  are **independent** when  $P(B | A) = P(B)$ . The multiplication rule for intersections then becomes  $P(A \text{ and } B) = P(A)P(B)$ .

In problems with several stages, draw a **tree diagram** to organize use of the multiplication and addition rules.

### SECTION 4.5 Exercises

*For Exercise 4.95, see page 284; for Exercise 4.96, see page 285; for Exercise 4.97, see page 287; for Exercise 4.98, see page 288; for Exercises 4.99 and 4.100, see page 289; and for Exercise 4.101, see page 292.*

#### 4.102 Find and explain some probabilities.

(a) Can we have an event  $A$  that has negative probability? Explain your answer.

(b) Suppose  $P(A) = 0.2$  and  $P(B) = 0.4$ . Explain what it means for  $A$  and  $B$  to be disjoint. Assuming that they are disjoint, find the probability that  $A$  or  $B$  occurs.

(c) Explain in your own words the meaning of the rule  $P(S) = 1$ .

(d) Consider an event  $A$ . What is the name for the event that  $A$  does not occur? If  $P(A) = 0.3$ , what is the probability that  $A$  does not occur?

(e) Suppose that  $A$  and  $B$  are independent and that  $P(A) = 0.2$  and  $P(B) = 0.5$ . Explain the meaning of the event  $\{A \text{ and } B\}$ , and find its probability.

#### **4.103 Unions.**

(a) Assume that  $P(A) = 0.4$ ,  $P(B) = 0.3$ , and  $P(C) = 0.1$ . If the events  $A$ ,  $B$ , and  $C$  are disjoint, find the probability that the union of these events occurs.

(b) Draw a Venn diagram to illustrate your answer to part (a).

(c) Find the probability of the complement of the union of  $A$ ,  $B$ , and  $C$ .

#### **4.104 Conditional probabilities.**

Suppose that  $P(A) = 0.5$ ,  $P(B) = 0.3$ , and  $P(B | A) = 0.2$ .

(a) Find the probability that both  $A$  and  $B$  occur.

(b) Use a Venn diagram to explain your calculation.

(c) What is the probability of the event that  $B$  occurs and  $B$  does not?

#### **4.105 Find the probabilities.**

Suppose that the probability that  $A$  occurs is 0.6 and the probability that  $A$  and  $B$  occur is 0.5.

(a) Find the probability that  $B$  occurs given that  $A$  occurs.

(b) Illustrate your calculations in part (a) using a Venn diagram.

#### **4.106 Why not?**

Suppose that  $P(A) = 0.4$ . Explain why  $P(A \text{ and } B)$  cannot be 0.5.

#### **4.107 Is the calcium intake adequate?**

In the population of young children eligible to participate in a study of whether or not their calcium intake is adequate, 52% are 5 to 10 years of age and 48% are 11 to 13 years of age. For those who are 5 to 10 years of age, 18% have inadequate calcium intake. For those who are 11 to 13 years of age, 57% have inadequate calcium intake.<sup>19</sup>

(a) Use letters to define the events of interest in this exercise.

(b) Convert the percents given to probabilities of the events you have defined.

(c) Use a tree diagram similar to Figure 4.19 (page 291) to calculate the probability that a randomly selected child from this population has an inadequate intake of calcium.

#### 4.108 Use Bayes's rule.

Refer to the previous exercise. Use Bayes's rule to find the probability that a child from this population who has inadequate intake is 11 to 13 years old.

#### 4.109 Are the events independent?

Refer to the previous two exercises. Are the age of the child and whether or not the child has adequate calcium intake independent? Calculate the probabilities that you need to answer this question, and write a short summary of your conclusion.

#### 4.110 What's wrong?

In each of the following scenarios, there is something wrong. Describe what is wrong and give a reason for your answer.

- (a)  $P(A \text{ or } B)$  is always equal to the sum of  $P(A)$  and  $P(B)$ .
- (b) The probability of an event minus the probability of its complement is always equal to 1.
- (c) Two events are disjoint if  $P(B | A) = P(B)$ .

#### 4.111 Exercise and sleep.

Suppose that 40% of adults get enough sleep, 46% get enough exercise, and 24% do both. Find the probabilities of the following events:

- (a) enough sleep and not enough exercise
- (b) not enough sleep and enough exercise
- (c) not enough sleep and not enough exercise
- (d) For each of parts (a), (b), and (c), state the rule that you used to find your answer.

#### 4.112 Exercise and sleep.

Refer to the previous exercise. Draw a Venn diagram showing the probabilities for exercise and sleep.

#### 4.113 Lying to a teacher.

Suppose that 48% of high school students would admit to lying at least once to a teacher during the past year and that 25% of students are male and would admit to lying at least once to a teacher during the past year.<sup>20</sup> Assume that 50% of the students are male. What is the probability that a randomly selected student is either male or would admit to lying to a teacher, during the past year? Be sure to show your work and indicate all the rules that you use to find your answer.

#### 4.114 Lying to a teacher.

Refer to the previous exercise. Suppose that you select a student from the subpopulation of those who would admit to lying to a teacher during the past year. What is the probability that the student is female? Be sure to show your work and indicate all the rules that you use to find your answer.

#### 4.115 Attendance at two-year and four-year colleges.

In a large national population of college students, 61% attend four-year institutions and the rest attend two-year institutions. Males make up 44% of the students in the four-year institutions and 41% of the students in the two-year institutions.

(a) Find the four probabilities for each combination of gender and type of institution in the following table. Be sure that your probabilities sum to 1.

	Men	Women
Four-year institution		
Two-year institution		

(b) Consider randomly selecting a female student from this population. What is the probability that she attends a four-year institution?

#### 4.116 Draw a tree diagram.

Refer to the previous exercise. Draw a tree diagram to illustrate the probabilities in a situation where you first identify the type of institution attended and then identify the gender of the student.

#### 4.117 Draw a different tree diagram for the same setting.

Refer to the previous two exercises. Draw a tree diagram to illustrate the probabilities in a situation where you first identify the gender of the student and then identify the type of institution attended. Explain why the probabilities in this tree diagram are different from those that you used in the previous exercise.

#### 4.118 Education and income.

Call a household prosperous if its income exceeds \$100,000. Call the household educated if the householder completed college. Select an American household at random, and let  $A$  be the event that the selected household is prosperous and  $B$  the event that it is educated. According to the Current Population Survey,  $P(A) = 0.138$ ,  $P(B) = 0.261$ , and the probability that a household is both prosperous and educated is  $P(A \text{ and } B) = 0.082$ . What is the probability  $P(A \text{ or } B)$  that the household selected is either prosperous or educated?

#### 4.119 Find a conditional probability.

In the setting of the previous exercise, what is the conditional probability that a household is prosperous, given that it is educated? Explain why your result shows that events  $A$  and  $B$  are not independent.

#### 4.120 Draw a Venn diagram.

Draw a Venn diagram that shows the relation between the events  $A$  and  $B$  in Exercise 4.118. Indicate each of the following events on your diagram and use the information in Exercise 4.118 to calculate the probability of each event. Finally, describe in words what each event is.

- (a)  $\{A \text{ and } B\}$
- (b)  $\{A^C \text{ and } B\}$
- (c)  $\{A \text{ and } B^C\}$
- (d)  $\{A^C \text{ and } B^C\}$

#### 4.121 Sales of cars and light trucks.

Motor vehicles sold to individuals are classified as either cars or light trucks (including SUVs) and as either domestic or imported. In a recent year, 69% of vehicles sold were light trucks, 78% were domestic, and 55% were domestic light trucks. Let  $A$  be the event that a vehicle is a car and  $B$  the event that it is imported. Write each of the following events in set notation and give its probability.

- (a) The vehicle is a light truck.
- (b) The vehicle is an imported car.

#### 4.122 Job offers.

Julie is graduating from college. She has studied biology, chemistry, and computing and hopes to work as a forensic scientist applying her science background to crime investigation. Late one night she thinks about some jobs she has applied for. Let  $A$ ,  $B$ , and  $C$  be the events that Julie is offered a job by

$A$  = the Connecticut Office of the Chief Medical Examiner

$B$  = the New Jersey Division of Criminal Justice

$C$  = the federal Disaster Mortuary Operations Response Team

Julie writes down her personal probabilities for being offered these jobs:

$$P(A) = 0.7$$

$$P(B) = 0.5$$

$$P(C) = 0.3$$

$$P(A \text{ and } B) = 0.3$$

$$P(A \text{ and } C) = 0.1$$

$$P(B \text{ and } C) = 0.1$$

$$P(A \text{ and } B \text{ and } C) = 0$$

Make a Venn diagram of the events  $A$ ,  $B$ , and  $C$ . As in Figure 4.18 (page 286), mark the probabilities of every intersection involving these events and their complements. Use this diagram for Exercises 4.123 to 4.125.

#### 4.123 Find the probability of at least one offer.

What is the probability that Julie is offered at least one of the three jobs?

#### 4.124 Find the probability of another event.

What is the probability that Julie is offered both the Connecticut and New Jersey jobs, but not the federal job?

#### 4.125 Find a conditional probability.

If Julie is offered the federal job, what is the conditional probability that she is also offered the New Jersey job? If Julie is offered the New Jersey job, what is the conditional probability that she is also offered the federal job?

#### 4.126 Academic degrees and gender.

Here are the projected numbers (in thousands) of earned degrees in the United States in the 2010–2011 academic year, classified by level and by the sex of the degree recipient:<sup>21</sup>

	Bachelor's	Master's	Professional	Doctorate
Female	933	502	51	26
Male	661	260	44	26

- Convert this table to a table giving the probabilities for selecting a degree earned and classifying the recipient by gender and the degree by the levels given above.
- If you choose a degree recipient at random, what is the probability that the person you choose is a woman?
- What is the conditional probability that you choose a woman, given that the person chosen received a professional degree?
- Are the events “choose a woman” and “choose a professional degree recipient” independent? How do you know?

#### 4.127 Find some probabilities.

The previous exercise gives the projected number (in thousands) of earned degrees in the United States in the 2010–2011 academic year. Use these data to answer the following questions.

- What is the probability that a randomly chosen degree recipient is a man?
- What is the conditional probability that the person chosen received a bachelor's degree, given that he is a man?
- Use the multiplication rule to find the probability of choosing a male bachelor's degree recipient. Check your result by finding this probability directly from the table of counts.

#### 4.128 Conditional probabilities and independence.

Using the information in Exercise 4.121, answer these questions.

- (a) Given that a vehicle is imported, what is the conditional probability that it is a light truck?
- (b) Are the events “vehicle is a light truck” and “vehicle is imported” independent? Justify your answer.

### ***Genetic counseling.***

*Conditional probabilities and Bayes’s rule are a basis for counseling people who may have genetic defects that can be passed to their children. Exercises 4.129 to 4.131 concern genetic counseling settings.*

#### **4.129 Albinism.**

People with albinism have little pigment in their skin, hair, and eyes. The gene that governs albinism has two forms (called alleles), which we denote by  $a$  and  $A$ . Each person has a pair of these genes, one inherited from each parent. A child inherits one of each parent’s two alleles independently with probability 0.5. Albinism is a recessive trait, so a person is albino only if the inherited pair is  $aa$ .

- (a) Beth’s parents are not albino but she has an albino brother. This implies that both of Beth’s parents have type  $Aa$ . Why?
- (b) Which of the types  $aa$ ,  $Aa$ ,  $AA$  could a child of Beth’s parents have? What is the probability of each type?
- (c) Beth is not albino. What are the conditional probabilities for Beth’s possible genetic types, given this fact? (Use the definition of conditional probability.)

#### **4.130 Find some conditional probabilities.**

Beth knows the probabilities for her genetic types from part (c) of the previous exercise. She marries Bob, who is albino. Bob’s genetic type must be  $aa$ .

- (a) What is the conditional probability that a child of Beth and Bob is non-albino if Beth has type  $Aa$ ? What is the conditional probability of a non-albino child if Beth has type  $AA$ ?
- (b) Beth and Bob’s first child is non-albino. What is the conditional probability that Beth is a carrier, type  $Aa$ ?

#### **4.131 Muscular dystrophy.**

Muscular dystrophy is an incurable muscle-wasting disease. The most common and serious type, called DMD, is caused by a sex-linked recessive mutation. Specifically, women can be carriers but do not get the disease; a son of a carrier has probability 0.5 of having DMD; a daughter has probability 0.5 of being a carrier. As many as one-third of DMD cases, however, are due to spontaneous mutations in sons of mothers who are not carriers. Toni has one son, who has DMD.

In the absence of other information, the probability is  $1/3$  that the son is the victim of a spontaneous mutation and  $2/3$  that Toni is a carrier. There is a screening test called the CK test that is positive with probability 0.7 if a woman is a carrier and with probability 0.1 if she is not. Toni’s CK test is positive. What is the probability that she is a carrier?

## CHAPTER 4 Exercises

### 4.132 Repeat the experiment many times.

Here is a probability distribution for a random variable  $X$ :

Value of $X$	-1	2
Probability	0.4	0.6

A single experiment generates a random value from this distribution. If the experiment is repeated many times, what will be the approximate proportion of times that the value is  $-1$ ? Give a reason for your answer.

### 4.133 Repeat the experiment many times and take the mean.

Here is a probability distribution for a random variable  $X$ :

Value of $X$	-1	2
Probability	0.2	0.8

A single experiment generates a random value from this distribution. If the experiment is repeated many times, what will be the approximate value of the mean of these random variables? Give a reason for your answer.

### 4.134 Work with a transformation.

Here is a probability distribution for a random variable  $X$

Value of $X$	1	2
Probability	0.4	0.6

- Find the mean and the standard deviation of this distribution.
- Let  $Y = 4X - 2$ . Use the rules for means and variances to find the mean and the standard deviation of the distribution of  $Y$ .
- For part (b) give the rules that you used to find your answer.

### 4.135 A different transformation.

Refer to the previous exercise. Now let  $Y = 4X^2 - 2$ .

- Find the distribution of  $Y$ .
- Find the mean and standard deviation for the distribution of  $Y$ .

(c) Explain why the rules that you used for part (b) of the previous exercise do not work for this transformation.

#### 4.136 Roll a pair of dice two times.

Consider rolling a pair of fair dice two times. Let  $A$  be the total on the up-faces for the first roll and let  $B$  be the total on the up-faces for the second roll. For each of the following pairs of events, tell whether they are disjoint, independent, or neither.

- (a)  $A = 2$  on the first roll,  $B = 8$  or more on the first roll.
- (b)  $A = 2$  on the first roll,  $B = 8$  or more on the second roll.
- (c)  $A = 5$  or less on the second roll,  $B = 4$  or less on the first roll.
- (d)  $A = 5$  or less on the second roll,  $B = 4$  or less on the second roll.

#### 4.137 Find the probabilities.

Refer to the previous exercise. Find the probabilities for each event.

#### 4.138 Some probability distributions.

Here is a probability distribution for a random variable  $X$ :

Value of $X$	2	3	4
Probability	0.2	0.4	0.4

- (a) Find the mean and standard deviation for this distribution.
- (b) Construct a different probability distribution with the same possible values, the same mean, and a larger standard deviation. Show your work and report the standard deviation of your new distribution.
- (c) Construct a different probability distribution with the same possible values, the same mean, and a smaller standard deviation. Show your work and report the standard deviation of your new distribution.

#### 4.139 A fair bet at craps.

Almost all bets made at gambling casinos favor the house. In other words, the difference between the amount bet and the mean of the distribution of the payoff is a positive number. An exception is “taking the odds” at the game of craps, a bet that a player can make under certain circumstances. The bet becomes available when a shooter throws a 4, 5, 6, 8, 9, or 10 on the initial roll. This number is called the “point”; when a point is rolled, we say that a point has been established. If a 4 is the point, an odds bet can be made that wins if a 4 is rolled before a 7 is rolled. The probability of winning this bet is  $1/3$  and the payoff for a \$10 bet is \$20 (you keep the \$10 you bet and you receive an additional \$20). The same probability of winning and the same payoff apply for an odds bet on a 10. For an initial roll of 5 or 9, the odds bet has a winning probability of  $2/5$  and the payoff for a \$10 bet is \$15. Similarly, when the initial roll is 6 or 8, the odds bet has a winning probability of  $5/11$  and the payoff for a \$10 bet is \$12. Find the mean of the payoff distribution for each of these bets. Then confirm that the bets are fair by showing that the difference between the amount bet and the mean of the distribution of the payoff is zero.

#### 4.140 An ancient Korean drinking game.

An ancient Korean drinking game involves a 14-sided die. The players roll the die in turn and must submit to whatever humiliation is written on the up-face: something like “Keep still when tickled on face.” Six of the 14 faces are squares. Let’s call them A, B, C, D, E, and F for short. The other eight faces are triangles, which we will call 1, 2, 3, 4, 5, 6, 7, and 8. Each of the squares is equally likely. Each of the triangles is also equally likely, but the triangle probability differs from the square probability. The probability of getting a square is 0.72. Give the probability model for the 14 possible outcomes.

#### 4.141 Wine tasters.

Two wine tasters rate each wine they taste on a scale of 1 to 5. From data on their ratings of a large number of wines, we obtain the following probabilities for both tasters’ ratings of a randomly chosen wine:

Taster 1	Taster 2				
	1	2	3	4	5
1	0.03	0.02	0.01	0.00	0.00
2	0.02	0.07	0.06	0.02	0.01
3	0.01	0.05	0.25	0.05	0.01
4	0.00	0.02	0.05	0.20	0.02
5	0.00	0.01	0.01	0.02	0.06

- (a) Why is this a legitimate assignment of probabilities to outcomes?
- (b) What is the probability that the tasters agree when rating a wine?
- (c) What is the probability that Taster 1 rates a wine higher than 3? What is the probability that Taster 2 rates a wine higher than 3?

#### 4.142 SAT scores.

The College Board finds that the distribution of students’ SAT scores depends on the level of education their parents have. Children of parents who did not finish high school have SAT Math scores  $X$  with mean 445 and standard deviation 106. Scores  $Y$  of children of parents with graduate degrees have mean 566 and standard deviation 109. Perhaps we should standardize to a common scale for equity. Find positive numbers  $a$ ,  $b$ , and  $c$ , such that  $a + bX$  and  $c + dY$  both have mean 500 and standard deviation 100.

#### 4.143 Lottery tickets.

Joe buys a ticket in the Tri-State Pick 3 lottery every day, always betting on 956. He will win something if the winning number contains 9, 5, and 6 in any order. Each day, Joe has probability 0.006 of winning, and he wins (or not) independently of other days because a new drawing is held each day. What is the probability that Joe’s first winning ticket comes on the 20th day?

#### 4.144 Slot machines.

Slot machines are now video games, with winning determined by electronic random number

generators. In the old days, slot machines were like this: you pull the lever to spin three wheels; each wheel has 20 symbols, all equally likely to show when the wheel stops spinning; the three wheels are independent of each other. Suppose that the middle wheel has 8 bells among its 20 symbols, and the left and right wheels have 1 bell each.

(a) You win the jackpot if all three wheels show bells. What is the probability of winning the jackpot?

(b) What is the probability that the wheels stop with exactly 2 bells showing?

*The following exercises require familiarity with the material presented in the optional Section 4.5.*

#### 4.145 Bachelor's degrees by gender.

Of the 2,325,000 bachelor's, master's, and doctoral degrees given by U.S. colleges and universities in a recent year, 69% were bachelor's degrees, 28% were master's degrees, and the rest were doctorates. Moreover, women earned 57% of the bachelor's degrees, 60% of the master's degrees, and 52% of the doctorates.<sup>22</sup> You choose a degree at random and find that it was awarded to a woman. What is the probability that it is a bachelor's degree?

#### 4.146 Higher education at two-year and four-year institutions.

The following table gives the counts of U.S. institutions of higher education classified as public or private and as two-year or four-year.<sup>23</sup>

	Public	Private
Two-year	1000	721
Four-year	2774	672

Convert the counts to probabilities and summarize the relationship between these two variables using conditional probabilities.

#### 4.147 Odds bets at craps.

Refer to the odds bets at craps in Exercise 4.139. Suppose that whenever the shooter has an initial roll of 4, 5, 6, 8, 9, or 10, you take the odds. Here are the probabilities for these initial rolls:

Point	4	5	6	8	9	10
Probability	3/36	4/36	5/36	5/36	4/36	3/36

Draw a tree diagram with the first stage showing the point rolled and the second stage showing whether the point is again rolled before a 7 is rolled. Include a firststage branch showing the outcome that a point is not established. In this case, the amount bet is zero and the distribution of the winnings is the special random variable that has  $P(X = 0) = 1$ . For the combined betting system where the player always makes a \$10 odds bet when it is available, show that the game is fair.

#### 4.148 Weights and heights of children adjusted for age.

The idea of conditional probabilities has many interesting applications, including the idea of a conditional distribution. For example, the National Center for Health Statistics produces distributions for weight and height for children while conditioning on other variables. Visit the website [cdc.gov/growthcharts/](http://cdc.gov/growthcharts/) and describe the different ways that weight and height distributions

are conditioned on other variables.

#### 4.149 Wine tasting.

In the setting of Exercise 4.141, Taster 1's rating for a wine is 3. What is the conditional probability that Taster 2's rating is higher than 3?

#### 4.150 An interesting case of independence.

Independence of events is not always obvious. Toss two balanced coins independently. The four possible combinations of heads and tails in order each have probability 0.25. The events

$$A = \text{head on the first toss}$$

$$B = \text{both tosses have the same outcome}$$

may seem intuitively related. Show that  $P(B | A) = P(B)$ , so that  $A$  and  $B$  are in fact independent.

#### 4.151 Find some conditional probabilities.

Choose a point at random in the square with sides  $0 \leq x \leq 1$  and  $0 \leq y \leq 1$ . This means that the probability that the point falls in any region within the square is the area of that region. Let  $X$  be the  $x$  coordinate and  $Y$  the  $y$  coordinate of the point chosen. Find the conditional probability  $P(Y < 1/3 | Y > X)$ . (*Hint*: Sketch the square and the events  $Y < 1/3$  and  $Y > X$ .)



#### 4.152 Sample surveys for sensitive issues.

It is difficult to conduct sample surveys on sensitive issues because many people will not answer questions if the answers might embarrass them. **Randomized response** is an effective way to guarantee anonymity while collecting information on topics such as student cheating or sexual behavior. Here is the idea. To ask a sample of students whether they have plagiarized a term paper while in college, have each student toss a coin in private. If the coin lands heads *and* they have not plagiarized, they are to answer "No." Otherwise, they are to give "Yes" as their answer. Only the student knows whether the answer reflects the truth or just the coin toss, but the researchers can use a proper random sample with follow-up for nonresponse and other good sampling practices.

Suppose that in fact the probability is 0.3 that a randomly chosen student has plagiarized a paper. Draw a tree diagram in which the first stage is tossing the coin and the second is the truth about plagiarism. The outcome at the end of each branch is the answer given to the randomized-response question. What is the probability of a "No" answer in the randomized-response poll? If the probability of plagiarism were 0.2, what would be the probability of a "No" response on the poll? Now suppose that you get 39% "No" answers in a randomized-response poll of a large sample of students at your college. What do you estimate to be the percent of the population who have plagiarized a paper?

# 5 Analysis of Two-Way Tables

CHAPTER



- 5.1 The Sampling Distribution of a Sample Mean
- 5.2 Sampling Distributions for Counts and Proportions

## Introduction

Statistical inference draws conclusions about a population or process from data. It emphasizes substantiating these conclusions via probability calculations, as probability allows us to take chance variation into account. We have already examined data and arrived at conclusions many times. How do we move from summarizing a single data set to formal inference involving probability calculations?



**parameters and statistics**, p. 206

The foundation for this was described in Section 3.4 (page 205). There, we not only discussed the use of *statistics* as estimates of population *parameters* but also described the chance variation of a statistic when the data are produced by random sampling or randomized experimentation.



**sampling distribution**, p. 208

The *sampling distribution* of a statistic shows how it would vary in these identical repeated data collections. That is, the sampling distribution is a probability distribution that answers the question “What would happen if we did this experiment or sampling many times?” It is these distributions that provide the necessary link between probability and the data in your sample or from your experiment. They are the key to understanding statistical inference.

Suppose that you plan to survey 1000 students at your university about their sleeping habits. The sampling distribution of the average hours of sleep per night describes what this average would be if many simple random samples of 1000 students were drawn from the population of students at your university. In other words, it gives you an idea of what you are likely to see from your survey. It tells you whether you should expect this average to be near the population mean and whether the variation of the statistic is roughly  $\pm 2$  hours or  $\pm 2$  minutes.

### THE DISTRIBUTION OF A STATISTIC

A statistic from a random sample or randomized experiment is a random variable. The probability distribution of the statistic is its **sampling distribution**.

To help in the transition from probability as a topic in itself to probability as a

foundation for inference, in this chapter we will study the sampling distributions of some common statistics. The general framework for constructing a sampling distribution is the same for all statistics, so our focus here will be on those statistics commonly used in inference.

 **LOOK BACK**  
density curves, p. 56

Before doing so, however, we need to consider another set of probability distributions that also play a role in statistical inference. Any quantity that can be measured on each member of a population is described by the distribution of its values for all members of the population. This is the context in which we first met distributions, as density curves that provide models for the overall pattern of data.

Imagine choosing one individual at random from a population and measuring a quantity. The quantities obtained from repeated draws of one individual from a population have a probability distribution that is the distribution of the population.

## Example

### 5.1 Total sleep time of college students.

A recent survey describes the distribution of total sleep time among college students as approximately Normal with a mean of 6.78 hours and standard deviation of 1.24 hours.<sup>1</sup> Suppose that we select a college student at random and obtain his or her sleep time. This result is a random variable  $X$  because prior to the random sampling, we don't know the sleep time. We do know, however, that in repeated sampling  $X$  will have the same  $N(6.78, 1.24)$  distribution that describes the pattern of sleep time in the entire population. We call  $N(6.78, 1.24)$  the *population distribution*.

## POPULATION DISTRIBUTION

The **population distribution** of a variable is the distribution of its values for all members of the population. The population distribution is also the probability distribution of the variable when we choose one individual at random from the population.

 **LOOK BACK**

In this example, the population of all college students actually exists, so that we can in principle draw an SRS of students from it. Sometimes our population of interest does not actually exist. For example, suppose that we are interested in studying final-exam scores in a statistics course, and we have the scores of the 34 students who took the course last semester. For the purposes of statistical inference, we might want to consider these 34 students as part of a hypothetical population of similar students who would take this course. In this sense, these 34 students represent not only themselves but also a larger population of similar students. The key idea is to think of the observations that you have as coming from a population with a probability distribution.

## USE YOUR KNOWLEDGE

### 5.1 Number of apps on an iOS device.

AppsFire is a service that shares the names of the apps on an iOS device with everyone else using the service. This, in a sense, creates an iOS device app recommendation system. Recently, the service drew a sample of 1000 AppsFire users and reported a median of 108 apps per device.<sup>2</sup> State the population that this survey describes, the statistic, and some likely values from the population distribution.

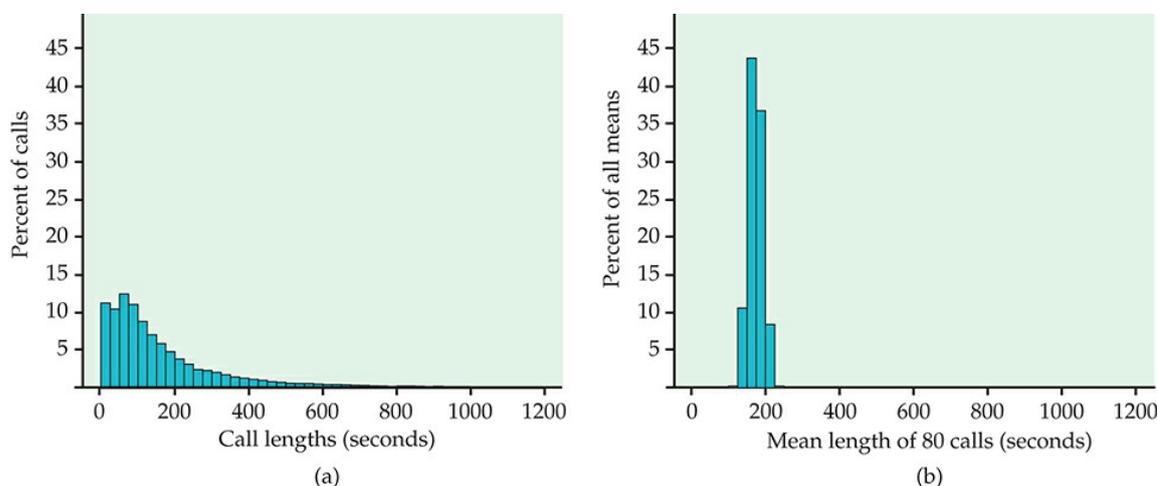
In the next two sections, we will study the sampling distributions of two common statistics, the sample mean and the sample proportion. The focus will be on the important features of these distributions so that we can quickly describe and use them in the later chapters on statistical inference. We will see that in each case the sampling distribution depends on **both** the population distribution and the way we collect the data from the population.

## 5.1 The Sampling Distribution of a Sample Mean

When you complete this section, you will be able to

- Explain the difference between the sampling distribution of  $\bar{x}$  and the population distribution.
- Determine the mean and standard deviation of  $\bar{x}$  for an SRS of size  $n$  from a population with mean  $\mu$  and standard deviation  $\sigma$ .
- Describe how much larger  $n$  has to be to reduce the standard deviation of  $\bar{x}$  by a certain factor.
- Utilize the central limit theorem to approximate the sampling distribution of  $\bar{x}$  and perform various probability calculations.

A variety of statistics are used to describe quantitative data. The sample mean, median, and standard deviation are all examples of statistics based on quantitative data. Statistical theory describes the sampling distributions of these statistics. However, the general framework for constructing a sampling distribution is the same for all statistics. In this section we will concentrate on the sample mean. Because sample means are just averages of observations, they are among the most frequently used statistics.



**FIGURE 5.1**

(a) The distribution of lengths of all customer service calls received by a bank in a month, for Example 5.2. (b) The distribution of the sample means  $\bar{x}$  for 500 random samples of size 80 from this population. The scales and histogram classes are exactly the same in both panels.

## Example

### 5.2 Sample means are approximately Normal.



CALLS80

Figure 5.1 illustrates two striking facts about the sampling distribution of a sample mean. Figure 5.1(a) displays the distribution of customer service call lengths for a bank service center for a month. There are more than 30,000 calls in this population.<sup>3</sup> (We omitted a few extreme outliers, calls that lasted more than 20 minutes.) The distribution is extremely skewed to the right. The population mean is  $\mu = 173.95$  seconds.

Table 1.2 (page 19) contains the lengths of a random sample of 80 calls from this population. The mean of these 80 calls is  $\bar{x} = 196.6$  seconds. If we were to take another sample of size 80, we would likely get a different value of  $\bar{x}$ . This is because this new sample would contain a different set of calls. To find the sampling distribution of  $\bar{x}$ , we take many SRSs of size 80 and calculate  $\bar{x}$  for each sample. Figure 5.1(b) is the distribution of the values of  $\bar{x}$  for 500 random samples. The scales and choice of classes are exactly the same as in Figure 5.1(a), so that we can make a direct comparison.

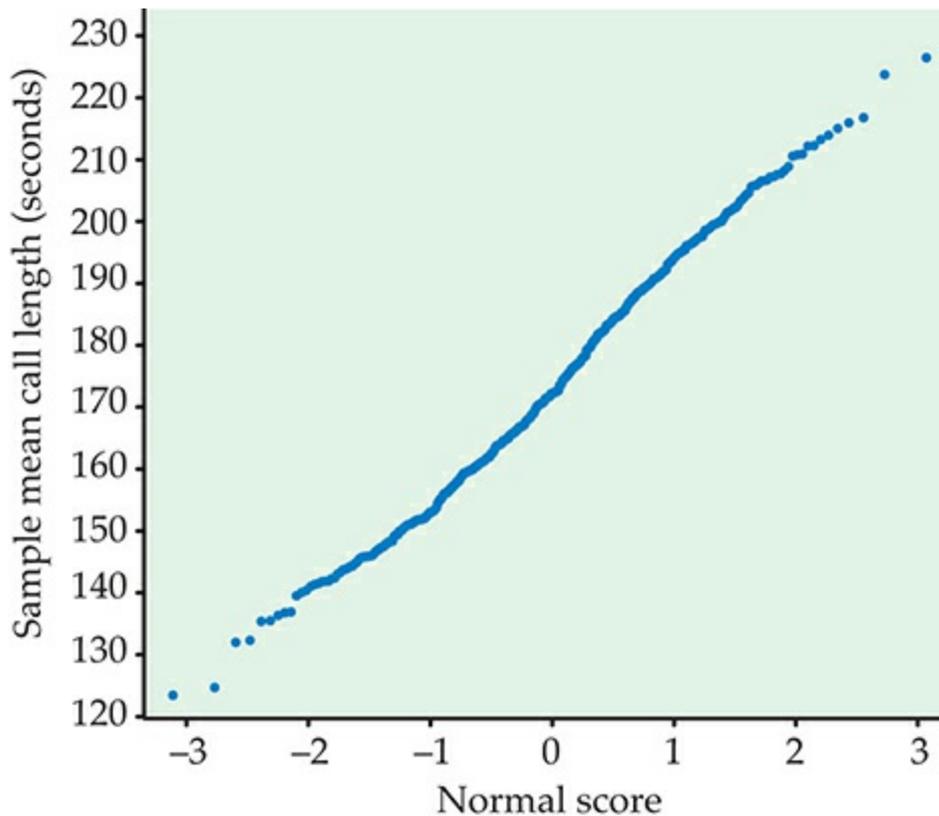
The sample means are much less spread out than the individual call lengths. What is more, the distribution in Figure 5.1(b) is roughly symmetric rather than skewed. The Normal quantile plot in Figure 5.2 confirms that the distribution is close to Normal.

This example illustrates two important facts about sample means that we will discuss in this section.

#### FACTS ABOUT SAMPLE MEANS

1. Sample means are less variable than individual observations.
2. Sample means are more Normal than individual observations.

These two facts contribute to the popularity of sample means in statistical inference.



**FIGURE 5.2**

Normal quantile plot of the 500 sample means in Figure 5.1(b). The distribution is close to Normal.

### The mean and standard deviation of $\bar{x}$

The sample mean  $\bar{x}$  from a sample or an experiment is an estimate of the mean  $\mu$  of the underlying population. The sampling distribution of  $\bar{x}$  is determined by the design used to produce the data, the sample size  $n$ , and the population distribution.

Select an SRS of size  $n$  from a population, and measure a variable  $X$  on each individual in the sample. The  $n$  measurements are values of  $n$  random variables  $X_1, X_2, \dots, X_n$ . A single  $X_i$  is a measurement on one individual selected at random from the population and therefore has the distribution of the population. If the population is large relative to the sample, we can consider  $X_1, X_2, \dots, X_n$  to be independent random variables each having the same distribution. This is our probability model for measurements on each individual in an SRS.

The sample mean of an SRS of size  $n$  is

$$\bar{x} = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$$

 **LOOK BACK**  
rules for means, p. 272

If the population has mean  $\mu$  then  $\mu$  is the mean of the distribution of each

observation  $X_j$ . To get the mean of  $\bar{x}$ , we use the rules for means of random variables. Specifically,

$$\begin{aligned}\mu_{\bar{x}} &= \frac{1}{n}(\mu_{X1} + \mu_{X2} + \dots + \mu_{Xn}) \\ &= \frac{1}{n}(\mu + \mu + \dots + \mu) = \mu\end{aligned}$$

 **LOOK BACK**  
unbiased estimator, p. 210

That is, *the mean of  $\bar{x}$  is the same as the mean of the population*. The sample mean  $\bar{x}$  is therefore an unbiased estimator of the unknown population mean  $\mu$

 **LOOK BACK**  
sampling distribution, p. 275

The observations are independent, so the addition rule for variances also applies:

$$\begin{aligned}\sigma_{\bar{x}}^2 &= \frac{1}{n^2}(\sigma_{X1}^2 + \sigma_{X2}^2 + \dots + \sigma_{Xn}^2) \\ &= \frac{1}{n^2}(\sigma^2 + \sigma^2 + \dots + \sigma^2) \\ &= \frac{\sigma^2}{n}\end{aligned}$$

With  $n$  in the denominator, the variability of  $\bar{x}$  about its mean decreases as the sample size grows. Thus, a sample mean from a large sample will usually be very close to the true population mean  $\mu$ . Here is a summary of these facts.

### MEAN AND STANDARD DEVIATION OF A SAMPLE MEAN

Let  $\bar{x}$  be the mean of an SRS of size  $n$  from a population having mean  $\mu$  and standard deviation  $\sigma$ . The mean and standard deviation of  $\bar{x}$  are

$$\begin{aligned}\mu_{\bar{x}} &= \mu \\ \sigma_{\bar{x}} &= \frac{\sigma}{n}\end{aligned}$$

How precisely does a sample mean  $\bar{x}$  estimate a population mean  $\mu$ ? Because the values of  $\bar{x}$  vary from sample to sample, we must give an answer in terms of the sampling distribution. We know that  $\bar{x}$  is an unbiased estimator of  $\mu$ , so its values in repeated samples are not systematically too high or too low. Most samples will give an  $\bar{x}$ -value close to  $\mu$  if the sampling distribution is concentrated close to its mean  $\mu$ . So the precision of estimation depends on the spread of the sampling distribution.

Because the standard deviation of  $\bar{x}$  is  $\sigma/n$ , the standard deviation of the statistic decreases in proportion to the square root of the sample size. This means, for example, that a sample size must be multiplied by 4 in order to divide the statistic's standard deviation in half. By comparison, a sample size must be multiplied by 100 in order to reduce the standard deviation by a factor of 10.

## Example

### 5.3 Standard deviations for sample means of service call lengths.



The standard deviation of the population of service call lengths in Figure 5.1(a) is  $\sigma = 184.81$  seconds. The length of a single call will often be far from the population mean. If we choose an SRS of 20 calls, the standard deviation of their mean length is

$$\sigma_{\bar{x}} = 184.81 / \sqrt{20} = 41.32 \text{ seconds}$$

Averaging over more calls reduces the variability and makes it more likely that  $\bar{x}$  is close to  $\mu$ . Our sample size of 80 calls is 4 times 20, so the standard deviation will be half as large:

$$\sigma_{\bar{x}} = 184.81 / \sqrt{80} = 20.66 \text{ seconds}$$

## USE YOUR KNOWLEDGE

### 5.2 Find the mean and the standard deviation of the sampling distribution.

Compute the mean and standard deviation of the sampling distribution of the sample mean when you plan to take an SRS of size 49 from a population with mean 420 and standard deviation 21.

### 5.3 The effect of increasing the sample size.

In the setting of the previous exercise, repeat the calculations for a sample size of 441. Explain the effect of the sample size increase on the mean and standard deviation of the sampling distribution.

## The central limit theorem

We have described the center and spread of the probability distribution of a sample mean  $\bar{x}$ , but not its shape. The shape of the distribution of  $\bar{x}$  depends on the shape of the population distribution. Here is one important case: if the population distribution is Normal, then so is the distribution of the sample mean.

### SAMPLING DISTRIBUTION OF A SAMPLE MEAN

If a population has the  $N(\mu, \sigma)$  distribution, then the sample mean  $\bar{x}$  of  $n$  independent observations has the  $N(\mu, \sigma/\sqrt{n})$  distribution.

This is a somewhat special result. Many population distributions are not Normal. The service call lengths in Figure 5.1(a), for example, are strongly skewed. Yet Figures 5.1(b) and 5.2 show that means of samples of size 80 are close to Normal. One of the most famous facts of probability theory says that, for large sample sizes, the distribution of  $\bar{x}$  is close to a Normal distribution. This is true no matter what shape the population distribution has, as long as the population has a finite standard deviation  $\sigma$ . This is the **central limit theorem**. It is much more useful than the fact that the distribution of  $\bar{x}$  is exactly Normal if the population is exactly Normal.

central limit theorem

## CENTRAL LIMIT THEOREM

Draw an SRS of size  $n$  from any population with mean  $\mu$  and finite standard deviation  $\sigma$ . When  $n$  is large, the sampling distribution of the sample mean  $\bar{x}$  is approximately Normal:

$$\bar{x} \text{ is approximately } N(\mu, \sigma/\sqrt{n})$$

### Example

#### 5.4 How close will the sample mean be to the population mean?

With the Normal distribution to work with, we can better describe how precisely a random sample of 80 calls estimates the mean length of all the calls in the population. The population standard deviation for the more than 30,000 calls in the population of Figure 5.1(a) is  $\sigma = 184.81$  seconds. From Example 5.3 we know  $\sigma_{\bar{x}} = 20.66$  seconds. By the 95 part of the 68–95–99.7 rule, about 95% of all samples will have mean  $\bar{x}$  within two standard deviations of  $\mu$  that is, with  $\pm 41.32$  seconds of  $\mu$ .

## USE YOUR KNOWLEDGE

#### 5.4 Use the 68–95–99.7 rule.

You take an SRS of size 49 from a population with mean 185 and standard deviation 70. According to the central limit theorem, what is the approximate sampling distribution of the sample mean? Use the 95 part of the 68–95–99.7 rule to describe the variability of  $\bar{x}$ .

For the sample size of  $n = 80$  in Example 5.4, the sample mean is not very precise. The population of service call lengths is very spread out, so the sampling distribution of  $\bar{x}$  has a large standard deviation.

## Example

### 5.5 How can we reduce the standard deviation?

In the setting of Example 5.4, if we want to reduce the standard deviation of  $\bar{x}$  by a factor of 4, we must take a sample 16 times as large,  $n = 16 \times 80$ , or 1280. Then

$$\sigma_{\bar{x}} = 184.81 / \sqrt{1280} = 5.166 \text{ seconds}$$

For samples of size 1280, about 95% of the sample means will be within twice 5.166, or 10.33 seconds, of the population mean  $\mu$ .

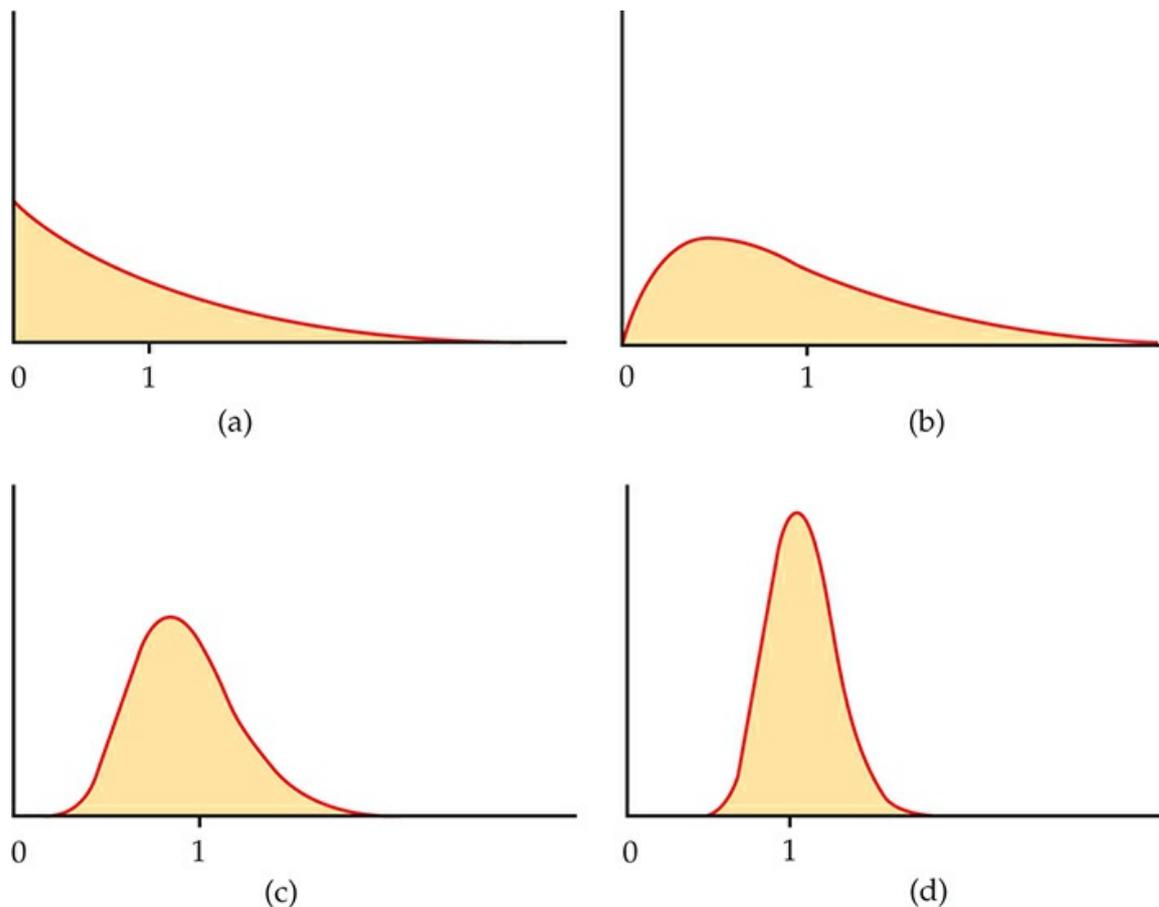
## USE YOUR KNOWLEDGE

### 5.5 The effect of increasing the sample size.

In the setting of Exercise 5.4, suppose that we increase the sample size to 1225. Use the 95 part of the 68–95–99.7 rule to describe the variability of this sample mean. Compare your results with those you found in Exercise 5.4.

Example 5.5 reminds us that if the population is very spread out, the  $n$  in the standard deviation of  $\bar{x}$  implies that very large samples are needed to estimate the population mean precisely. The main point of the example, however, is that the central limit theorem allows us to use Normal probability calculations to answer questions about sample means even when the population distribution is not Normal.

How large a sample size  $n$  is needed for  $\bar{x}$  to be close to Normal depends on the population distribution. More observations are required if the shape of the population distribution is far from Normal. For the very skewed call length population, samples of size 80 are large enough. Further study would be needed to see if the distribution of  $\bar{x}$  is close to Normal for smaller samples like  $n = 20$  or  $n = 40$ . Here is a more detailed study of another skewed distribution.



**FIGURE 5.3**

The central limit theorem in action: the sampling distribution of sample means from a strongly non-Normal population becomes more Normal as the sample size increases. (a) The distribution of 1 observation. (b) The distribution of  $\bar{x}$  for 2 observations. (c) The distribution of  $\bar{x}$  for 10 observations. (d) The distribution of  $\bar{x}$  for 25 observations.

## Example

### 5.6 The central limit theorem in action.

Figure 5.3 shows the central limit theorem in action for another very non-Normal population. Figure 5.3(a) displays the density curve of a single observation from the population. The distribution is strongly right-skewed, and the most probable outcomes are near 0. The mean  $\mu$  of this distribution is 1, and its standard deviation  $\sigma$  is also 1. This particular continuous distribution is called an **exponential distribution**. Exponential distributions are used as models for how long an iOS device, for example, will last and for the time between text messages sent on your cell phone.

Figures 5.3(b), (c), and (d) are the density curves of the sample means of 2, 10, and 25 observations from this population. As  $n$  increases, the shape becomes more Normal. The mean remains at  $\mu = 1$ , but the standard deviation decreases, taking the value  $1/n$ . The density curve for 10 observations is still somewhat skewed to the right but already resembles a Normal curve having  $\mu = 1$  and  $\sigma = 1/10 = 0.32$ . The density curve for  $n = 25$  is yet more Normal. The contrast between the shape of the population distribution and of the distribution of the mean of 10 or 25 observations is striking.

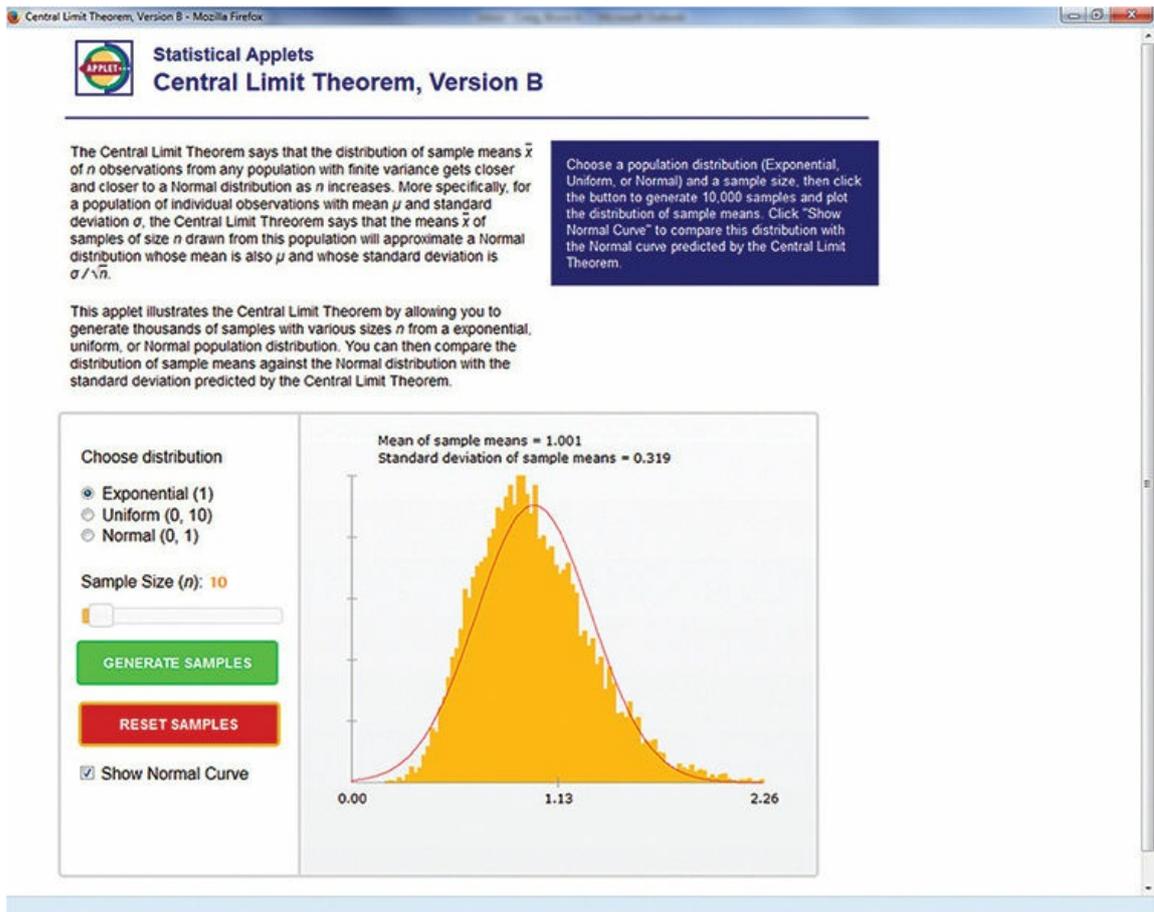


You can also use the *Central Limit Theorem* applet to study the sampling distribution of  $\bar{x}$ . From one of three population distributions, 10,000 SRSs of a user-specified sample size  $n$  are generated, and a histogram of the sample means is constructed. You can then compare this estimated sampling distribution with the Normal curve that is based on the central limit theorem.

## Example

### 5.7 Using the *Central Limit Theorem* applet.

In Example 5.6, we considered sample sizes of  $n = 2, 10,$  and  $25$  from an exponential distribution. Figure 5.4 shows a screenshot of the *Central Limit Theorem* applet for the exponential distribution when  $n = 10$ . The mean and standard deviation of this sampling distribution are 1 and  $1/10 = 0.316$ , respectively. From the 10,000 SRSs, the mean is estimated to be 1.001 and the estimated standard deviation is 0.319. These are both quite close to the true values. In Figure 5.3(c) we saw that the density curve for 10 observations is still somewhat skewed to the right. We can see this same behavior in Figure 5.4 when we compare the histogram with the Normal curve based on the central limit theorem.



**FIGURE 5.4**

Screenshot of the *Central Limit Theorem* applet for the exponential distribution when  $n = 10$ , for Example 5.7.

Try using the applet for the other sample sizes in Example 5.6. You should get histograms shaped like the density curves shown in Figure 5.3. You can also consider other sample sizes by sliding  $n$  from 1 to 100. As you increase  $n$ , the shape of the histogram moves closer to the Normal curve that is based on the central limit theorem.

## USE YOUR KNOWLEDGE

### 5.6 Use the *Central Limit Theorem* applet.



Let's consider the uniform distribution between 0 and 10. For this distribution, all intervals of the same length between 0 and 10 are

equally likely. This distribution has a mean of 5 and standard deviation of 2.89.

- (a) Approximate the population distribution by setting  $n = 1$  and clicking the “Generate samples” button.
- (b) What are your estimates of the population mean and population standard deviation based on the 10,000 SRSs? Are these population estimates close to the true values?
- (c) Describe the shape of the histogram and compare it with the Normal curve.

### **5.7 Use the *Central Limit Theorem* applet again.**

Refer to the previous exercise. In the setting of Example 5.6, let’s approximate the sampling distribution for samples of size  $n = 2, 10,$  and 25 observations.

- (a) For each sample size, compute the mean and standard deviation of  $\bar{x}$ .
- (b) For each sample size, use the applet to approximate the sampling distribution. Report the estimated mean and standard deviation. Are they close to the true values calculated in (a)?
- (c) For each sample size, compare the shape of the sampling distribution with the Normal curve based on the central limit theorem.
- (d) For this population distribution, what sample size do you think is needed to make you feel comfortable using the central limit theorem to approximate the sampling distribution of  $\bar{x}$ ? Explain your answer.

Now that we know that the sampling distribution of the sample mean  $\bar{x}$  is approximately Normal for a sufficiently large  $n$  let’s consider some probability calculations.

## **Example**

### **5.8 Time between sent text messages.**

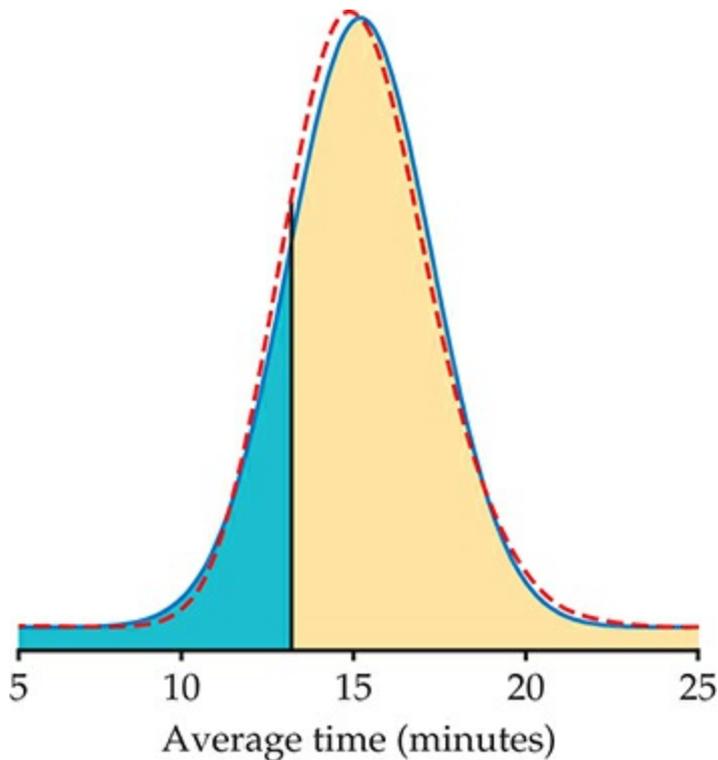
Americans aged 18 to 29 years send an average of almost 88 text messages a day.<sup>4</sup> Suppose that the time  $X$  between text messages sent from your cell phone is governed by the exponential distribution with mean  $\mu = 15$  minutes and standard deviation  $\sigma = 15$  minutes. You record the next 50 times between sent text messages. What is the probability that their average exceeds 13 minutes?

The central limit theorem says that the sample mean time  $\bar{x}$  (in minutes)

between text messages has approximately the Normal distribution with mean equal to the population mean  $\mu = 15$  minutes and standard deviation

$$\sigma = \sqrt{1550} = 39.12 \text{ minutes}$$

The sampling distribution of  $\bar{x}$  is therefore approximately  $N(15, 2, 12)$ . Figure 5.5 shows this Normal curve (solid) and also the actual density curve of  $\bar{x}$  (dashed).



**FIGURE 5.5**

The exact distribution (dashed) and the Normal approximation from the central limit theorem (solid) for the average time between text messages sent on your cell phone, for Example 5.8.

The probability we want is  $P(\bar{x} > 13.0)$ . This is the area to the right of 13 under the solid Normal curve in Figure 5.5. A Normal distribution calculation gives

$$\begin{aligned} P(\bar{x} > 13.0) &= P(\bar{x} - 15.2 > 13.0 - 15.2) \\ &= P(Z > -0.942) = 0.8264 \end{aligned}$$

The exactly correct probability is the area under the dashed density curve in the figure. It is 0.8271. The central limit theorem Normal approximation is off by only about 0.0007.

We can also use this sampling distribution to talk about the total time between the 1st and 51st text message sent from your phone.

## Example

### 5.9 Convert the results to the total time.

There are 50 time intervals between the 1st and 51st text message. According to the central limit theorem calculations in Example 5.8,

$$P(\bar{x} > 13.0) = 0.8264$$

We know that the sample mean is the total time divided by 50, so the event  $\{\bar{x} > 13.0\}$  is the same as the event  $\{50\bar{x} > 50(13.0)\}$ . We can say that the probability is 0.8264 that the total time is  $50(13.0) = 650$  minutes (10.8 hours) or greater.

## USE YOUR KNOWLEDGE

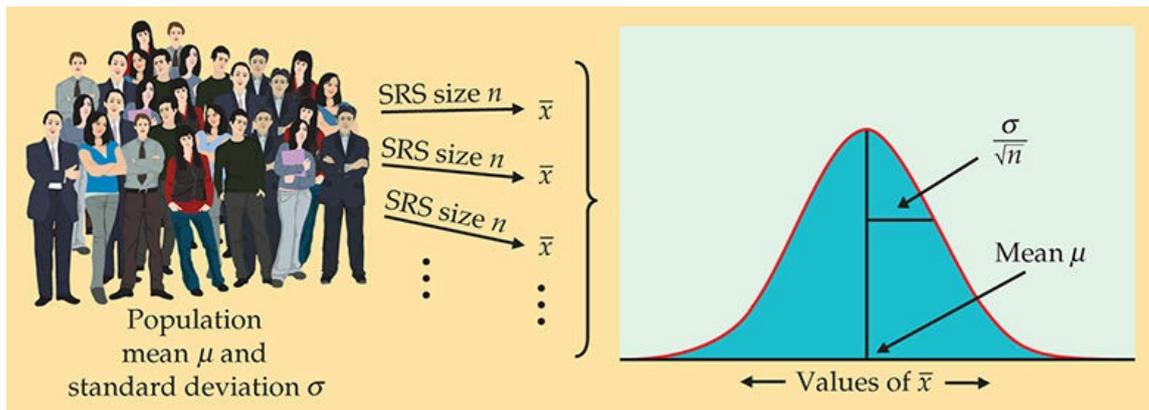
### 5.8 Find a probability.

Refer to Example 5.8. Find the probability that the mean time between text messages is less than 16 minutes. The exact probability is 0.6944. Compare your answer with the exact one.

Figure 5.6 summarizes the facts about the sampling distribution of  $\bar{x}$  in a way that emphasizes the big idea of a sampling distribution. The general framework for constructing the sampling distribution of  $\bar{x}$  is shown on the left.

- **Take many random samples of size  $n$  from a population with mean  $\mu$  and standard deviation  $\sigma$ .**
- **Find the sample mean  $\bar{x}$  for each sample.**
- **Collect all the  $\bar{x}$ 's and display their distribution.**

The sampling distribution of  $\bar{x}$  is shown on the right. Keep this figure in mind as you go forward.



**FIGURE 5.6**

The sampling distribution of a sample mean  $\bar{x}$  has mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$ . The sampling distribution is Normal if the population distribution is Normal; it is approximately Normal for large samples in any case.

## A few more facts

The central limit theorem is the big fact of probability theory in this section. Here are three additional facts related to our investigations that will be useful in describing methods of inference in later chapters.

### ← LOOK BACK

rules for means, p. 272

rules for variances, p. 275

The fact that the sample mean of an SRS from a Normal population has a Normal distribution is a special case of a more general fact: **any linear combination of independent Normal random variables is also Normally distributed.** That is, if  $X$  and  $Y$  are independent Normal random variables and  $a$  and  $b$  are any fixed numbers,  $aX + bY$  is also Normally distributed, and this is true for any number of Normal random variables. In particular, the sum or difference of independent Normal random variables has a Normal distribution. The mean and standard deviation of  $aX + bY$  are found as usual from the rules for means and variances. These facts are often used in statistical calculations. Here is an example.

### Example

#### 5.10 Getting to and from campus.

You live off campus and take the shuttle, provided by your apartment

complex, to and from campus. Your time on the shuttle in minutes varies from day to day. The time going to campus  $X$  has the  $N(20, 4)$  distribution, and the time returning from campus  $Y$  varies according to the  $N(18, 8)$  distribution. If they vary independently, what is the probability that you will be on the shuttle for less time going to campus?

The difference in times  $X - Y$  is Normally distributed, with mean and variance

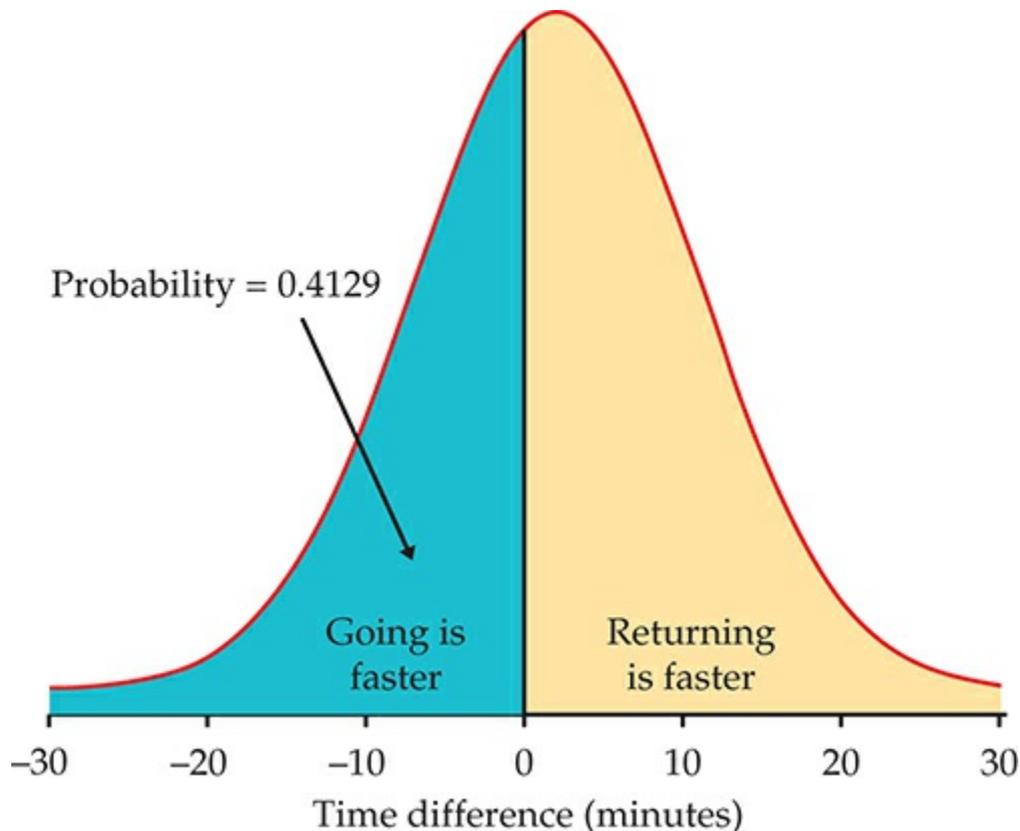
$$\mu_{X-Y} = \mu_X - \mu_Y = 20 - 18 = 2$$

$$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 = 4 + 8 = 12$$

Because  $\sigma_{X-Y} = 3.46$ ,  $X - Y$  has the  $N(2, 12)$  distribution. Figure 5.7 illustrates the probability computation:

$$\begin{aligned} P(X < Y) &= P(X - Y < 0) \\ &= P((X - Y) - 2 < 0 - 2) \\ &= P(Z < -0.58) = 0.2810 \end{aligned}$$

Although on average it takes longer to go to campus than return, the trip to campus will take less time on roughly two of every five days.



**FIGURE 5.7**

The Normal probability calculation for Example 5.10. The difference in times going to campus and returning from campus ( $X - Y$ ) is Normal with mean 2 minutes and standard deviation 3.46

minutes.

The second useful fact is that **more general versions of the central limit theorem say that the distribution of a sum or average of many small random quantities is close to Normal.** This is true even if the quantities are not independent (as long as they are not too highly correlated) and even if they have different distributions (as long as no single random quantity is so large that it dominates the others). These more general versions of the central limit theorem suggest why the Normal distributions are common models for observed data. Any variable that is a sum of many small random influences will have approximately a Normal distribution.

Finally, **the central limit theorem also applies to discrete random variables.** An average of discrete random variables will never result in a continuous sampling distribution, but the Normal distribution often serves as a good approximation. In Section 5.2, we will discuss the sampling distribution and Normal approximation for counts and proportions. This Normal approximation is just an example of the central limit theorem applied to these discrete random variables.

---

## BEYOND THE BASICS

### Weibull distributions

Our discussion of sampling distributions so far has concentrated on the Normal model to approximate the sampling distribution of the sample mean  $\bar{x}$ . This model is important in statistical practice because of the central limit theorem and the fact that sample means are among the most frequently used statistics. Simplicity also contributes to its popularity. The parameter  $\mu$  is easy to understand, and to estimate it, we use a statistic  $\bar{x}$  that is also easy to understand and compute.

There are, however, many other probability distributions that are used to model data in various circumstances. The time that a product, such as a computer hard drive, lasts before failing rarely has a Normal distribution. Earlier we mentioned the use of the exponential distribution to model time to failure. Another class of continuous distributions, the **Weibull distributions**, is more commonly used in these situations.

### Weibull distributions

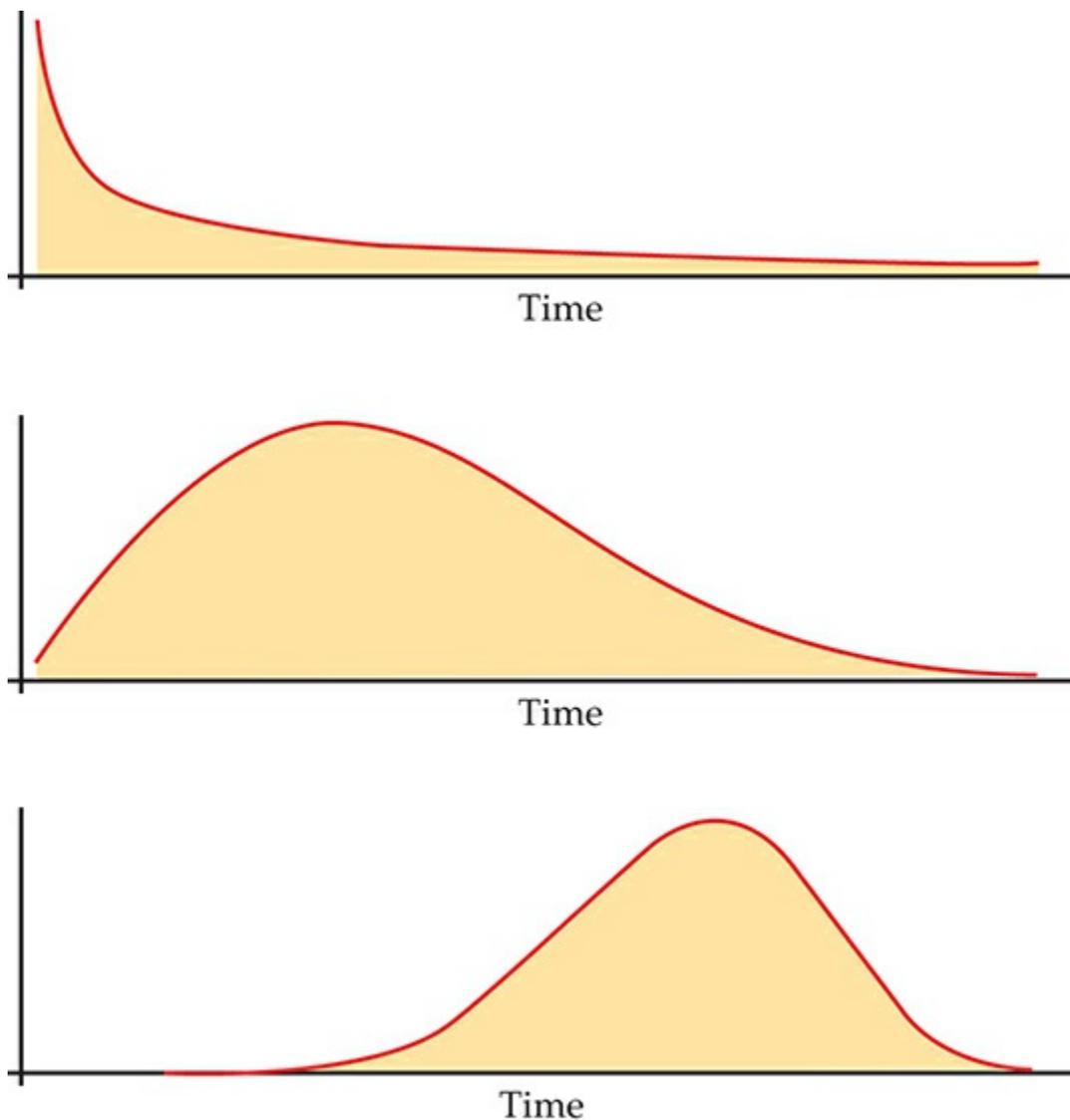
---

## Example

### 5.11 Weibull density curves.

Figure 5.8 shows the density curves of three members of the Weibull family. Each describes a different type of distribution for the time to failure of a product.

1. The top curve in Figure 5.8 is a model for *infant mortality*. This describes products that often fail immediately, prior to delivery to the customer. However, if the product does not fail right away, it will likely last a long time. For products like this, a manufacturer might test them and ship only the ones that do not fail immediately.



**FIGURE 5.8**

Density curves for three members of the Weibull family of distributions, for Example 5.11.

2. The middle curve in Figure 5.8 is a model for *early failure*. These products do not fail immediately, but many fail early in their lives after they are in the hands of customers. This is disastrous—the product or the process that makes it must be changed at once.
3. The bottom curve in Figure 5.8 is a model for *old-age wear-out*. Most of these products fail only when they begin to wear out, and then many fail at about the same age.

A manufacturer certainly wants to know to which of these classes a new product belongs. To find out, engineers operate a random sample of products until they fail. From the failure time data we can estimate the parameter (called the “shape parameter”) that distinguishes among the three Weibull distributions in Figure 5.8. The shape parameter has no simple definition like that of a population proportion or mean, and it cannot be estimated by a simple statistic such as  $\hat{p}$  or  $\bar{x}$ .

Two things save the situation. First, statistical theory provides general approaches for finding good estimates of any parameter. These general methods not only tell us how to use  $\bar{x}$  in the Normal settings but also tell us how to estimate the Weibull shape parameter. Second, software can calculate the estimate from data even though there is no algebraic formula that we can write for the estimate. Statistical practice often relies on both mathematical theory and methods of computation more elaborate than the ones we will meet in this book. Fortunately, big ideas such as sampling distributions carry over to more complicated situations.<sup>5</sup>

## SECTION 5.1 Summary

The **sample mean**  $\bar{x}$  of an SRS of size  $n$  drawn from a large population with mean  $\mu$  and standard deviation  $\sigma$  has a sampling distribution with mean and standard deviation

$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The sample mean  $\bar{x}$  is an unbiased estimator of the population mean  $\mu$  and is less variable than a single observation. The standard deviation decreases in proportion to the square root of the sample size  $n$ . This means that to reduce the standard deviation by a factor of  $C$ , we need to increase the sample size by a factor of  $C^2$ .

The **central limit theorem** states that for large  $n$  the sampling distribution of  $\bar{x}$  is approximately  $N(\mu, \sigma/\sqrt{n})$  for any population with mean  $\mu$  and finite standard deviation  $\sigma$ . This allows us to approximate probability calculations about  $\bar{x}$  using

the Normal distribution.

Linear combinations of independent Normal random variables have Normal distributions. In particular, if the population has a Normal distribution, so does  $\bar{x}$ .

## SECTION 5.1 Exercises

*For Exercise 5.1, see page 303; for Exercises 5.2 and 5.3, see page 307; for Exercise 5.4, see page 308; for Exercise 5.5, see page 308; for Exercises 5.6 and 5.7, see pages 310–311; and for Exercise 5.8, see page 312.*

### 5.9 What is wrong?

Explain what is wrong in each of the following statements.

- (a) If the population standard deviation is 20, then the standard deviation of  $\bar{x}$  for an SRS of 10 observations will be  $20/10 = 2$ .
- (b) When taking SRSs from a large population, larger sample sizes will result in larger standard deviations of  $\bar{x}$ .
- (c) For an SRS from a large population, both the mean and the standard deviation of  $\bar{x}$  depend on the sample size  $n$ .

### 5.10 What is wrong?

Explain what is wrong in each of the following statements.

- (a) The central limit theorem states that for large  $n$ , the population mean  $\mu$  is approximately Normal.
- (b) For large  $n$ , the distribution of observed values will be approximately Normal.
- (c) For sufficiently large  $n$ , the 68–95–99.7 rule says that  $\bar{x}$  should be within  $\mu \pm 2\sigma$  about 95% of the time.

### 5.11 Generating a sampling distribution.

Let's illustrate the idea of a sampling distribution in the case of a very small sample from a very small population. The population is the 10 scholarship players currently on your women's basketball team. For convenience, the 10 players have been labeled with the integers 0 to 9. For each player, the total amount of time spent (in minutes) on Facebook during the last week is recorded in the table below.

Player	0	1	2	3	4	5	6	7	8	9
Total time (min)	108	63	127	210	92	88	161	133	105	168

The parameter of interest is the average amount of time on Facebook. The sample is an SRS of size  $n = 3$  drawn from this population of players. Because the players are labeled 0 to 9, a single random digit from Table B chooses one player for the sample.

- (a) Find the mean for the 10 players in the population. This is the population mean  $\mu$ .
- (b) Use Table B to draw an SRS of size 3 from this population. (*Note:* You may sample the same player's time more than once.) Write down the three times in your sample and calculate the sample mean  $\bar{x}$ . This

statistic is an estimate of  $\mu$ .

(c) Repeat this process 9 more times using different parts of Table B. Make a histogram of the 10 values of  $\bar{x}$ . You are approximating the sampling distribution of  $\bar{x}$ .

(d) Is the center of your histogram close to  $\mu$ ? Explain why you'd expect it to get closer to  $\mu$  the more times you repeated this sampling process.

## 5.12 Number of apps on a Smartphone.

At a recent Appnation conference, Nielsen reported an average of 41 apps per smartphone among U.S. smartphone subscribers.<sup>6</sup> State the population for this survey, the statistic, and some likely values from the population distribution.

## 5.13 Why the difference?

Refer to the previous exercise. In Exercise 5.1 (page 303), a survey by AppsFire reported a median of 108 apps per device. This is very different from the average reported in the previous exercise.

(a) Do you think that the two populations are comparable? Explain your answer.

(b) The AppsFire report provides a footnote stating that their data exclude users who do not use any apps at all. Explain how this might contribute to the difference in the two reported statistics.

## 5.14 Total sleep time of college students.

In Example 5.1, the total sleep time per night among college students was approximately Normally distributed with mean  $\mu = 6.78$  hours and standard deviation  $\sigma = 1.24$  hours. You plan to take an SRS of size  $n = 150$  and compute the average total sleep time.

(a) What is the standard deviation for the average time?

(b) Use the 95 part of the 68–95–99.7 rule to describe the variability of this sample mean.

(c) What is the probability that your average will be below 6.9 hours?

## 5.15 Determining sample size.

Refer to the previous exercise. Now you want to use a sample size such that about 95% of the averages fall within  $\pm 10$  minutes (0.17 hours) of the true mean  $\mu = 6.78$ .

(a) Based on your answer to part (b) in Exercise 5.14, should the sample size be larger or smaller than 150? Explain.

(b) What standard deviation of  $\bar{x}$  do you need such that 95% of all samples will have a mean within 10 minutes of  $\mu$ ?

(c) Using the standard deviation you calculated in part (b), determine the number of students you need to sample.

## 5.16 File size on a tablet PC.

A tablet PC contains 8152 music and video files. The distribution of file size is highly skewed. Assume that the standard deviation for this population is 0.82 megabytes (MB).

- (a) What is the standard deviation of the average file size when you take an SRS of 16 files from this population?
- (b) How many files would you need to sample if you wanted the standard deviation of  $\bar{x}$  to be no larger than 0.10 MB?

### 5.17 Bottling an energy drink.

A bottling company uses a filling machine to fill cans with an energy drink. The cans are supposed to contain 250 milliliters (ml). The machine, however, has some variability, so the standard deviation of the volume is  $\sigma = 0.5$  ml. A sample of 4 cans is inspected each hour for process control purposes, and records are kept of the sample mean volume. If the process mean is exactly equal to the target value, what will be the mean and standard deviation of the numbers recorded?

### 5.18 Average file size on a tablet.

Refer to Exercise 5.16. Suppose that the true mean file size of the music and video files on the tablet is 7.4 MB and you plan to take an SRS of  $n = 40$  files.

- (a) Explain why it may be reasonable to assume that the average  $\bar{x}$  is approximately Normal even though the population distribution is highly skewed.
- (b) Sketch the approximate Normal curve for the sample mean, making sure to specify the mean and standard deviation.
- (c) What is the probability that your sample mean will differ from the population mean by more than 0.15 MB?

### 5.19 Can volumes.

Averages are less variable than individual observations. It is reasonable to assume that the can volumes in Exercise 5.17 vary according to a Normal distribution. In that case, the mean  $\bar{x}$  of an SRS of cans also has a Normal distribution.

- (a) Make a sketch of the Normal curve for a single can. Add the Normal curve for the mean of an SRS of 4 cans on the same sketch.
- (b) What is the probability that the volume of a single randomly chosen can differs from the target value by 1 ml or more?
- (c) What is the probability that the mean volume of an SRS of 4 cans differs from the target value by 1 ml or more?

### 5.20 Number of friends on Facebook.

Facebook recently examined all active Facebook users (more than 10% of the global population) and determined that the average user has 190 friends. This distribution takes only integer values, so it is certainly not Normal. It is also highly skewed to the right, with a median of 100 friends.<sup>7</sup> Suppose that  $\sigma = 288$  and you take an SRS of 70 Facebook users.

- (a) For your sample, what are the mean and standard deviation of  $\bar{x}$ , the mean number of friends per user?

(b) Use the central limit theorem to find the probability that the average number of friends for 70 Facebook users is greater than 250.

(c) What are the mean and standard deviation of the total number of friends in your sample?

(d) What is the probability that the total number of friends among your sample of 70 Facebook users is greater than 17,500?



### 5.21 Cholesterol levels of teenagers.

A study of the health of teenagers plans to measure the blood cholesterol level of an SRS of 13- to 16-year-olds. The researchers will report the mean  $\bar{x}$  from their sample as an estimate of the mean cholesterol level  $\mu$  in this population.

(a) Explain to someone who knows no statistics what it means to say that  $\bar{x}$  is an “unbiased” estimator of  $\mu$ .

(b) The sample result  $\bar{x}$  is an unbiased estimator of the population truth  $\mu$  no matter what size SRS the study chooses. Explain to someone who knows no statistics why a large sample gives more trustworthy results than a small sample.

### 5.22 ACT scores of high school seniors.

The scores of your state’s high school seniors on the ACT college entrance examination in a recent year had mean  $\mu = 22.3$  and standard deviation  $\sigma = 5.2$ . The distribution of scores is only roughly Normal.

(a) What is the approximate probability that a single student randomly chosen from all those taking the test scores 27 or higher?

(b) Now consider an SRS of 16 students who took the test. What are the mean and standard deviation of the sample mean score  $\bar{x}$  of these 16 students?

(c) What is the approximate probability that the mean score  $\bar{x}$  of these 16 students is 27 or higher?

(d) Which of your two Normal probability calculations in parts (a) and (c) is more accurate? Why?

### 5.23 Monitoring the emerald ash borer.

The emerald ash borer is a beetle that poses a serious threat to ash trees. Purple traps are often used to detect or monitor populations of this pest. In the counties of your state where the beetle is present, thousands of traps are used to monitor the population. These traps are checked periodically. The distribution of beetle counts per trap is discrete and strongly skewed. A majority of traps have no beetles, and only a few will have more than 1 beetle. For this exercise, assume that the mean number of beetles trapped is 0.3 with a standard deviation of 0.8.

(a) Suppose that your state does not have the resources to check all the traps, and so it plans to check only an SRS of  $n = 100$  traps. What are the mean and standard deviation of the average number of beetles  $\bar{x}$  in 100 traps?

(b) Use the central limit theorem to find the probability that the average number of beetles in 100 traps is greater than 0.5.

(c) Do you think it is appropriate in this situation to use the central limit theorem? Explain your answer.

### 5.24 Grades in a math course.

Indiana University posts the grade distributions for its courses online.<sup>8</sup> Students in one section of Math 118 in the fall 2012 semester received 33% A's, 33% B's, 20% C's, 12% D's, and 2% F's.

- (a) Using the common scale  $A = 4, B = 3, C = 2, D = 1, F = 0$ , take  $X$  to be the grade of a randomly chosen Math 118 student. Use the definitions of the mean (page 265) and standard deviation (page 273) for discrete random variables to find the mean  $\mu$  and the standard deviation  $\sigma$  of grades in this course.
- (b) Math 118 is a large enough course that we can take the grades of an SRS of 25 students to be independent of each other. If  $\bar{x}$  is the average of these 25 grades, what are the mean and standard deviation of  $\bar{x}$ ?
- (c) What is the probability that a randomly chosen Math 118 student gets a B or better,  $P(X \geq 3)$ ?
- (d) What is the approximate probability  $P(\bar{x} \geq 3)$  that the grade point average for 25 randomly chosen Math 118 students is B or better?

### 5.25 Diabetes during pregnancy.

Sheila's doctor is concerned that she may suffer from gestational diabetes (high blood glucose levels during pregnancy). There is variation both in the actual glucose level and in the results of the blood test that measures the level. A patient is classified as having gestational diabetes if her glucose level is above 140 milligrams per deciliter (mg/dl) one hour after a sugary drink is ingested. Sheila's measured glucose level one hour after ingesting the sugary drink varies according to the Normal distribution with  $\mu = 125$  mg/dl and  $\sigma = 10$  mg/dl.

- (a) If a single glucose measurement is made, what is the probability that Sheila is diagnosed as having gestational diabetes?
- (b) If measurements are made instead on three separate days and the mean result is compared with the criterion 140 mg/dl, what is the probability that Sheila is diagnosed as having gestational diabetes?

### 5.26 A roulette payoff.

A \$1 bet on a single number on a casino's roulette wheel pays \$35 if the ball ends up in the number slot you choose. Here is the distribution of the payoff  $X$ :

Payoff $X$	\$0	\$35
Probability	0.974	0.026

Each spin of the roulette wheel is independent of other spins.

- (a) What are the mean and standard deviation of  $X$ ?
- (b) Sam comes to the casino weekly and bets on 10 spins of the roulette wheel. What does the law of large numbers say about the average payoff Sam receives from his bets each visit?
- (c) What does the central limit theorem say about the distribution of Sam's average payoff after betting on 520 spins in a year?
- (d) Sam comes out ahead for the year if his average payoff is greater than \$1 (the amount he bet on each spin). What is the probability that Sam ends the year ahead? The true probability is 0.396. Does using the central limit theorem provide a reasonable approximation? We will return to this problem in the next section.

### 5.27 Defining a high glucose reading.

In Exercise 5.25, Sheila's measured glucose level one hour after ingesting the sugary drink varies according to the Normal distribution with  $\mu = 125$  mg/dl and  $\sigma = 10$  mg/dl. What is the level  $L$  such that there is probability only 0.05 that the mean glucose level of three test results falls above  $L$  for Sheila's glucose level distribution?

### 5.28 Risks and insurance.

The idea of insurance is that we all face risks that are unlikely but carry high cost. Think of a fire destroying your home. So we form a group to share the risk: we all pay a small amount, and the insurance policy pays a large amount to those few of us whose homes burn down. An insurance company looks at the records for millions of homeowners and sees that the mean loss from fire in a year is  $\mu = \$250$  per house and that the standard deviation of the loss is  $\sigma = \$1000$ . (The distribution of losses is extremely right-skewed: most people have \$0 loss, but a few have large losses.) The company plans to sell fire insurance for \$250 plus enough to cover its costs and profit.

- Explain clearly why it would be unwise to sell only 12 policies. Then explain why selling many thousands of such policies is a safe business.
- If the company sells 25,000 policies, what is the approximate probability that the average loss in a year will be greater than \$270?

### 5.29 Weights of airline passengers.

In response to the increasing weight of airline passengers, the Federal Aviation Administration told airlines to assume that passengers average 190 pounds in the summer, including clothing and carry-on baggage. But passengers vary: the FAA gave a mean but not a standard deviation. A reasonable standard deviation is 35 pounds. Weights are not Normally distributed, especially when the population includes both men and women, but they are not very non-Normal. A commuter plane carries 25 passengers. What is the approximate probability that the total weight of the passengers exceeds 5200 pounds? (*Hint:* To apply the central limit theorem, restate the problem in terms of the mean weight.)

### 5.30 Trustworthiness and eye color.

Various studies have shown that facial appearance affects social interactions. One recent study looked at the relationship between eye color and trustworthiness.<sup>9</sup> In this study, there were 238 participants, 78 with brown eyes and 160 with blue or green eyes. Each participant was asked to rate a set of student photos in terms of trustworthiness on a 10-point scale, where 1 means very trustworthy and 10 very untrustworthy. All photos showed a student who was seated in front of a white background and looking directly at the camera with a neutral expression. The photos were cropped so that the eyes were at the same height on each photo and a neckline was visible.

Suppose that for the population of all brown-eyed participants, a photo of a blue-eyed female student has a mean score of 5.8 and a standard deviation of 2.5. That same photo for the population of all blue- or green-eyed participants has a mean score of 6.3 and a standard deviation of 2.2.

- Although each participant's score is discrete, the mean score for each eye color group will be close to Normal. Why?
- What are the means and standard deviations of the sample means of the scores for the two eye color groups in this study?

### 5.31 Trustworthiness and eye color, continued.

Refer to the previous exercise.

- (a) We can take all 238 scores to be independent because participants are not told each other's scores. What is the distribution of the difference between the mean scores in the two groups?
- (b) Find the probability that the mean score for the brown-eyed group is less than the mean score for the other group.

### 5.32 Iron depletion without anemia and physical performance.

Several studies have shown a link between iron depletion without anemia (IDNA) and physical performance. In one recent study, the physical performance of 24 female collegiate rowers with IDNA was compared with 24 female collegiate rowers with normal iron status.<sup>10</sup> Several different measures of physical performance were studied, but we'll focus here on training-session duration. Assume that training-session duration of female rowers with IDNA is Normally distributed with mean 58 minutes and standard deviation 11 minutes. Training-session duration of female rowers with normal iron status is Normally distributed with mean 69 minutes and standard deviation 18 minutes.

- (a) What is the probability that the mean duration of the 24 rowers with IDNA exceeds 63 minutes?
- (b) What is the probability that the mean duration of the 24 rowers with normal iron status is less than 63 minutes?
- (c) What is the probability that the mean duration of the 24 rowers with IDNA is greater than the mean duration of the 24 rowers with normal iron status?



### 5.33 Treatment and control groups.

The previous exercise illustrates a common setting for statistical inference. This exercise gives the general form of the sampling distribution needed in this setting. We have a sample of  $n$  observations from a treatment group and an independent sample of  $m$  observations from a control group. Suppose that the response to the treatment has the  $N(\mu_X, \sigma_X)$  distribution and that the response of control subjects has the  $N(\mu_Y, \sigma_Y)$  distribution. Inference about the difference  $\mu_Y - \mu_X$  between the population means is based on the difference  $\bar{y} - \bar{x}$  between the sample means in the two groups.

- (a) Under the assumptions given, what is the distribution of  $\bar{y}$ ? Of  $\bar{x}$ ?
- (b) What is the distribution of  $\bar{y} - \bar{x}$ ?



### 5.34 Investments in two funds.

Jennifer invests her money in a portfolio that consists of 70% Fidelity 500 Index Fund and 30% Fidelity Diversified International Fund. Suppose that in the long run the annual real return  $X$  on the 500 Index Fund has mean 9% and standard deviation 19%, the annual real return  $Y$  on the Diversified International Fund has mean 11% and standard deviation 17%, and the correlation between  $X$  and  $Y$  is 0.6.

- (a) The return on Jennifer's portfolio is  $R = 0.7X + 0.3Y$ . What are the mean and standard deviation of  $R$ ?
- (b) The distribution of returns is typically roughly symmetric but with more extreme high and low observations than a Normal distribution. The average return over a number of years, however, is close to Normal. If Jennifer holds her portfolio for 20 years, what is the approximate probability that her average return is less than 5%?
- (c) The calculation you just made is not overly helpful, because Jennifer isn't really concerned about the mean return  $R$ . To see why, suppose that her portfolio returns 12% this year and 6% next year. The mean return for the two years is 9%. If Jennifer starts with \$1000, how much does she have at the end of the first

year? At the end of the second year? How does this amount compare with what she would have if both years had the mean return, 9%? Over 20 years, there may be a large difference between the ordinary mean  $\bar{R}$  and the *geometric mean*, which reflects the fact that returns in successive years multiply rather than add.

## 5.2 Sampling Distributions for Counts and Proportions

When you complete this section, you will be able to

- Determine when the count  $X$  can be modeled using the binomial distribution.
- Determine when the sampling distribution of  $X$  can be modeled using the binomial distribution.
- Calculate the mean and standard deviation of  $X$  when it has the  $B(n,p)$  distribution.
- Explain the differences in the sampling distributions of a count  $X$  and the associated sample proportion  $\hat{p}=X/n$ .
- Determine when one can utilize the Normal approximation to describe the sampling distribution of the count or the sampling distribution of the sample proportion.
- Use the Normal approximation for counts and proportions to perform probability calculations about the statistics.



**categorical variable**, p. 3

In the previous section, we discussed the probability distribution of the sample mean, which meant a focus on population values that were quantitative. We will now shift our focus to population values that are categorical. Counts and proportions are discrete statistics that describe categorical data. We focus our discussion on the simplest case of a random variable with only two possible categories. Here is an example.

### Example

**5.12 Work hours make it difficult to spend time with children.**



A sample survey asks 1006 British parents whether they think long working hours are making it difficult to spend enough time with their children.<sup>11</sup> We would like to view the responses of these parents as representative of a larger population of British parents who hold similar beliefs. That is, we will view the responses of the sampled parents as an SRS from a population.

When there are only two possible outcomes for a random variable, we can summarize the results by giving the count for one of the possible outcomes. We let  $n$  represent the sample size, and we use  $X$  to represent the random variable that gives the count for the outcome of interest.

## Example

### 5.13 The random variable of interest.

In our sample survey of British parents,  $n = 1006$ . We will ask each parent in our sample whether he or she feels long working hours make it difficult to spend enough time with their children. The variable  $X$  is the number of parents who think that long working hours make it difficult to spend enough time with their children. In this case,  $X = 755$ .

In our example, we chose the random variable  $X$  to be the number of parents who think that long working hours make it difficult to spend enough time with their children. We could have chosen  $X$  to be the number of parents who do not think that long working hours make it difficult to spend enough time with their children. The choice is yours. Often we make the choice based on how we would like to describe the results in a summary. Which choice do you prefer in this case?

When a random variable has only two possible outcomes, we can also use the **sample proportion**  $\hat{p} = X/n$  as a summary.

## Example

### 5.14 The sample proportion.

The sample proportion of parents surveyed who think that long working hours make it difficult to spend enough time with their children is

$$\hat{p} = \frac{755}{1006} = 0.75$$

Notice that this summary takes into account the sample size  $n$ . We need to know  $n$  in order to properly interpret the meaning of the random variable  $X$ . For example, the conclusion we would draw about parent opinions in this survey would be quite different if we had observed  $X = 755$  from a sample twice as large,  $n = 2012$ .

## USE YOUR KNOWLEDGE

### 5.35 Sexual harassment in middle school and high school.

A survey of 1965 students in grades 7 to 12 reports that 48% of the students say they have encountered some type of sexual harassment while at school.<sup>12</sup> Give  $n$ ,  $X$ , and  $\hat{p}$  for this survey.

### 5.36 Seniors who have taken a statistics course.

In a random sample of 300 senior students from your college, 63% reported that they had taken a statistics course. Give  $n$ ,  $X$ , and  $\hat{p}$  for this setting.

### 5.37 Use of the Internet to find a place to live.

A poll of 1500 college students asked whether or not they have used the Internet to find a place to live sometime within the past year. There were 1025 students who answered “Yes”; the other 475 answered “No.”

- (a) What is  $n$ ?
- (b) Choose one of the two possible outcomes to define the random variable,  $X$ . Give a reason for your choice.
- (c) What is the value of  $X$ ?
- (d) Find the sample proportion,  $\hat{p}$ .

Just like the sample mean, sample counts and sample proportions are commonly used statistics, and understanding their sampling distributions is important for statistical inference. These statistics, however, are discrete random variables and thus introduce us to a new family of probability distributions.

## The binomial distributions for sample counts

The distribution of a count  $X$  depends on how the data are produced. Here is a simple but common situation.

### THE BINOMIAL SETTING

1. There is a fixed number of observations  $n$
2. The  $n$  observations are all independent.
3. Each observation falls into one of just two categories, which for convenience we call “success” and “failure.”
4. The probability of a success, call it  $p$  is the same for each observation.

Think of tossing a coin  $n$  times as an example of the binomial setting. Each toss gives either heads or tails and the outcomes of successive tosses are independent. If we call heads a success, then  $p$  is the probability of a head and remains the same as long as we toss the same coin. The number of heads we count is a random variable  $X$ . The distribution of  $X$  (and, more generally, the distribution of the count of successes in any binomial setting) is completely determined by the number of observations  $n$  and the success probability  $p$

### BINOMIAL DISTRIBUTIONS

The distribution of the count  $X$  of successes in the binomial setting is called the **binomial distribution** with parameters  $n$  and  $p$ . The parameter  $n$  is the number of observations, and  $p$  is the probability of a success on any one

observation. The possible values of  $X$  are the whole numbers from 0 to  $n$ . As an abbreviation, we say that the distribution of  $X$  is  $B(n, p)$ .



The binomial distributions are an important class of discrete probability distributions. Later in this section we will learn how to assign probabilities to outcomes and how to find the mean and standard deviation of binomial distributions. That said, *the most important skill for using binomial distributions is the ability to recognize situations to which they do and do not apply*. This can be done by checking all the facets of the binomial setting.

## Example

### 5.15 Binomial examples?

(a) Genetics says that children receive genes from each of their parents independently. Each child of a particular pair of parents has probability 0.25 of having type O blood. If these parents have 3 children, the number who have type O blood is the count  $X$  of successes in 3 independent trials with probability 0.25 of a success on each trial. So  $X$  has the  $B(3, 0.25)$  distribution.

(b) Engineers define reliability as the probability that an item will perform its function under specific conditions for a specific period of time. Replacement heart valves made of animal tissue, for example, have probability 0.77 of performing well for 15 years.<sup>13</sup> The probability of failure within 15 years is therefore 0.23. It is reasonable to assume that valves in different patients fail (or not) independently of each other. The number of patients in a group of 500 who will need another valve replacement within 15 years has the  $B(500, 0.23)$  distribution.

(c) A multicenter trial is designed to assess a new surgical procedure. A total of 540 patients will undergo the procedure, and the count of patients  $X$  who suffer a major adverse cardiac event (MACE) within 30 days of surgery will be recorded. Because these patients will receive this procedure from different surgeons at different hospitals, it may not be true that the probability of a MACE is the same for each patient. Thus,  $X$  may not have the binomial distribution.

## USE YOUR KNOWLEDGE

### 5.38 Genetics and blood types.

Genetics says that children receive genes from each of their parents independently. Suppose that each child of a particular pair of parents has probability 0.5 of having type AB blood. If these parents have 4 children, what is the distribution of the number who have type AB blood? Explain your answer.

### 5.39 Toss a coin.

Toss a fair coin 10 times. Give the distribution of  $X$ , the number of heads that you observe.

## Binomial distributions in statistical sampling

The binomial distributions are important in statistics when we wish to make inferences about the proportion  $p$  of “successes” in a population. Here is a typical example.

## Example

### 5.16 Audits of financial records.



The financial records of businesses may be audited by state tax authorities to

test compliance with tax laws. It is too time-consuming to examine all sales and purchases made by a company during the period covered by the audit. Suppose that the auditor examines an SRS of 150 sales records out of 10,000 available. One issue is whether each sale was correctly classified as subject to state sales tax or not. Suppose that 800 of the 10,000 sales are incorrectly classified. Is the count  $X$  of misclassified records in the sample a binomial random variable?

*Choosing an SRS from a population is not quite a binomial setting.* Removing one record in Example 5.16 changes the proportion of bad records in the remaining population, so the state of the second record chosen is not independent of the first. Because the population is large, however, removing a few items has a very small effect on the composition of the remaining population. Successive inspection results are very nearly independent. The population proportion of misclassified records is

$$p = 800/10,000 = 0.08$$

If the first record chosen is bad, the proportion of bad records remaining is  $799/9999 = 0.079908$ . If the first record is good, the proportion of bad records left is  $800/9999 = 0.080008$ . These proportions are so close to 0.08 that for practical purposes we can act as if removing one record has no effect on the proportion of misclassified records remaining. We act as if the count  $X$  of misclassified sales records in the audit sample has the binomial distribution  $B(150, 0.08)$ .



**stratified sample**, p. 197

Populations like the one described in Example 5.16 often contain a relatively small number of items with very large values. For this example, these values would be very large sale amounts and likely represent an important group of items to the auditor. An SRS taken from such a population will likely include very few items of this type. Therefore, it is common to use a stratified sample in settings like this. Strata are defined based on dollar value of the sale, and within each stratum, an SRS is taken. The results are then combined to obtain an estimate for the entire population.

## SAMPLING DISTRIBUTION OF A COUNT

A population contains proportion  $p$  of successes. If the population is much larger than the sample, the count  $X$  of successes in an SRS of size  $n$  has approximately the binomial distribution  $B(n, p)$ .

The accuracy of this approximation improves as the size of the population increases relative to the size of the sample. As a rule of thumb, we will use the binomial sampling distribution for counts

when the population is at least 20 times as large as the sample.

## Finding binomial probabilities

We will later give a formula for the probability that a binomial random variable takes any of its values. In practice, you will rarely have to use this formula for calculations because some calculators and most statistical software packages will calculate binomial probabilities for you.

### Example

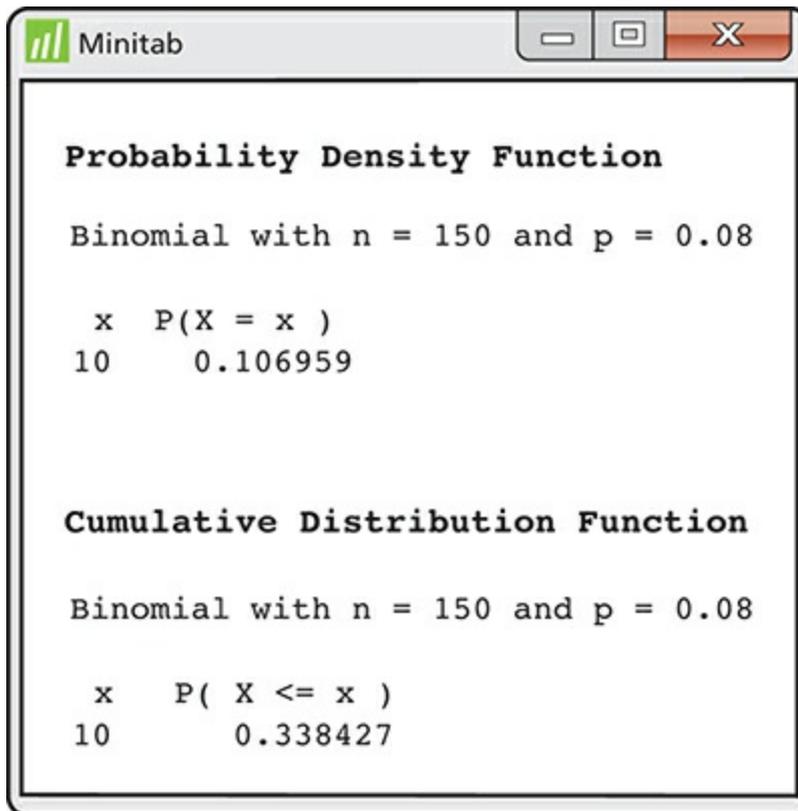
#### 5.17 Probabilities for misclassified sales records.

In the audit setting of Example 5.16, what is the probability that the audit finds exactly 10 misclassified sales records? What is the probability that the audit finds no more than 10 misclassified records? Figure 5.9 shows the output from one statistical software system. You see that if the count  $X$  has the  $B(150, 0.08)$  distribution,

$$P(X = 10) = 0.106959$$

$$P(X \leq 10) = 0.338427$$

It was easy to request these calculations in the software's menus. For the TI-83/84 calculator, the functions **binompdf** and **binomcdf** would be used. In R, the functions **dbinom** and **pbinom** would be used. Typically, the output supplies more decimal places than we need and uses labels that may not be helpful (for example, "Probability Density Function" when the distribution is discrete, not continuous). But, as usual with software, we can ignore distractions and find the results we need.



**FIGURE 5.9**

Binomial probabilities for Example 5.17: output from the Minitab statistical software.

If you do not have suitable computing facilities, you can still shorten the work of calculating binomial probabilities for some values of  $n$  and  $p$  by looking up probabilities in Table C in the back of this book. The entries in the table are the probabilities  $P(X = k)$  of individual outcomes for a binomial random variable  $X$ .

## Example

### 5.18 The probability histogram.

Suppose that the audit in Example 5.16 chose just 15 sales records. What is the probability that no more than 1 of the 15 is misclassified? The count  $X$  of misclassified records in the sample has approximately the  $B(15, 0.08)$  distribution. Figure 5.10 is a probability histogram for this distribution. The distribution is strongly skewed. Although  $X$  can take any whole-number value from 0 to 15, the probabilities of values larger than 5 are so small that they do not appear in the histogram.

We want to calculate

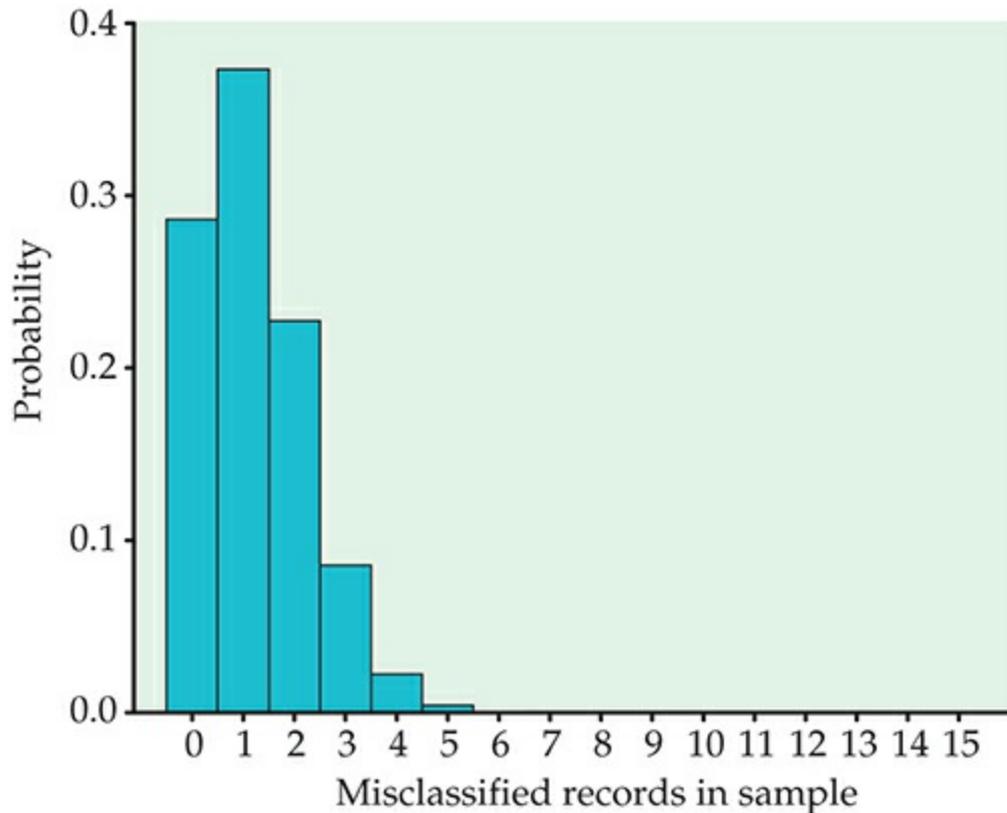
$$P(X \leq 1) = P(X = 0) + P(X = 1)$$

when  $X$  has the  $B(15, 0.08)$  distribution. To use Table C for this calculation, look opposite  $n = 15$  and under  $p = 0.08$ . The entries in the rows for each  $k$  are  $P(X = k)$ . Blank cells in the table are 0 to four decimal places. You see that

$$\begin{aligned} P(X \leq 1) &= P(X = 0) + P(X = 1) \\ &= 0.2863 + 0.3734 = 0.6597 \end{aligned}$$

About two-thirds of all samples will contain no more than 1 bad record. In fact, almost 29% of the samples will contain no bad records. The sample of size 15 cannot be trusted to provide adequate evidence about misclassified sales records. A larger number of observations is needed.

		<i>p</i>
<i>n</i>	<i>k</i>	<b>.08</b>
15	0	.2863
	1	.3734
	2	.2273
	3	.0857
	4	.0223
	5	.0043
	6	.0006
	7	.0001
	8	
	9	



**FIGURE 5.10**

Probability histogram for the binomial distribution with  $n = 15$  and  $p = 0.08$ , for Example 5.18.

The values of  $p$  that appear in Table C are all 0.5 or smaller. When the probability of a success is greater than 0.5, restate the problem in terms of the number of failures. The probability of a failure is less than 0.5 when the probability of a success exceeds 0.5. When using the table, always stop to ask whether you must count successes or failures.

## Example

### 5.19 Falling asleep in class.

In the survey of 4513 college students described in Example 5.1, 46% of the respondents reported falling asleep in class due to poor sleep. You randomly sample 12 students in your dormitory, and 9 state that they fell asleep in class during the last week due to poor sleep. Relative to the survey results, is this an unusually high number of students?

To answer this question, assume that the students' actions (falling asleep or not) are independent, with the probability of falling asleep equal to 0.46. This

independence assumption may not be reasonable if the students study and socialize together or if there is a loud student in the dormitory who keeps everyone up. We'll assume this is not an issue here, so the number  $X$  of students who fell asleep in class out of 12 students has the  $B(12, 0.46)$  distribution.

We want the probability of classifying at least 9 students as having fallen asleep in class. Using software, we find

$$\begin{aligned}P(X \geq 9) &= P(X = 9) + P(X = 10) + P(X = 11) + P(X = 12) \\ &= 0.0319 + 0.0082 + 0.0013 + 0.0001 = 0.0415\end{aligned}$$

We would expect to find 9 or more students falling asleep in class about 4% of the time, in fewer than 1 of every 20 surveys. This is a pretty rare outcome and falls outside the range of the usual chance variation due to random sampling.

## USE YOUR KNOWLEDGE

### 5.40 Free-throw shooting.

Courtney is a basketball player who makes 90% of her free throws. In a recent game, she had 10 free throws and missed 3 of them. How unusual is this outcome? Using software, calculator, or Table C, compute  $1 - P(X \leq 2)$ , where  $X$  is the number of free throws missed in 10 shots. Explain your answer.

### 5.41 Find the probabilities.

(a) Suppose that  $X$  has the  $B(6, 0.4)$  distribution. Use software, calculator, or Table C to find  $P(X = 0)$  and  $P(X \geq 4)$ .

(b) Suppose that  $X$  has the  $B(6, 0.6)$  distribution. Use software, calculator, or Table C to find  $P(X = 6)$  and  $P(X \leq 2)$ .

(c) Explain the relationship between your answers to parts (a) and (b) of this exercise.

## Binomial mean and standard deviation

If a count  $X$  has the  $B(n, p)$  distribution, what are the mean  $\mu_X$  and the standard deviation  $\sigma_X$ ? We can guess the mean. If we expect 46% of the students to have fallen asleep in class due to poor sleep, the mean number in 12 students should be 46% of 12, or 5.5. That's  $\mu_X$  when  $X$  has the  $B(12, 0.46)$  distribution.

 **LOOK BACK**

**means and variances of random variables**, p. 263

Intuition suggests more generally that the mean of the  $B(n, p)$  distribution should be  $np$ . Can we show that this is correct and also obtain a short formula for the standard deviation? Because binomial distributions are discrete probability distributions, we could find the mean and variance by using the definitions in Section 4.4. Here is an easier way.

A binomial random variable  $X$  is the count of successes in  $n$  independent observations that each have the same probability  $p$  of success. Let the random variable  $S_i$  indicate whether the  $i$ th observation is a success or failure by taking the values  $S_i = 1$  if a success occurs and  $S_i = 0$  if the outcome is a failure. The  $S_i$  are independent because the observations are, and each  $S_i$  has the same simple distribution:

Outcome	1	0
Probability	$p$	$1-p$

 **LOOK BACK**

**mean and variance of a discrete random variable**, p. 279

From the definition of the mean of a discrete random variable, we know that the mean of each  $S_i$  is

$$\mu_s = (1)(p) + (0)(1-p) = p$$

Similarly, the definition of the variance shows that  $\sigma_{S_i}^2 = p(1-p)$ . Because each  $S_i$  is 1 for a success and 0 for a failure, to find the total number of successes  $X$  we add the  $S_i$ 's:

$$X = S_1 + S_2 + \dots + S_n$$

Apply the addition rules for means and variances to this sum. To find the mean of  $X$  we add the means of the  $S_i$ 's:

$$\begin{aligned}\mu_X &= \mu_{S_1} + \mu_{S_2} + \dots + \mu_{S_n} \\ &= n\mu_S = np\end{aligned}$$

Similarly, the variance is  $n$  times the variance of a single  $S$ , so that  $\sigma_X^2 = np(1-p)$ . The standard deviation  $\sigma_X$  is the square root of the variance. Here is the result.

## **BINOMIAL MEAN AND STANDARD DEVIATION**

If a count  $X$  has the binomial distribution  $B(n, p)$ , then

$$\mu_X = np$$

$$\sigma_X = np(1-p)$$

## Example

### 5.20 The Helsinki Heart Study.

The Helsinki Heart Study asked whether the anticholesterol drug gemfibrozil reduces heart attacks. In planning such an experiment, the researchers must be confident that the sample sizes are large enough to enable them to observe enough heart attacks. The Helsinki study planned to give gemfibrozil to about 2000 men aged 40 to 55 and a placebo to another 2000. The probability of a heart attack during the five-year period of the study for men this age is about 0.04. What are the mean and standard deviation of the number of heart attacks that will be observed in one group if the treatment does not change this probability?

There are 2000 independent observations, each having probability  $p = 0.04$  of a heart attack. The count  $X$  of heart attacks has the  $B(2000, 0.04)$  distribution, so that

$$\mu_X = np = (2000)(0.04) = 80$$

$$\sigma_X = np(1-p) = (2000)(0.04)(0.96) = 8.76$$

The expected number of heart attacks is large enough to permit conclusions about the effectiveness of the drug. In fact, there were 84 heart attacks among the 2035 men actually assigned to the placebo, quite close to the mean. The gemfibrozil group of 2046 men suffered only 56 heart attacks. This is evidence that the drug reduces the chance of a heart attack. In a later chapter we will learn how to determine if this is strong enough evidence to conclude that the drug is effective.

## Sample proportions

What proportion of a company's sales records have an incorrect sales tax classification? What percent of adults favor stronger laws restricting firearms? In

statistical sampling we often want to estimate the **proportion**  $p$  of “successes” in a population. Our estimator is the sample proportion of successes:

proportion

$$\hat{p} = \frac{\text{count of successes in sample}}{\text{size of sample}} = \frac{X}{n}$$



*Be sure to distinguish between the proportion  $\hat{p}$  and the count  $X$ .* The count takes whole-number values between 0 and  $n$ , but a proportion is always a number between 0 and 1. In the binomial setting, the count  $X$  has a binomial distribution. The proportion  $\hat{p}$  does *not* have a binomial distribution. We can, however, do probability calculations about  $\hat{p}$  by restating them in terms of the count  $X$  and using binomial methods. In Example 5.9 (page 312) we took a similar approach for the sum, restating the problem in terms of the sample mean and then using the Normal distribution to calculate the probability.

## Example

### 5.21 Buying clothes online.



A survey by the Consumer Reports National Research Center revealed that 85% of all respondents were very or completely satisfied with their online clothes-shopping experience.<sup>14</sup> It was also reported, however, that people over the age of 40 were generally more satisfied than younger respondents. You decide to take a nationwide random sample of 2500 college students and ask if

they agree or disagree that “I am very or completely satisfied with my online clothes-shopping experience.” Suppose that 60% of all college students would agree if asked this question. What is the probability that the sample proportion who agree is at least 58%?

The count  $X$  who agree has the binomial distribution  $B(2500, 0.6)$ . The sample proportion  $\hat{p}=X/2500$  does *not* have a binomial distribution, because it is not a count. But we can translate any question about a sample proportion  $\hat{p}$  into a question about the count  $X$ . Because 58% of 2500 is 1450,

$$\begin{aligned} P(\hat{p} \geq 0.58) &= P(X \geq 1450) \\ &= P(X = 1450) + P(X = 1451) + \dots + P(X = 2500) \end{aligned}$$

This is a rather elaborate calculation. We must add more than 1000 binomial probabilities. Software tells us that  $P(\hat{p} \geq 0.58) = 0.9802$ . But what do we do if we don't have access to software?



**rules for means**, p. 272

**rules for variances**, p. 275

As a first step, find the mean and standard deviation of a sample proportion. We know the mean and standard deviation of a sample count, so apply the rules from Section 4.4 for the mean and variance of a constant times a random variable. Here is the result.

## MEAN AND STANDARD DEVIATION OF A SAMPLE PROPORTION

Let  $\hat{p}$  be the sample proportion of successes in an SRS of size  $n$  drawn from a large population having population proportion  $p$  of successes. The mean and standard deviation of  $\hat{p}$  are

$$\begin{aligned} \mu_{\hat{p}} &= p \\ \sigma_{\hat{p}} &= \sqrt{p(1-p)/n} \end{aligned}$$

The formula for  $\sigma_{\hat{p}}$  is exactly correct in the binomial setting. It is approximately correct for an SRS from a large population. We will use it when the population is at least 20 times as large as the sample.

Let's now use these formulas to calculate the mean and standard deviation for Example 5.21.

## Example

### 5.22 The mean and the standard deviation.

The mean and standard deviation of the proportion of the survey respondents in Example 5.21 who are satisfied with their online clothes-shopping experience are

$$\mu_{\hat{p}} = p = 0.6$$

$$\sigma_{\hat{p}} = \sqrt{p(1-p)/n} = \sqrt{(0.6)(0.4)/2500} = 0.0098$$

## USE YOUR KNOWLEDGE

### 5.42 Find the mean and the standard deviation.

If we toss a fair coin 200 times, the number of heads is a random variable that is binomial.

- Find the mean and the standard deviation of the sample proportion of heads.
- Is your answer to part (a) the same as the mean and the standard deviation of the sample count of heads? Explain your answer.

## LOOK BACK

**unbiased estimator**, p. 210

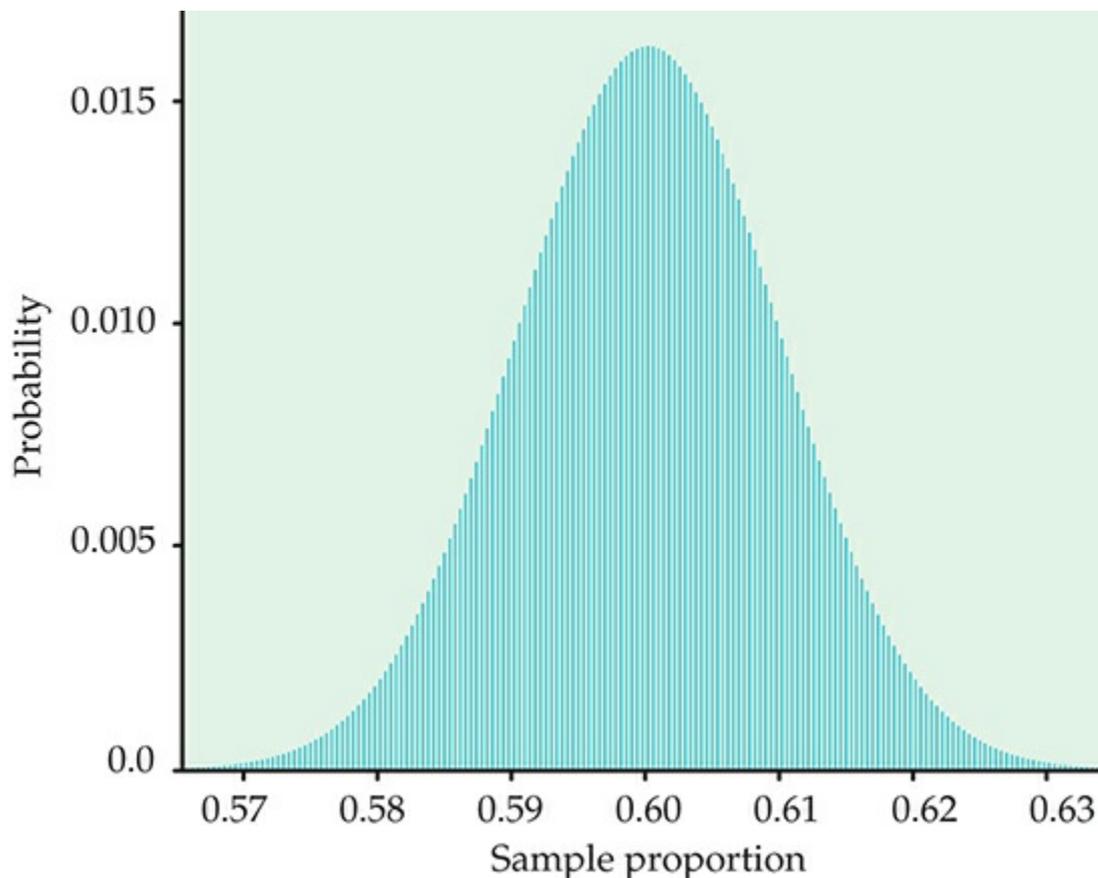
The fact that the mean of  $\hat{p}$  is  $p$  states in statistical language that the sample proportion  $\hat{p}$  in an SRS is an *unbiased estimator* of the population proportion  $p$ . When a sample is drawn from a new population having a different value of the population proportion  $p$ , the sampling distribution of the unbiased estimator  $\hat{p}$  changes so that its mean moves to the new value of  $p$ . We observed this fact empirically in Section 3.4 and have now verified it from the laws of probability.

The variability of  $\hat{p}$  about its mean, as described by the variance or standard deviation, gets smaller as the sample size increases. So a sample proportion from a large sample will usually lie quite close to the population proportion  $p$ . We observed this in the simulation experiment on page 208 in Section 3.4. Now we have discovered exactly how the variability decreases: the standard deviation is  $\sqrt{p(1-p)/n}$ . Similar to what we observed in the previous section, the  $n$  in the

denominator means that the sample size must be multiplied by 4 if we wish to divide the standard deviation in half.

## Normal approximation for counts and proportions

Using simulation, we discovered in Section 3.4 that the sampling distribution of a sample proportion  $\hat{p}$  is close to Normal. Now we know that the distribution of  $\hat{p}$  is that of a binomial count divided by the sample size  $n$ . This seems at first to be a contradiction. To clear up the matter, look at Figure 5.11. This is a probability histogram of the exact distribution of the proportion of frustrated shoppers  $\hat{p}$  based on the binomial distribution  $B(2500, 0.6)$ . There are hundreds of narrow bars, one for each of the 2501 possible values of  $\hat{p}$ . Most have probabilities too small to show in a graph. *The probability histogram looks very Normal!* In fact, both the count  $X$  and the sample proportion  $\hat{p}$  are approximately Normal in large samples.



**FIGURE 5.11**

Probability histogram of the sample proportion  $\hat{p}$  based on a binomial count with  $n = 2500$  and  $p = 0.6$ . The distribution is very close to Normal.

 **LOOK BACK**  
central limit theorem, p. 307

We also know this to be true as a result of the central limit theorem discussed in the previous section. Recall that we can consider the count  $X$  as a sum

$$X = S_1 + S_2 + \dots + S_n$$

of independent random variables  $S_i$  that take the value 1 if a success occurs on the  $i$ th trial and the value 0 otherwise. The proportion of successes  $\hat{p} = X/n$  can then be thought of as the sample mean of the  $S_i$  and, like all sample means, is approximately Normal when  $n$  is large. Given that  $\hat{p}$  is approximately Normal, the count will also be approximately Normal since it is just a constant  $n$  times  $\hat{p}$ , an approximately Normal random variable.

## NORMAL APPROXIMATION FOR COUNTS AND PROPORTIONS

Draw an SRS of size  $n$  from a large population having population proportion  $p$  of successes. Let  $X$  be the count of successes in the sample and  $\hat{p} = X/n$  be the sample proportion of successes. When  $n$  is large, the sampling distributions of these statistics are approximately Normal:

$$X \text{ is approximately } N(np, np(1-p))$$

$$\hat{p} \text{ is approximately } N(p, p(1-p)/n)$$

As a rule of thumb, we will use this approximation for values of  $n$  and  $p$  that satisfy  $np \geq 10$  and  $n(1-p) \geq 10$ .

These Normal approximations are easy to remember because they say that  $\hat{p}$  and  $X$  are Normal, with their usual means and standard deviations. Whether or not you use the Normal approximations should depend on how accurate your calculations need to be. For most statistical purposes great accuracy is not required. Our “rule of thumb” for use of the Normal approximations reflects this judgment.

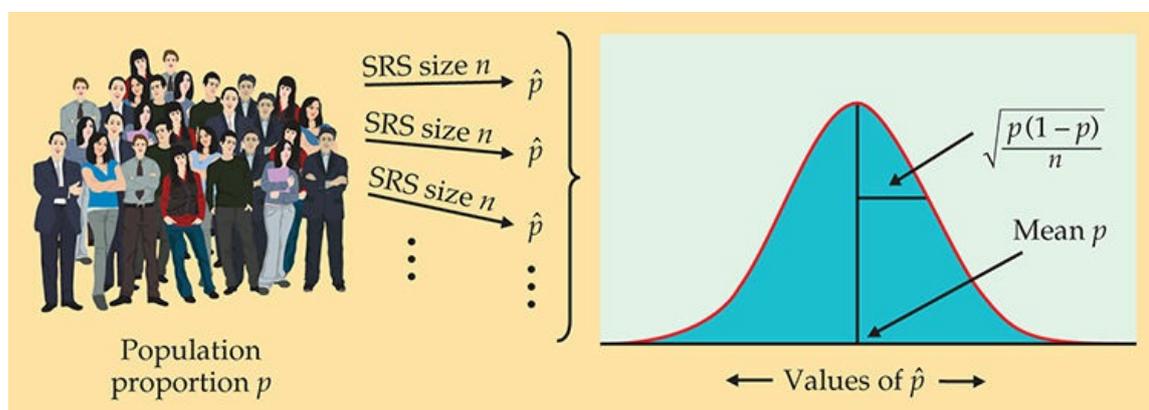


FIGURE 5.12

The sampling distribution of a sample proportion  $\hat{p}$  is approximately Normal with mean  $p$  and standard deviation  $\sqrt{p(1-p)/n}$ .



The accuracy of the Normal approximations improves as the sample size  $n$  increases. They are most accurate for any fixed  $n$  when  $p$  is close to  $1/2$ , and least accurate when  $p$  is near  $0$  or  $1$ . You can compare binomial distributions with their Normal approximations by using the *Normal Approximation to Binomial* applet. This applet allows you to change  $n$  or  $p$  while watching the effect on the binomial probability histogram and the Normal curve that approximates it.

Figure 5.12 summarizes the distribution of a sample proportion in a form that emphasizes the big idea of a sampling distribution. Just as with Figure 5.6, the general framework for constructing a sampling distribution is shown on the left.

- **Take many random samples of size  $n$  from a population that contains proportion  $p$  of successes.**
- **Find the sample proportion  $\hat{p}$  for each sample.**
- **Collect all the  $\hat{p}$ 's and display their distribution.**

The sampling distribution of  $\hat{p}$  is shown on the right. Keep this figure in mind as you move toward statistical inference.

## Example

### 5.23 Compare the Normal approximation with the exact calculation.

Let's compare the Normal approximation for the calculation of Example 5.21 with the exact calculation from software. We want to calculate  $P(\hat{p} \geq 0.58)$  when the sample size is  $n = 2500$  and the population proportion is  $p = 0.6$ . Example 5.22 shows that

$$\mu_{\hat{p}} = p = 0.6$$

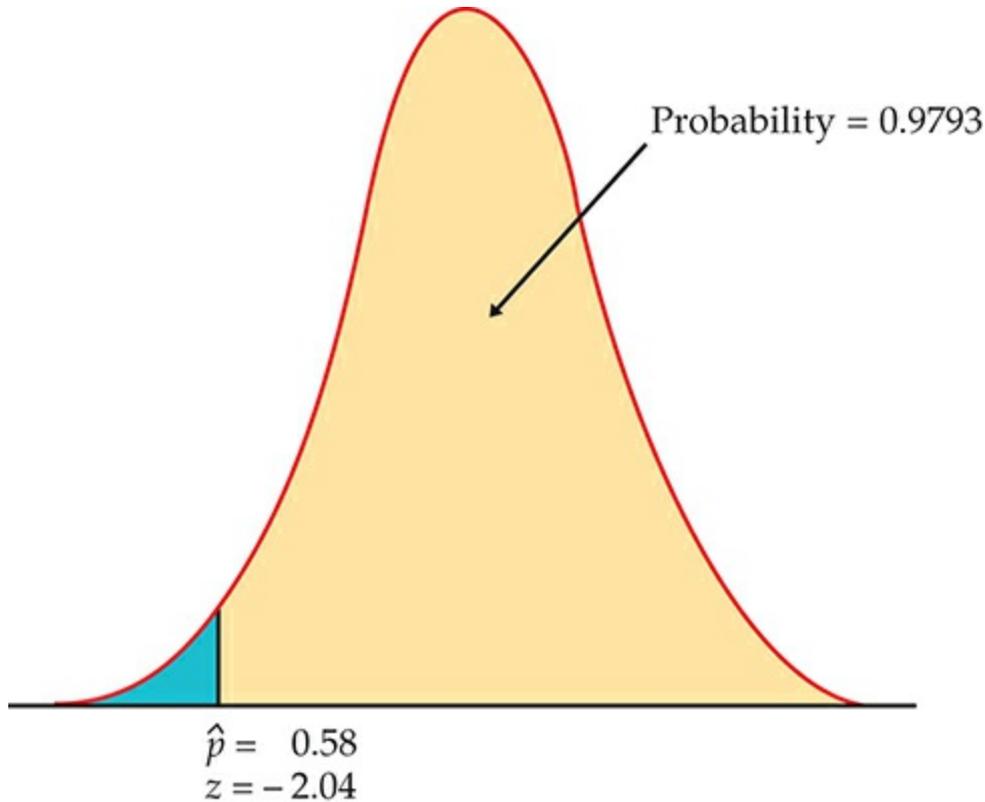
$$\sigma_{\hat{p}} = \sqrt{p(1-p)/n} = 0.0098$$

Act as if  $\hat{p}$  were Normal with mean  $0.6$  and standard deviation  $0.0098$ . The approximate probability, as illustrated in Figure 5.13, is

$$P(\hat{p} \geq 0.58) = P(\hat{p} - 0.60 / 0.0098 \geq (0.58 - 0.60) / 0.0098)$$

$$\doteq P(Z \geq -2.04) = 0.9793$$

That is, about 98% of all samples have a sample proportion that is at least 0.58. Because the sample was large, this Normal approximation is quite accurate. It misses the software value 0.9802 by only 0.0009.



**FIGURE 5.13**

The Normal probability calculation for Example 5.23.

## Example

### 5.24 Using the Normal approximation.

The audit described in Example 5.16 examined an SRS of 150 sales records for compliance with sales tax laws. In fact, 8% of all the company's sales records have an incorrect sales tax classification. The count  $X$  of bad records in the sample has approximately the  $B(150, 0.08)$  distribution.

According to the Normal approximation to the binomial distributions, the count  $X$  is approximately Normal with mean and standard deviation

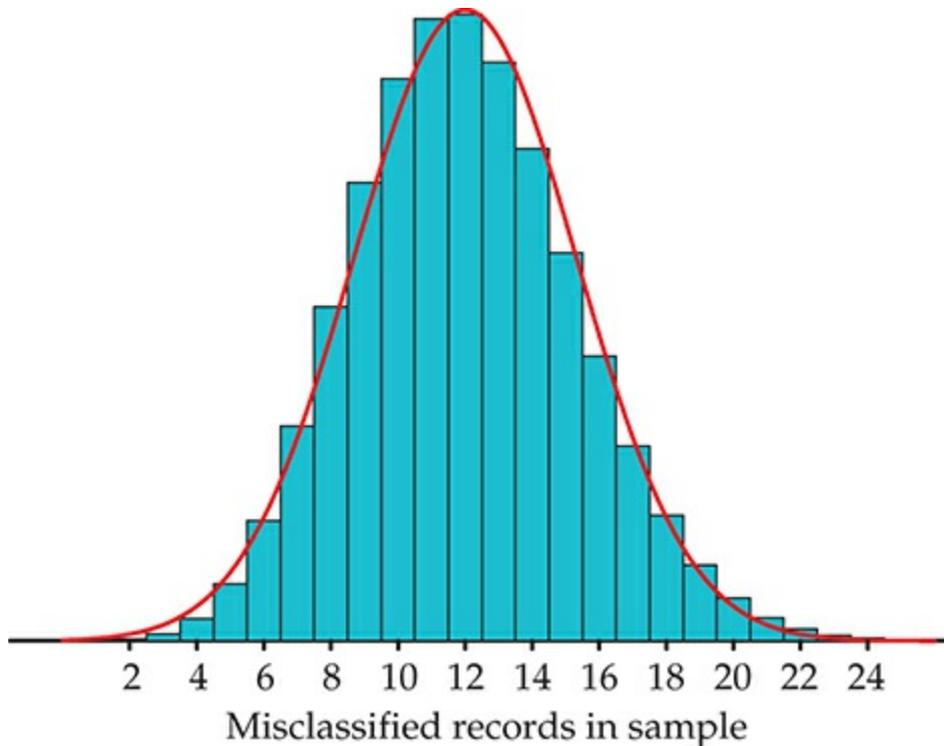
$$\mu_X = np = (150)(0.08) = 12$$

$$\sigma_X = np(1-p) = (150)(0.08)(0.92) = 3.3226$$

The Normal approximation for the probability of no more than 10 misclassified records is the area to the left of  $X = 10$  under the Normal curve. Using Table A,

$$\begin{aligned} P(X \leq 10) &= P(X - 123.3226 \leq 10 - 123.3226) \\ &\doteq P(Z \leq -0.60) = 0.2743 \end{aligned}$$

Software tells us that the actual binomial probability that no more than 10 of the records in the sample are misclassified is  $P(X \leq 10) = 0.3384$ . The Normal approximation is only roughly accurate. Because  $np = 12$ , this combination of  $n$  and  $p$  is close to the border of the values for which we are willing to use the approximation.



**FIGURE 5.14**

Probability histogram and Normal approximation for the binomial distribution with  $n = 150$  and  $p = 0.08$ , for Example 5.24.

The distribution of the count of bad records in a sample of 15 is distinctly non-Normal, as Figure 5.10 showed. When we increase the sample size to 150, however, the shape of the binomial distribution becomes roughly Normal. Figure 5.14 displays the probability histogram of the binomial distribution with the density curve of the approximating Normal distribution superimposed. Both distributions have the same mean and standard deviation, and both the area under the histogram and the area under the curve are 1. The Normal curve fits the histogram reasonably well. Look closely: the histogram is slightly skewed to the

right, a property that the symmetric Normal curve can't match.

## USE YOUR KNOWLEDGE

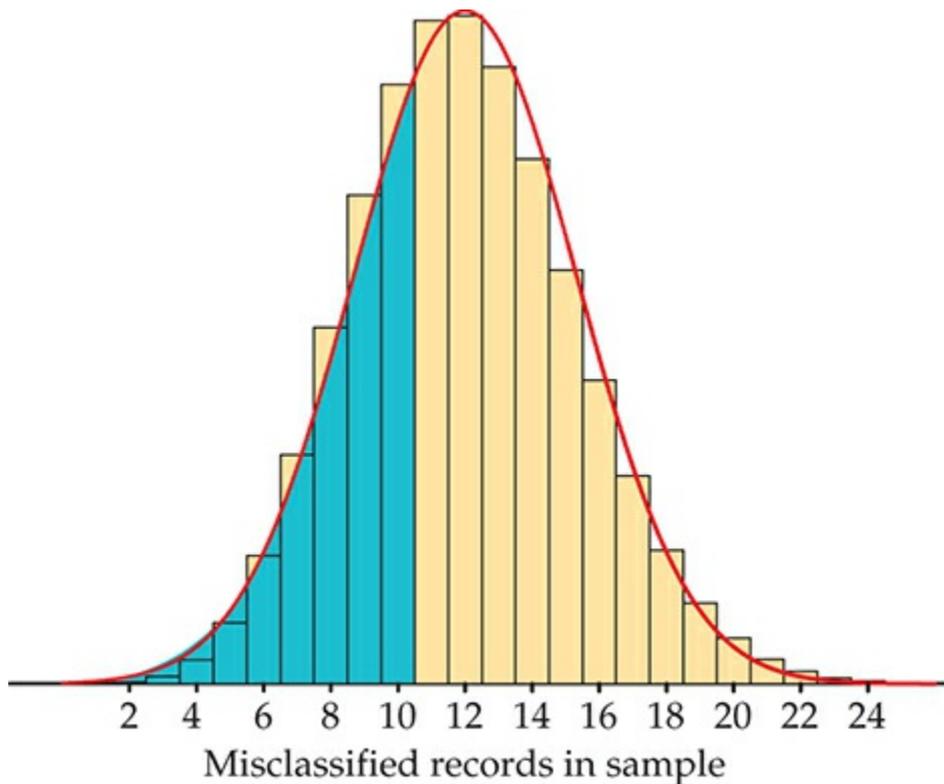
### 5.43 Use the Normal approximation.

Suppose that we toss a fair coin 200 times. Use the Normal approximation to find the probability that the sample proportion of heads is

- (a) between 0.4 and 0.6.
- (b) between 0.45 and 0.55.

## The continuity correction

Figure 5.15 illustrates an idea that greatly improves the accuracy of the Normal approximation to binomial probabilities. The binomial probability  $P(X \leq 10)$  is the area of the histogram bars for values 0 to 10. The bar for  $X = 10$  actually extends from 9.5 to 10.5. Because the discrete binomial distribution puts probability only on whole numbers, the probabilities  $P(X \leq 10)$  and  $P(X \leq 10.5)$  are the same. The Normal distribution spreads probability continuously, so these two Normal probabilities are different. The Normal approximation is more accurate if we consider  $X = 10$  to extend from 9.5 to 10.5, matching the bar in the probability histogram.



**FIGURE 5.15**

Area under the Normal approximation curve for the probability in Example 5.24

The event  $\{X \leq 10\}$  includes the outcome  $X = 10$ . Figure 5.15 shades the area under the Normal curve that matches all the histogram bars for outcomes 0 to 10, bounded on the right not by 10, but by 10.5. So  $P(X \leq 10)$  is calculated as  $P(X \leq 10.5)$ . On the other hand,  $P(X < 10)$  excludes the outcome  $X = 10$ , so we exclude the entire interval from 9.5 to 10.5 and calculate  $P(X \leq 9.5)$  from the Normal table. Here is the result of the Normal calculation in Example 5.24 improved in this way:

$$\begin{aligned}
 P(X \leq 10) &= P(X \leq 10.5) \\
 &= P(X - 123.3226 \leq 10.5 - 123.3226) \\
 &\doteq P(Z \leq -0.45) = 0.3264
 \end{aligned}$$

The improved approximation misses the binomial probability by only 0.012. Acting as though a whole number occupies the interval from 0.5 below to 0.5 above the number is called the **continuity correction** to the Normal approximation. If you need accurate values for binomial probabilities, try to use software to do exact calculations. If no software is available, use the continuity correction unless  $n$  is very large. Because most statistical purposes do not require extremely accurate probability calculations, we do not emphasize use of the continuity correction.

continuity correction

## Binomial formula

We can find a formula for the probability that a binomial random variable takes any value by adding probabilities for the different ways of getting exactly that many successes in  $n$  observations. Here is the example we will use to show the idea.

### Example

#### 5.25 Blood types of children.

Each child born to a particular set of parents has probability 0.25 of having blood type O. If these parents have 5 children, what is the probability that exactly 2 of them have type O blood?

The count of children with type O blood is a binomial random variable  $X$  with  $n = 5$  tries and probability  $p = 0.25$  of a success on each try. We want  $P(X = 2)$ .

Because the method doesn't depend on the specific example, we will use "S" for success and "F" for failure. In Example 5.25, "S" would stand for type O blood. Do the work in two steps.

*Step 1:* Find the probability that a specific 2 of the 5 tries give successes, say the first and the third. This is the outcome SFSFF. The multiplication rule for independent events tells us that

$$\begin{aligned}P(\text{SFSFF}) &= P(S)P(F)P(S)P(F)P(F) \\ &= (0.25)(0.75)(0.25)(0.75)(0.75) \\ &= (0.25)^2 (0.75)^3\end{aligned}$$

*Step 2:* Observe that the probability of *any one* arrangement of 2 S's and 3 F's has this same probability. That's true because we multiply together 0.25 twice and 0.75 three times whenever we have 2 S's and 3 F's. The probability that  $X = 2$  is the probability of getting 2 S's and 3 F's in any arrangement whatsoever. Here are all the possible arrangements:

SSFFF SFSFF SFFSF SFFFS FSSFF  
FSFSF FSFFS FFSSF FFSFS FFFSS

There are 10 of them, all with the same probability. The overall probability of 2

successes is therefore

$$P(X = 2) = 10 (0.25)^2 (0.75)^3 = 0.2637$$

The pattern of this calculation works for any binomial probability. To use it, we need to be able to count the number of arrangements of  $k$  successes in  $n$  observations without actually listing them. We use the following fact to do the counting.

## BINOMIAL COEFFICIENT

The number of ways of arranging  $k$  successes among  $n$  observations is given by the **binomial coefficient**

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

for  $k = 0, 1, 2, \dots, n$ .

The formula for binomial coefficients uses the **factorial** notation. The factorial  $n!$  for any positive whole number  $n$  is

**factorial**

$$n! = n \times (n - 1) \times (n - 2) \times \dots \times 3 \times 2 \times 1$$

Also,  $0! = 1$ . Notice that the larger of the two factorials in the denominator of a binomial coefficient will cancel much of the  $n!$  in the numerator. For example, the binomial coefficient we need for Example 5.25 is

$$\begin{aligned} \binom{5}{2} &= \frac{5!}{2!3!} \\ &= \frac{(5)(4)(3)(2)(1)(2)(1)}{(2)(1) \times (3)(2)(1)} \\ &= \frac{(5)(4)(2)(1)}{(2)(1)} = 2 \times 2 = 10 \end{aligned}$$

This agrees with our previous calculation.



*The notation  $\binom{n}{k}$  is not related to the fraction  $\frac{n}{k}$ . A helpful way to remember its meaning is to read it as “binomial coefficient  $n$  choose  $k$ ” Binomial coefficients have many uses in mathematics, but we are interested in them only as an aid to finding binomial probabilities. The binomial coefficient  $\binom{n}{k}$  counts the number of ways in which  $k$  successes can be distributed among  $n$  observations. The binomial probability  $P(X = k)$  is this count multiplied by the probability of any specific*

arrangement of the  $k$  successes. Here is the formula we seek.

## **BINOMIAL PROBABILITY**

If  $X$  has the binomial distribution  $B(n, p)$  with  $n$  observations and probability  $p$  of success on each observation, the possible values of  $X$  are  $0, 1, 2, \dots, n$ . If  $k$  is any one of these values, the **binomial probability** is

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Here is an example of the use of the binomial probability formula.

### **Example**

#### **5.26 Using the binomial probability formula.**

The number  $X$  of misclassified sales records in the auditor's sample in Example 5.18 has the  $B(15, 0.08)$  distribution. The probability of finding no more than 1 misclassified record is

$$\begin{aligned} P(X \leq 1) &= P(X = 0) + P(X = 1) \\ &= \binom{15}{0} (0.08)^0 (0.92)^{15} + \binom{15}{1} (0.08)^1 (0.92)^{14} \\ &= 15! / 0! 15! (1) (0.2863) + 15! / 1! 14! (0.08) (0.3112) \\ &= (1) (1) (0.2863) + (15) (0.08) (0.3112) \\ &= 0.2863 + 0.3734 = 0.6597 \end{aligned}$$

The calculation used the facts that  $0! = 1$  and that  $a^0 = 1$  for any number  $a \neq 0$ . The result agrees with that obtained from Table C in Example 5.18.

## **USE YOUR KNOWLEDGE**

### **5.44 An unfair coin.**

A coin is slightly bent, and as a result the probability of a head is 0.54. Suppose that you toss the coin four times.

- (a) Use the binomial formula to find the probability of 3 or more heads.
- (b) Compare your answer with the one that you would obtain if the coin were fair.

## The Poisson distributions

A count  $X$  has a binomial distribution when it is produced under the binomial setting. If one or more facets of this setting do not hold, the count  $X$  will have a different distribution. In this subsection, we discuss one of these distributions.

Frequently, we meet counts that are open-ended, that is, are not based on a fixed number of  $n$  observations: the number of customers at a popular cafe between 12:00 P.M. and 1:00 P.M.; the number of dings on your car door; the number of reported pedestrian/bicyclist collisions on campus during the academic year. These are all counts that could be 0, 1, 2, 3, and so on indefinitely.

The Poisson distribution is another model for a count and can often be used in these open-ended situations. The count represents the number of events (call them “successes”) that occur in some fixed unit of measure such as a period of time or region of space. The Poisson distribution is appropriate under the following conditions.

### THE POISSON SETTING

1. The number of successes that occur in two nonoverlapping units of measure are **independent**.
2. The probability that a success will occur in a unit of measure is the same for all units of equal size and is proportional to the size of the unit.
3. The probability that more than one event occurs in a unit of measure is negligible for very small-sized units. In other words, the events occur one at a time.

For binomial distributions, the important quantities were  $n$ , the fixed number of observations, and  $p$ , the probability of success on any given observation. For Poisson distributions, the only important quantity is the mean number of successes  $\mu$  occurring per unit of measure.

### POISSON DISTRIBUTION

The distribution of the count  $X$  of successes in the Poisson setting is the **Poisson distribution** with **mean**  $\mu$ . The parameter  $\mu$  is the mean number of successes per unit of measure. The possible values of  $X$  are the whole numbers 0, 1, 2, 3, ... If  $k$  is any whole number, then\*

$$P(X=k)=e^{-\mu}\mu^k k!$$

The **standard deviation** of the distribution is  $\mu$ .

---

\* The  $e$  in the Poisson probability formula is a mathematical constant equal to 2.71828 to six decimal places. Many calculators have an  $e^x$  function.

## Example

### 5.27 Number of dropped calls.

Suppose that the number of dropped calls on your cell phone varies, with an average of 2.1 calls per day. If we assume that the Poisson setting is reasonable for this situation, we can model the daily count of dropped calls  $X$  using the Poisson distribution with  $\mu = 2.1$ . What is the probability of having no more than 2 dropped calls tomorrow?

We can calculate  $P(X \leq 2)$  using either software or the Poisson probability formula. Using the probability formula:

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= e^{-2.1} \frac{(2.1)^0}{0!} + e^{-2.1} \frac{(2.1)^1}{1!} + e^{-2.1} \frac{(2.1)^2}{2!} \\ &= 0.1225 + 0.2572 + 0.2700 \\ &= 0.6497 \end{aligned}$$

Using the R software, the probability is

$$\begin{aligned} &\text{dpois}(0,2.1) + \text{dpois}(1,2.1) + \text{dpois}(2,2.1) \\ &[1] 0.6496314 \end{aligned}$$

These two answers differ slightly due to roundoff error in the hand calculation. There is roughly a 65% chance that you will have no more than 2 dropped calls tomorrow.

Similar to the binomial, Poisson probability calculations are rarely done by hand if the event includes numerous possible values for  $X$ . Most software provides functions to calculate  $P(X = k)$  and the cumulative probabilities of the form  $P(X \leq k)$ . These cumulative probability calculations make solving many problems less tedious. Here's an example.

## Example

### 5.28 Counting software remote users.

Your university supplies online remote access to various software programs used in courses. Suppose that the number of students remotely accessing these programs in any given hour can be modeled by a Poisson distribution with  $\mu = 17.2$ . What is the probability that more than 25 students will remotely access these programs in the next hour?

Calculating this probability requires two steps.

1. Write  $P(X > 25)$  as an expression involving a cumulative probability:  
$$P(X > 25) = 1 - P(X \leq 25)$$
2. Obtain  $P(X \leq 25)$  and subtract the value from 1. Again using R,  
`1- ppois(25,17.2)`  
`[1] 0.02847261`

The probability that more than 25 students will use this remote access in the next hour is only 0.028. Relying on software to get the cumulative probability is much quicker and less prone to error than the method of Example 5.27. For this case, that method would involve determining 26 probabilities and then summing their values.

Under the Poisson setting, this probability of 0.028 applies not only to the next hour but any other hour in the future. The probability does not change because the units of measure are the same size and nonoverlapping.

## USE YOUR KNOWLEDGE

### 5.45 Number of aphids.

The milkweed aphid is a common pest to many ornamental plants. Suppose that the number of aphids on a shoot of a Mexican butterfly weed follows a Poisson distribution with  $\mu = 4$  aphids.

- (a) What is the probability of observing exactly 5 aphids on a shoot?
- (b) What is the probability of observing 5 or fewer aphids on a shoot?

#### 5.46 Number of aphids, continued.

Refer to the previous exercise.

- (a) What proportion of shoots would you expect to have no aphids present?
- (b) If you do not observe any aphids on a shoot, is the probability that a nearby shoot has no aphids smaller than, equal to, or larger than your answer in part (a)? Explain your reasoning.

If we add counts from successive nonoverlapping areas, we are just counting the successes in a larger area. That count still meets the conditions of the Poisson setting. However, since our unit of measure has doubled, the mean of this new count is twice as large. Put more formally, if  $X$  is a Poisson random variable with mean  $\mu_X$  and  $Y$  is a Poisson random variable with mean  $\mu_Y$  and  $Y$  is independent of  $X$ , then  $X + Y$  is a Poisson random variable with mean  $\mu_X + \mu_Y$ . This fact means that we can combine areas or look at a portion of an area and still use Poisson distributions to model the count.

### Example

#### 5.29 Number of potholes.

The Automobile Association (AA) in Britain had member volunteers make a 60-minute, two-mile walk around their neighborhoods and survey the condition of their roads and sidewalks. One outcome was the number of potholes, defined as being at least 2 inches deep and at least 6 inches in diameter, in their roads.<sup>15</sup> It was reported that Scotland averages 8.9 potholes per mile of road and London averages 4.9 potholes per mile of road. Suppose that the number of potholes per mile in each of these two regions follow the Poisson distribution. Then

- The number of potholes per 20 miles of road in Scotland is a Poisson random variable with mean  $20 \times 8.9 = 178$ .

- The number of potholes per half mile of road in London is a Poisson random variable with mean  $0.5 \times 4.9 = 2.45$ .
- The number of potholes per 500 miles of road in Scotland is a Poisson random variable with mean  $500 \times 8.9 = 4450$ .
- If we examined 2 miles of road in Scotland and 5 miles of road in London, the total number of potholes would be a Poisson random variable with mean  $2 \times 8.9 + 5 \times 4.9 = 42.3$ .

When the mean of the Poisson distribution is large, it may be difficult to calculate Poisson probabilities using a calculator or software. Fortunately, when  $\mu$  is large, Poisson probabilities can be approximated using the Normal distribution with mean  $\mu$  and standard deviation  $\mu$ . Here is an example.

## Example

### 5.30 Number of text messages sent.

In Example 5.8 (page 311), it was reported that Americans aged 18 to 29 years send an average of almost 88 text messages a day. Suppose that the number of text messages you send per day follows a Poisson distribution with mean 88. What is the probability that over a week you would send more than 650 text messages?

To answer this using software, we first compute the mean number of text messages sent per week. Since there are 7 days in a week, the mean is  $7 \times 88 = 616$ . Plugging this into R tells us that there is slightly more than an 8% chance of sending this many texts:

```
1-ppois(650,616)
[1] 0.08317643
```

For the Normal approximation we compute

$$\begin{aligned}
 P(X > 650) &= P(X - 616 > 650 - 616) \\
 &= P(Z > 1.37) \\
 &= 1 - P(Z < 1.37) \\
 &= 1 - 0.9147 = 0.0853
 \end{aligned}$$

The approximation is quite accurate, differing from the actual probability by

| only 0.0021.

While the Normal approximation is adequate for many practical purposes, we recommend using statistical software when possible so you can get exact Poisson probabilities.

There is one other approximation associated with the Poisson distribution that is worth mentioning. It is related to the binomial distribution. Previously, we recommended using the Normal distribution to approximate the binomial distribution when  $n$  and  $p$  satisfy  $np \geq 10$  and  $n(1 - p) \geq 10$ . In cases where  $n$  is large but  $p$  is so small that  $np < 10$ , the Poisson distribution with  $\mu = np$  yields more accurate results. For example, suppose that you wanted to calculate  $P(X \leq 2)$  when  $X$  has the  $B(1000, 0.001)$  distribution. Using R, the actual binomial probability and the Poisson approximation are

```
pbinom(2,1000,.001) ppois(2,1)
[1] 0.9197907 [1] 0.9196986
```

The Poisson approximation gives a very accurate probability calculation for the binomial distribution in this case.

## SECTION 5.2 Summary

A **count**  $X$  of successes has the **binomial distribution**  $B(n, p)$  in the **binomial setting**: there are  $n$  trials, all independent, each resulting in a success or a failure, and each having the same probability  $p$  of a success.

The binomial distribution  $B(n, p)$  is a good approximation to the **sampling distribution of the count of successes** in an SRS of size  $n$  from a large population containing proportion  $p$  of successes. We will use this approximation when the population is at least 20 times larger than the sample.

The **sample proportion** of successes  $\hat{p} = X/n$  is an estimator of the population proportion  $p$ . It does not have a binomial distribution, but we can do probability calculations about  $\hat{p}$  by restating them in terms of  $X$ .

**Binomial probabilities** are most easily found by software. There is an exact formula that is practical for calculations when  $n$  is small. Table C contains binomial probabilities for some values of  $n$  and  $p$ . For large  $n$ , you can use the Normal approximation.

The mean and standard deviation of a **binomial count**  $X$  and a **sample proportion**  $\hat{p} = X/n$  are

$$\mu_X = np$$

$$\mu_{\hat{p}} = p$$

$$\sigma_X = np(1-p)$$

$$\sigma_{\hat{p}} = p(1-p)/n$$

The sample proportion  $\hat{p}$  is therefore an unbiased estimator of the population proportion  $p$ .

The **Normal approximation** to the binomial distribution says that if  $X$  is a count having the  $B(n, p)$  distribution, then when  $n$  is large,

$$X \text{ is approximately } N(np, np(1-p))$$

$$\hat{p} \text{ is approximately } N(p, p(1-p)/n)$$

We will use this approximation when  $np \geq 10$  and  $n(1-p) \geq 10$ . It allows us to approximate probability calculations about  $X$  and  $\hat{p}$  using the Normal distribution.

The **continuity correction** improves the accuracy of the Normal approximations. The exact **binomial probability formula** is

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where the possible values of  $X$  are  $k = 0, 1, \dots, n$ . The binomial probability formula uses the **binomial coefficient**

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Here the **factorial**  $n!$  is

$$n! = n \times (n-1) \times (n-2) \times \dots \times 3 \times 2 \times 1$$

for positive whole numbers  $n$  and  $0! = 1$ . The binomial coefficient counts the number of ways of distributing  $k$  successes among  $n$  trials.

A count  $X$  of successes has a **Poisson distribution** in the **Poisson setting**: the number of successes that occur in two nonoverlapping units of measure are independent; the probability that a success will occur in a unit of measure is the same for all units of equal size and is proportional to the size of the unit; the probability that more than one event occurs in a unit of measure is negligible for very small-sized units. In other words, the events occur one at a time.

If  $X$  has the Poisson distribution with mean  $\mu$ , then the standard deviation of  $X$  is  $\mu$ , and the possible values of  $X$  are the whole numbers 0, 1, 2, 3, and so on.

The **Poisson probability** that  $X$  takes any of these values is

$$P(X=k) = \frac{e^{-\mu} \mu^k}{k!} \quad k=0,1,2,3,\dots$$

Sums of independent Poisson random variables also have the Poisson distribution. For example, in a Poisson model with mean  $\mu$  per unit of measure, the count of successes in  $a$  units is a Poisson random variable with mean  $a\mu$ .

## SECTION 5.2 Summary

*For Exercises 5.35 to 5.37, see page 322; for Exercises 5.38 and 5.39, see page 324; for Exercises 5.40 and 5.41, see page 327; for Exercise 5.42, see page 331; for Exercise 5.43, see page 335; for Exercise 5.44, see page 339; and for Exercises 5.45 and 5.46, see page 341.*

*Most binomial probability calculations required in these exercises can be done by using Table C or the Normal approximation. Your instructor may request that you use the binomial probability formula or*

software. In exercises requiring the Normal approximation, you should use the continuity correction if you studied that topic.

### 5.47 What is wrong?

Explain what is wrong in each of the following scenarios.

- (a) If you toss a fair coin four times and a head appears each time, then the next toss is more likely to be a tail than a head.
- (b) If you toss a fair coin four times and observe the pattern HTHT, then the next toss is more likely to be a head than a tail.
- (c) The quantity  $\hat{p}$  is one of the parameters for a binomial distribution.
- (d) The binomial distribution can be used to model the daily number of pedestrian/cyclist near-crash events on campus.

### 5.48 What is wrong?

Explain what is wrong in each of the following scenarios.

- (a) In the binomial setting,  $X$  is a proportion.
- (b) The variance for a binomial count is  $p(1-p)/n$ .
- (c) The Normal approximation to the binomial distribution is always accurate when  $n$  is greater than 1000.
- (d) We can use the binomial distribution to approximate the sampling distribution of  $\hat{p}$  when we draw an SRS of size  $n = 50$  students from a population of 500 students.

### 5.49 you use the binomial distribution?

In each of the following situations, is it reasonable to use a binomial distribution for the random variable  $X$ ? Give reasons for your answer in each case. If a binomial distribution applies, give the values of  $n$  and  $p$ .

- (a) A poll of 200 college students asks whether or not you usually feel irritable in the morning.  $X$  is the number who reply that they do usually feel irritable in the morning.
- (b) You toss a fair coin until a head appears.  $X$  is the count of the number of tosses that you make.
- (c) Most calls made at random by sample surveys don't succeed in talking with a person. Of calls to New York City, only one-twelfth succeed. A survey calls 500 randomly selected numbers in New York City.  $X$  is the number of times that a person is reached.
- (d) You deal 10 cards from a shuffled deck of standard playing cards and count the number  $X$  of black cards.

### 5.50 Should you use the binomial distribution?

In each of the following situations, is it reasonable to use a binomial distribution for the random variable  $X$ ? Give reasons for your answer in each case.

- (a) In a random sample of students in a fitness study,  $X$  is the mean daily exercise time of the sample.

(b) A manufacturer of running shoes picks a random sample of 20 shoes from the production of shoes each day for a detailed inspection.  $X$  is the number of pairs of shoes with a defect.

(c) A nutrition study chooses an SRS of college students. They are asked whether or not they usually eat at least five servings of fruits or vegetables per day.  $X$  is the number who say that they do.

(d)  $X$  is the number of days during the school year when you skip a class.

### 5.51 Stealing from a store.

A survey of over 20,000 U.S. high school students revealed that 20% of the students say that they stole something from a store in the past year.<sup>16</sup> This is down 7% from the last survey, which was performed two years earlier. You decide to take a random sample of 10 high school students from your city and ask them this question.

(a) If the high school students in your city match this 20% rate, what is the distribution of the number of students who say that they stole something from a store in the past year? What is the distribution of the number of students who do not say that they stole something from a store in the past year?

(b) What is the probability that 4 or more of the 10 students in your sample say that they stole something from a store in the past year?

### 5.52 Paying for music downloads.

A survey of Canadian teens aged 12 to 17 years reported that roughly 75% of them used a fee-based website to download music.<sup>17</sup> You decide to interview a random sample of 15 U.S. teenagers. For now, assume that they behave similarly to the Canadian teenagers.

(a) What is the distribution of the number  $X$  who used a fee-based website to download music? Explain your answer.

(b) What is the probability that at least 12 of the 15 teenagers in your sample used a fee-based website to download music.

### 5.53 Stealing from a store, continued.

Refer to Exercise 5.51.

(a) What is the mean number of students in your sample who say that they stole something from a store in the past year? What is the mean number of students who do not say that they stole? You should see that these two means add to 10, the total number of students.

(b) What is the standard deviation  $\sigma$  of the number of students in your sample who say that they stole something?

(c) Suppose that you live in a city where only 10% of the high school students say that they stole something from a store in the past year. What is  $\sigma$  in this case? What is  $\sigma$  if  $p = 0.01$ ? What happens to the standard deviation of a binomial distribution as the probability of a success gets close to 0?

### 5.54 Paying for music downloads, continued.

Refer to Exercise 5.52. Suppose that only 60% of the U.S. teenagers used a fee-based website to download music.

(a) If you interview 15 U.S. teenagers at random, what is the mean of the count  $X$  who used a fee-based website to download music? What is the mean of the proportion  $\hat{p}$  in your sample who used a fee-based website to download music?

(b) Repeat the calculations in part (a) for samples of size 150 and 1500. What happens to the mean count of successes as the sample size increases? What happens to the mean proportion of successes?

### 5.55 More on paying for music downloads.

Consider the settings of Exercises 5.52 and 5.54.

(a) Using the 75% rate of the Canadian teenagers, what is the smallest number  $m$  out of  $n = 15$  U.S. teenagers such that  $P(X \leq m)$  is no larger than 0.05? You might consider  $m$  or fewer students as evidence that the rate in your sample is lower than the 75% rate of the Canadian teenagers.

(b) Now using the 60% rate of the U.S. teenagers and your answer to part (a), what is  $P(X \leq m)$ ? This represents the chance of obtaining enough evidence given that the rate is 60%.

(c) If you were to increase the sample size from  $n = 15$  to  $n = 100$  and repeat parts (a) and (b), would you expect the probability in (b) to increase or decrease? Explain your answer.

### 5.56 Attitudes toward drinking and studies of behavior.

Some of the methods in this section are approximations rather than exact probability results. We have given rules of thumb for safe use of these approximations.

(a) You are interested in attitudes toward drinking among the 75 members of a fraternity. You choose 30 members at random to interview. One question is “Have you had five or more drinks at one time during the last week?” Suppose that in fact 30% of the 75 members would say “Yes.” Explain why you *cannot* safely use the  $B(30, 0.3)$  distribution for the count  $X$  in your sample who say “Yes.”

(b) The National AIDS Behavioral Surveys found that 0.2% (that’s 0.002 as a decimal fraction) of adult heterosexuals had both received a blood transfusion and had a sexual partner from a group at high risk of AIDS. Suppose that this national proportion holds for your region. Explain why you *cannot* safely use the Normal approximation for the sample proportion who fall in this group when you interview an SRS of 1000 adults.

### 5.57 Random digits.

Each entry in a table of random digits like Table B has probability 0.1 of being a 0, and digits are independent of each other.

(a) What is the probability that a group of six digits from the table will contain at least one 5?

(b) What is the mean number of 5s in lines 40 digits long?



### 5.58 Use the *Probability* applet.

The *Probability* applet simulates tosses of a coin. You can choose the number of tosses  $n$ , and the probability  $p$  of a head. You can therefore use the applet to simulate binomial random variables.

The count of misclassified sales records in Example 5.18 (page 326) has the binomial distribution with  $n = 15$  and  $p = 0.08$ . Set these values for the number of tosses and probability of heads in the applet. Table C shows that the probability of getting a sample with exactly 0 misclassified records is 0.2863. This is the

long-run proportion of samples with no bad records. Click “Toss” and “Reset” repeatedly to simulate 25 samples of 15 tosses. Record the number of bad records (the count of heads) in each of the 25 samples.

- (a) What proportion of the 25 samples had exactly 0 bad records? Do you think this sample proportion is close to the probability?
- (b) Remember that this probability of 0.2863 tells us only what happens in the long run. Here we’re considering only 25 samples. If  $X$  is the number of samples out of 25 with exactly 0 misclassified records, what is the distribution of  $X$ ?
- (c) Explain how to use the distribution in part (b) to describe the sampling distribution of  $\hat{p}$  in part (a).

### 5.59 Illegal file sharing.

Would you stop illegal file sharing if you received a warning with a penalty notice attached? More than 1000 adult New Zealanders (aged 15 to 50 years) were asked this. Of those who have illegally file-shared content 70% said that they would stop.<sup>18</sup> You randomly sample 4 New Zealanders who have illegally file-shared content and ask them this question. Let  $X$  be the number who say “Yes.”

- (a) What are  $n$  and  $p$  in the binomial distribution of  $X$ ?
- (b) Find the probability of each possible value of  $X$ , and draw a probability histogram for this distribution.
- (c) Find the mean number of positive responders and mark the location of this value on your histogram.

### 5.60 The ideal number of children.

“What do you think is the ideal number of children for a family to have?” A Gallup Poll asked this question of 1020 randomly chosen adults. Over half (53%) thought that a total of two children was ideal.<sup>19</sup> Suppose that  $p = 0.53$  is exactly true for the population of all adults. Gallup announced a margin of error of  $\pm 4$  percentage points for this poll. What is the probability that the sample proportion  $\hat{p}$  for an SRS of size  $n = 1020$  falls between 0.49 and 0.57? You see that it is likely, but not certain, that polls like this give results that are correct within their margin of error. We will say more about margins of error in Chapter 6.

### 5.61 Illegal file sharing, continued.

Refer to Exercise 5.59. Roughly 30% of those surveyed had illegally file-shared content. Assume that you sample  $n = 300$  New Zealanders.

- (a) What is the probability that the sample proportion  $\hat{p}$  of those who would stop after being given a warning is between 0.67 and 0.73 if the population proportion is  $p = 0.70$ ?
- (b) What is the probability that the sample proportion  $\hat{p}$  is between 0.87 and 0.93 if the population proportion is  $p = 0.90$ ?
- (c) Using the results from parts (a) and (b), how does the probability that  $\hat{p}$  falls within  $\pm 0.03$  of the true  $p$  change as  $p$  gets closer to 1?

### 5.62 How do the results depend on the sample size?

Return to the Gallup Poll setting of Exercise 5.60. We are supposing that the proportion of all adults who think that having two children is ideal is  $p = 0.53$ . What is the probability that a sample proportion  $\hat{p}$  falls

between 0.49 and 0.57 (that is, within  $\pm 4$  percentage points of the true  $p$ ) if the sample is an SRS of size  $n = 300$ ? Of size  $n = 5000$ ? Combine these results with your work in Exercise 5.60 to make a general statement about the effect of larger samples in a sample survey.

### 5.63 Shooting free throws.

Since the mid-1960s, the overall free-throw percent at all college levels, for both men and women, has remained pretty consistent. For men, players have been successful on roughly 69% of their free throws, with the season percent never falling below 67% or above 70%.<sup>20</sup> Assume that 300,000 free throws will be attempted in the upcoming season.

- (a) What are the mean and standard deviation of  $\hat{p}$  if the population proportion is  $p = 0.69$ ?
- (b) Using the 68–95–99.7 rule, we expect  $\hat{p}$  to fall between what two percents about 95% of the time?
- (c) Given the width of the interval in part (b) and the range of season percents, do you think that it is reasonable to assume that the population proportion has been the same over the last 50 seasons? Explain your answer.

### 5.64 Online learning.

Recently the U.S. Department of Education released a report on online learning stating that blended instruction, a combination of conventional face-to-face and online instruction, appears more effective in terms of student performance than conventional teaching.<sup>21</sup> You decide to poll the incoming students at your institution to see if they prefer courses that blend face-to-face instruction with online components. In an SRS of 400 incoming students, you find that 311 prefer this type of course.

- (a) What is the sample proportion who prefer this type of blended instruction?
- (b) If the population proportion for all students nationwide is 85%, what is the standard deviation of  $\hat{p}$ ?
- (c) Using the 68–95–99.7 rule, if you had drawn an SRS from the United States, you would expect  $\hat{p}$  to fall between what two percents about 95% of the time?
- (d) Based on your result in part (a), do you think that the incoming students at your institution prefer this type of instruction more, less, or about the same as students nationally? Explain your answer.



### 5.65 Binge drinking among women.

The Centers for Disease Control and Prevention finds that 24.2% of women aged 18 to 24 years binge drank. Binge drinking for women is defined as consuming at least 4 alcoholic drinks per episode during the past 30 days. Those who binge drank averaged 6.4 drinks per episode and 3.6 episodes per month. The study took a sample of almost 11,000 women aged 18 to 24 years, so the population proportion of women who binge drank is very close to  $p = 0.24$ .<sup>22</sup> The administration of your college surveys an SRS of 200 female students and finds that 56 binge drink.

- (a) What is the sample proportion of women at your college who binge drink?
- (b) If, in fact, the proportion of all women on your campus who binge drink is the same as the national 24%, what is the probability that the proportion in an SRS of 200 students is as large or larger than the result of the administration's sample?
- (c) A writer for the student paper says that the percent of women who binge drink is higher on your campus than nationally. Write a short letter to the editor explaining why the survey does not support this

conclusion.

### 5.66 How large a sample is needed?

The changing probabilities you found in Exercises 5.60 and 5.62 are due to the fact that the standard deviation of the sample proportion  $\hat{p}$  gets smaller as the sample size  $n$  increases. If the population proportion is  $p = 0.53$ , how large a sample is needed to reduce the standard deviation of  $\hat{p}$  to  $\sigma_{\hat{p}} = 0.005$ ? (The 68–95–99.7 rule then says that about 95% of all samples will have  $\hat{p}$  within 0.01 of the true  $p$ .)

### 5.67 A test for ESP.

In a test for ESP (extrasensory perception), the experimenter looks at cards that are hidden from the subject. Each card contains either a star, a circle, a wave, or a square. As the experimenter looks at each of 20 cards in turn, the subject names the shape on the card.

- (a) If a subject simply guesses the shape on each card, what is the probability of a successful guess on a single card? Because the cards are independent, the count of successes in 20 cards has a binomial distribution.
- (b) What is the probability that a subject correctly guesses at least 10 of the 20 shapes?
- (c) In many repetitions of this experiment with a subject who is guessing, how many cards will the subject guess correctly on the average? What is the standard deviation of the number of correct guesses?
- (d) A standard ESP deck actually contains 25 cards. There are five different shapes, each of which appears on 5 cards. The subject knows that the deck has this makeup. Is a binomial model still appropriate for the count of correct guesses in one pass through this deck? If so, what are  $n$  and  $p$ ? If not, why not?

### 5.68 Admitting students to college.

A selective college would like to have an entering class of 950 students. Because not all students who are offered admission accept, the college admits more than 950 students. Past experience shows that about 75% of the students admitted will accept. The college decides to admit 1200 students. Assuming that students make their decisions independently, the number who accept has the  $B(1200, 0.75)$  distribution. If this number is less than 950, the college will admit students from its waiting list.

- (a) What are the mean and the standard deviation of the number  $X$  of students who accept?
- (b) Use the Normal approximation to find the probability that at least 800 students accept.
- (c) The college does not want more than 950 students. What is the probability that more than 950 will accept?
- (d) If the college decides to increase the number of admission offers to 1300, what is the probability that more than 950 will accept?

### 5.69 Is the ESP result better than guessing?

When the ESP study of Exercise 5.67 discovers a subject whose performance appears to be better than guessing, the study continues at greater length. The experimenter looks at many cards bearing one of five shapes (star, square, circle, wave, and cross) in an order determined by random numbers. The subject cannot see the experimenter as the experimenter looks at each card in turn, in order to avoid any possible nonverbal clues. The answers of a subject who does not have ESP should be independent observations, each with probability  $1/5$  of success. We record 900 attempts.

- (a) What are the mean and the standard deviation of the count of successes?
- (b) What are the mean and the standard deviation of the proportion of successes among the 900 attempts?
- (c) What is the probability that a subject without ESP will be successful in at least 24% of 900 attempts?
- (d) The researcher considers evidence of ESP to be a proportion of successes so large that there is only probability 0.01 that a subject could do this well or better by guessing. What proportion of successes must a subject have to meet this standard? (Example 1.45, on page 67, shows how to do an inverse calculation for the Normal distribution that is similar to the type required here.)



### 5.70 Show that these facts are true.

Use the definition of binomial coefficients to show that each of the following facts is true. Then restate each fact in words in terms of the number of ways that  $k$  successes can be distributed among  $n$  observations.

- (a)  $\binom{n}{n} = 1$  for any whole number  $n \geq 1$ .
- (b)  $\binom{n}{n-1} = n$  for any whole number  $n \geq 1$ .
- (c)  $\binom{n}{k} = \binom{n}{n-k}$  for any  $n$  and  $k$  with  $k \leq n$ .

### 5.71 Multiple-choice tests.

Here is a simple probability model for multiple-choice tests. Suppose that each student has probability  $p$  of correctly answering a question chosen at random from a universe of possible questions. (A strong student has a higher  $p$  than a weak student.) The correctness of an answer to a question is independent of the correctness of answers to other questions. Jodi is a good student for whom  $p = 0.88$ .

- (a) Use the Normal approximation to find the probability that Jodi scores 85% or lower on a 100-question test.
- (b) If the test contains 250 questions, what is the probability that Jodi will score 85% or lower?
- (c) How many questions must the test contain in order to reduce the standard deviation of Jodi's proportion of correct answers to half its value for a 100-item test?
- (d) Laura is a weaker student for whom  $p = 0.72$ . Does the answer you gave in part (c) for the standard deviation of Jodi's score apply to Laura's standard deviation also?

### 5.72 Tossing a die.

You are tossing a balanced die that has probability  $1/6$  of coming up 1 on each toss. Tosses are independent. We are interested in how long we must wait to get the first 1.

- (a) The probability of a 1 on the first toss is  $1/6$ . What is the probability that the first toss is not a 1 and the second toss is a 1?
- (b) What is the probability that the first two tosses are not 1s and the third toss is a 1? This is the probability that the first 1 occurs on the third toss.
- (c) Now you see the pattern. What is the probability that the first 1 occurs on the fourth toss? On the fifth toss?

### 5.73 The geometric distribution.

Generalize your work in Exercise 5.72. You have independent trials, each resulting in a success or a failure. The probability of a success is  $p$  on each trial. The binomial distribution describes the count of successes in a fixed number of trials. Now the number of trials is not fixed; instead, continue until you get a success. The random variable  $Y$  is the number of the trial on which the first success occurs. What are the possible values of  $Y$ ? What is the probability  $P(Y = k)$  for any of these values? (*Comment:* The distribution of the number of trials to the first success is called a **geometric distribution**.)

### 5.74 Number of colony-forming units.

In microbiology, colony-forming units (CFUs) are used to measure the number of microorganisms present in a sample. To determine the number of CFUs, the sample is prepared, spread uniformly on an agar plate, and then incubated at some suitable temperature. Suppose that the number of CFUs that appear after incubation follows a Poisson distribution with  $\mu = 15$ .

- If the area of the agar plate is 75 square centimeters ( $\text{cm}^2$ ), what is the probability of observing fewer than 4 CFUs in a  $25 \text{ cm}^2$  area of the plate?
- If you were to count the total number of CFUs in 5 plates, what is the probability you would observe more than 90 CFUs? Use the Poisson distribution to obtain this probability.
- Repeat the probability calculation in part (b) but now use the Normal approximation. How close is your answer to your answer in part (b)?

### 5.75 Metal fatigue.

Metal fatigue refers to the gradual weakening and eventual failure of metal that undergoes cyclic loads. The wings of an aircraft, for example, are subject to cyclic loads when in the air, and cracks can form. It is thought that these cracks start at large particles found in the metal. Suppose that the number of particles large enough to initiate a crack follows a Poisson distribution with mean  $\mu = 0.5$  per square centimeter ( $\text{cm}^2$ ).

- What is the mean of the Poisson distribution if we consider a  $100 \text{ cm}^2$  area?
- Using the Normal approximation, what is the probability that this section has more than 60 of these large particles?

## CHAPTER 5 Exercises

### 5.76 The cost of Internet access.

In Canada, households spent an average of \$68 monthly for high-speed broadband access.<sup>23</sup> Assume that the standard deviation is \$22. If you ask an SRS of 500 Canadian households with broadband access how much they pay, what is the probability that the average amount will exceed \$70?

### 5.77 Dust in coal mines.

A laboratory weighs filters from a coal mine to measure the amount of dust in the mine atmosphere. Repeated measurements of the weight of dust on the same filter vary Normally with standard deviation  $\sigma = 0.08$  milligram (mg) because the weighing is not perfectly precise. The dust on a particular filter actually weighs 123 mg.

- The laboratory reports the mean of 3 weighings of this filter. What is the distribution of this mean?
- What is the probability that the laboratory reports a weight of 124 mg or higher for this filter?

### 5.78 The effect of sample size on the standard deviation.

Assume that the standard deviation in a very large population is 100.

- Calculate the standard deviation for the sample mean for samples of size 1, 4, 25, 100, 250, 500, 1000, and 5000.
- Graph your results with the sample size on the  $x$  axis and the standard deviation on the  $y$  axis.
- Summarize the relationship between the sample size and the standard deviation that your graph shows.

### 5.79 Marks per round in cricket.

Cricket is a dart game that uses the numbers 15 to 20 and the bull's-eye. Each time you hit one of these regions you score either 0, 1, 2, or 3 marks. Thus, in a round of three throws, a person can score 0 to 9 marks. Lex plans to play 20 games. Her distribution of marks per round is discrete and strongly skewed. A majority of her rounds result in 0, 1, or 2 marks and only a few are more than 4 marks. Assume that her mean is 2.13 marks per round with a standard deviation of 1.88.

- Her 20 games involve 140 rounds of three throws each. What are the mean and standard deviation of the average number of marks  $\bar{x}$  in 140 rounds?
- Using the central limit theorem, what is the probability that she averages fewer than 2 marks per round?
- Do you think that the central limit theorem can be used in this setting? Explain your answer.

### 5.80 Common last names.

The U.S. Census Bureau says that the 10 most common names in the United States are (in order) Smith, Johnson, Williams, Brown, Jones, Miller, Davis, Garcia, Rodriguez, and Wilson.<sup>24</sup> These names account for 4.9% of all U.S. residents. Out of curiosity, you look at the authors of the textbooks for your current courses. There are 12 authors in all. Would you be surprised if none of the names of these authors were among the 10 most common? Give a probability to support your answer and explain the reasoning behind your calculation.

### 5.81 Benford's law.

It is a striking fact that the first digits of numbers in legitimate records often follow a distribution known as Benford's law. Here it is:

First digit	1	2	3	4	5	6	7	8	9
Proportion	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Fake records usually have fewer first digits 1, 2, and 3. What is the approximate probability, if Benford's law holds, that among 1000 randomly chosen invoices there are 560 or fewer in amounts with first digit 1, 2, or 3?

### 5.82 Genetics of peas.

According to genetic theory, the blossom color in the second generation of a certain cross of sweet peas should be red or white in a 3:1 ratio. That is, each plant has probability  $\frac{3}{4}$  of having red blossoms, and the blossom colors of separate plants are independent.

- (a) What is the probability that exactly 9 out of 12 of these plants have red blossoms?
- (b) What is the mean number of red-blossomed plants when 120 plants of this type are grown from seeds?
- (c) What is the probability of obtaining at least 80 red-blossomed plants when 120 plants are grown from seeds?

### 5.83 The weight of a dozen eggs.

The weight of the eggs produced by a certain breed of hen is Normally distributed with mean 66 grams (g) and standard deviation 6 g. If cartons of such eggs can be considered to be SRSs of size 12 from the population of all eggs, what is the probability that the weight of a carton falls between 755 and 830 g?

### 5.84 Plastic caps for motor oil containers.

A machine fastens plastic screw-on caps onto containers of motor oil. If the machine applies more torque than the cap can withstand, the cap will break. Both the torque applied and the strength of the caps vary. The capping-machine torque has the Normal distribution with mean 7.0 inch-pounds and standard deviation 0.9 inch-pounds. The cap strength (the torque that would break the cap) has the Normal distribution with mean 10.1 inch-pounds and standard deviation 1.2 inch-pounds.

- (a) Explain why it is reasonable to assume that the cap strength and the torque applied by the machine are independent.

(b) What is the probability that a cap will break while being fastened by the capping machine?

### 5.85 A roulette payoff revisited.

Refer to Exercise 5.26 (page 319). In part (d), the central limit theorem was used to approximate the probability that Sam ends the year ahead. The estimate was about 0.10 too large. Let's see if we can get closer using the Normal approximation to the binomial with the continuity correction.

(a) If Sam plans to bet on 520 roulette spins, he needs to win at least \$520 to break even. If each win gives him \$35, what is the minimum number of wins  $m$  he must have?

(b) Given  $p = 1/38 = 0.026$ , what are the mean and standard deviation of  $X$ , the number of wins in 520 roulette spins?

(c) Use the information in the previous two parts to compute  $P(X \geq m)$  with the continuity correction. Does your answer get closer to the exact probability 0.396?

### 5.86 Learning a foreign language.

Does delaying oral practice hinder learning a foreign language? Researchers randomly assigned 25 beginning students of Russian to begin speaking practice immediately and another 25 to delay speaking for four weeks. At the end of the semester both groups took a standard test of comprehension of spoken Russian. Suppose that in the population of all beginning students, the test scores for early speaking vary according to the  $N(32, 6)$  distribution and scores for delayed speaking have the  $N(29, 5)$  distribution.

(a) What is the sampling distribution of the mean score  $\bar{x}$  in the early-speaking group in many repetitions of the experiment? What is the sampling distribution of the mean score  $\bar{y}$  in the delayed-speaking group?

(b) If the experiment were repeated many times, what would be the sampling distribution of the difference  $\bar{y} - \bar{x}$  between the mean scores in the two groups?

(c) What is the probability that the experiment will find (misleadingly) that the mean score for delayed speaking is at least as large as that for early speaking?

### 5.87 Summer employment of college students.

Suppose (as is roughly true) that 88% of college men and 82% of college women were employed last summer. A sample survey interviews SRSs of 400 college men and 400 college women. The two samples are of course independent.

(a) What is the approximate distribution of the proportion  $p^F$  of women who worked last summer? What is the approximate distribution of the proportion  $p^M$  of men who worked?

(b) The survey wants to compare men and women. What is the approximate distribution of the difference in the proportions who worked,  $p^M - p^F$ ? Explain the reasoning behind your answer.

(c) What is the probability that in the sample a higher proportion of women than men worked last summer?

### 5.88 Income of working couples.

A study of working couples measures the income  $X$  of the husband and the income  $Y$  of the wife in a large number of couples in which both partners are employed. Suppose that you knew the means  $\mu_X$  and  $\mu_Y$  and the variances  $\sigma_X^2$  and  $\sigma_Y^2$  of both variables in the population.

- (a) Is it reasonable to take the mean of the total income  $X + Y$  to be  $\mu_X + \mu_Y$ ? Explain your answer.
- (b) Is it reasonable to take the variance of the total income to be  $\sigma_X^2 + \sigma_Y^2$ ? Explain your answer.

### 5.89 A random walk.

A particle moves along the line in a random walk. That is, the particle starts at the origin (position 0) and moves either right or left in independent steps of length 1. If the particle moves to the right with probability 0.6, its movement at the  $i$ th step is a random variable  $X_i$  with distribution

$$P(X_i = 1) = 0.6$$

$$P(X_i = -1) = 0.4$$

The position of the particle after  $k$  steps is the sum of these random movements,

$$Y = X_1 + X_2 + \dots + X_k$$

Use the central limit theorem to find the approximate probability that the position of the particle after 500 steps is at least 200 to the right.

### 5.90 A lottery payoff.

A \$1 bet in a state lottery's Pick 3 game pays \$500 if the three-digit number you choose exactly matches the winning number, which is drawn at random. Here is the distribution of the payoff  $X$ :

Payoff $X$	\$0	\$500
Probability	0.999	0.001

Each day's drawing is independent of other drawings.

- (a) Joe buys a Pick 3 ticket twice a week. The number of times he wins follows a  $B(104, 0.001)$  distribution. Using the Poisson approximation to the binomial, what is the probability that he wins at least once?
- (b) The exact binomial probability is 0.0988. How accurate is the Poisson approximation here?
- (c) If Joe pays \$5 a ticket, he needs to win at least twice a year to come out ahead. Using the Poisson approximation, what is the probability that Joe comes out ahead?

# 6 Introduction to Inference

CHAPTER



- 6.1 Estimating with Confidence**
- 6.2 Tests of Significance**
- 6.3 Use and Abuse of Tests**
- 6.4 Power and Inference as a Decision**

## Introduction

Statistical inference draws conclusions about a population or process from sample data. It also provides a statement of how much confidence we can place in our conclusions. Although there are numerous methods for inference, there are only a few general types of statistical inference. This chapter introduces the two most common types: *confidence intervals* and *tests of significance*.

Because the underlying reasoning for these two types of inference remains the same across different settings, this chapter considers just one simple setting that is closely related to our study of the sampling distributions of  $\bar{x}$  in Section 5.1 (page 303): inference about the mean of a large population whose standard deviation is known. This setting, although unrealistic, allows us the opportunity to focus on the underlying rationale of these types of statistical inference rather than the calculations.

Later chapters will present inference methods to use in most of the settings we met in learning to explore data. In fact, there are libraries—both of books and of computer software—full of more elaborate statistical techniques. Informed use of any of these methods, however, requires a firm understanding of the underlying reasoning. That is the goal of this chapter. A computer or calculator will do the arithmetic, but *you must still exercise sound judgment based on understanding*.

## Overview of Inference

The purpose of statistical inference is to draw conclusions from data. Formal inference emphasizes substantiating our conclusions via probability calculations. Probability allows us to take chance variation into account. Here is an example.

### Example

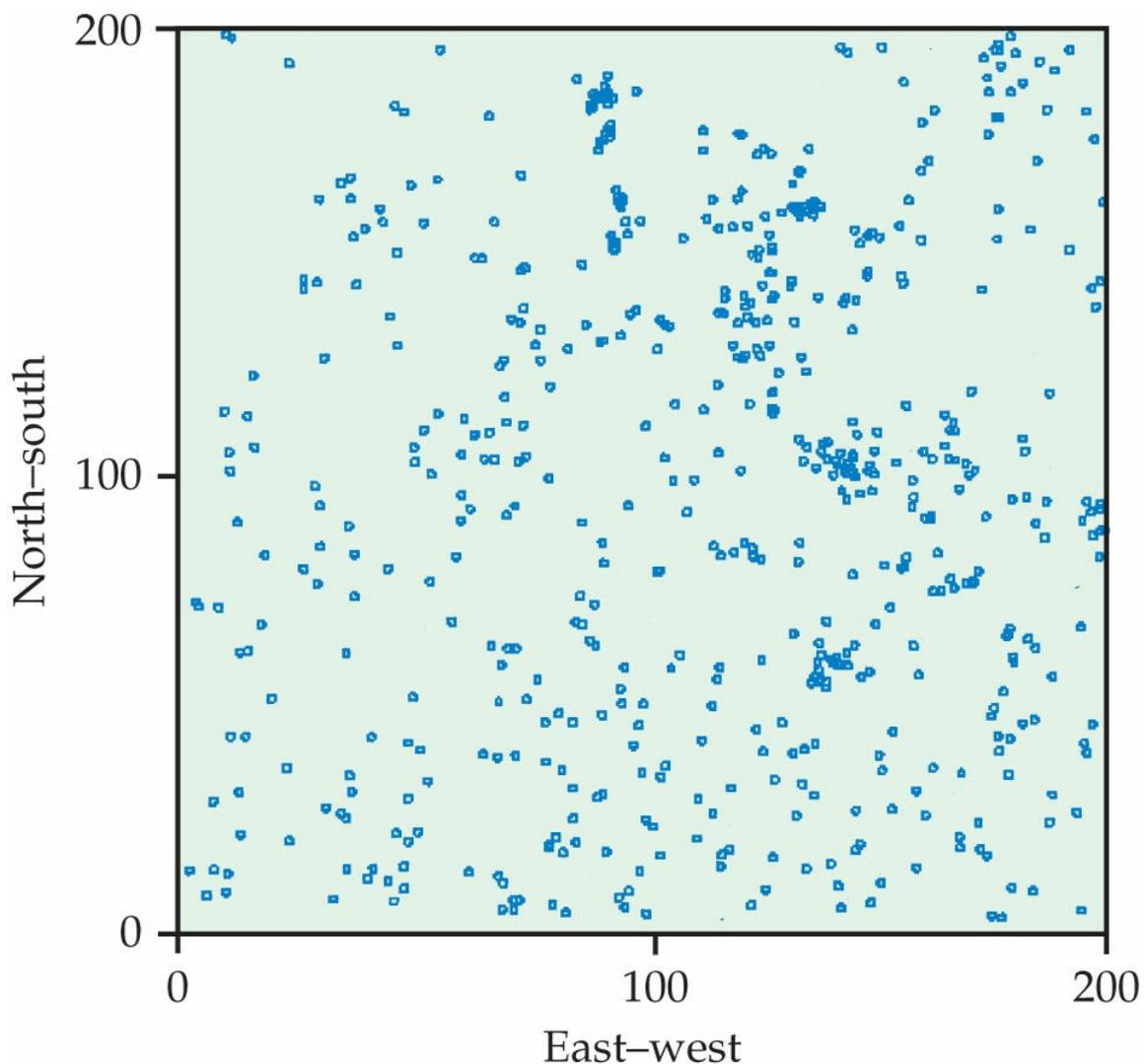
#### 6.1 Clustering of trees in a forest



The Wade Tract in Thomas County, Georgia, is an old-growth forest of longleaf pine trees (*Pinus palustris*) that has survived in a relatively

undisturbed state since before the settlement of the area by Europeans. Foresters who study these trees are interested in how the trees are distributed in the forest. Is there some sort of clustering, resulting in regions of the forest with more trees than others? Or are the tree locations random, resulting in no particular patterns? Figure 6.1 gives a plot of the locations of all 584 longleaf pine trees in a 200-meter by 200-meter region in the Wade Tract.<sup>1</sup>

Do the locations appear to be random, or do there appear to be clusters of trees? One approach to the analysis of these data indicates that a pattern as clustered as, or more clustered than, the one in Figure 6.1 would occur only 4% of the time if, in fact, the locations of longleaf pine trees in the Wade Tract are random. Because this chance is fairly small, we conclude that there is some clustering of these trees.



**Figure 6.1**

The distribution of longleaf pine trees, for Example 6.1.

This probability calculation helps us to distinguish between patterns that are consistent or inconsistent with the random location scenario. Here is an example

assessing a new cold medication—with a different conclusion.

## Example

### 6.2 Effectiveness of a new cold medication

Researchers want to know if a new medication is more effective in relieving cold symptoms than a popular over-the-counter medication. Twenty patients are randomly assigned to receive the new medication, and another 20 receive the popular over-the-counter medication. Fifteen (75%) of those taking the new medication find satisfactory symptom relief versus only 11 (55%) of the popular medication patients.

Our unaided judgment suggests that the new medication is better. However, probability calculations tell us that a difference this large or larger between the results in the two groups of 20 patients would occur about one time in five simply because of chance variation. In this case, it is better to conclude that the data fail to establish a real difference between the two treatments. This probability (nearly 0.19) is too large to ignore.

In this chapter we introduce the two most prominent types of formal statistical inference. Section 6.1 concerns *confidence intervals* for estimating the value of a population parameter. Section 6.2 presents *tests of significance*, which assess the evidence for a claim, such as those in Examples 6.1 and 6.2. Both types of inference are based on the sampling distributions of statistics. That is, both report probabilities that state *what would happen if we used the inference method many times*.



**sampling distribution, p. 208**

This kind of probability statement is characteristic of standard statistical inference. Users of statistics must understand the nature of this reasoning and the meaning of the probability statements that appear, for example, online and in journal articles and statistical software output.

Because the methods of formal inference are based on sampling distributions, they require a probability model for the data. Trustworthy probability models can arise in many ways, but the model is most secure and inference is most reliable when the data are produced by a properly randomized design.



*When you use statistical inference, you are acting as if the data come from a random sample or a randomized experiment. If this is not true, your conclusions may be open to challenge. Do not be overly impressed by the complex details of formal inference. This elaborate machinery cannot remedy basic flaws in producing the data such as voluntary response samples and confounded experiments. Use the common sense developed in your study of the first three chapters of this book, and proceed to detailed formal inference only when you are satisfied that the data deserve such analysis.*

## 6.1 Estimating with Confidence

When you complete this section, you will be able to

- Describe a level  $C$  confidence interval for a population parameter in terms of an estimate and its margin of error.
- Construct a level  $C$  confidence interval for  $\mu$  from an SRS of size  $n$  from a large population having known standard deviation  $\sigma$ .
- Explain how the margin of error changes with a change in the confidence level  $C$ .
- Determine the sample size needed to obtain a specified margin of error for a level  $C$  confidence interval for  $\mu$ .
- Identify situations where inference about  $\mu$  based on the confidence interval  $\bar{x} \pm z^* \sigma / \sqrt{n}$  may be suspect.

The SAT is a widely used measure of readiness for college study. It consists of three sections, one for mathematical reasoning ability (SATM), one for verbal reasoning ability (SATV), and one for writing ability (SATW). Possible scores on each section range from 200 to 800, for a total range of 600 to 2400. Since 1995, section scores have been *recentered* so that the mean is approximately 500 with a standard deviation of 100 in a large “standardized group.” This scale has been maintained so that scores have a constant interpretation.



linear transformations, p. 45

### Example

#### 6.3 Estimating the mean SATM score for seniors in California



← **LOOK BACK**

**law of large numbers, p. 268**

Suppose that you want to estimate the mean SATM score for the 486, 549 high school seniors in California.<sup>2</sup> You know better than to trust data from the students who choose to take the SAT. Only about 38% of California students typically take the SAT. These self-selected students are planning to attend college and are not representative of all California seniors. At considerable effort and expense, you give the test to a simple random sample (SRS) of 500 California high school seniors. The mean score for your sample is  $\bar{x} = 485$

What can you say about the mean score  $\mu$  in the population of all 486, 549 seniors?

The sample mean  $\bar{x}$  is the natural estimator of the unknown population mean  $\mu$ . We know that  $\bar{x}$  is an unbiased estimator of  $\mu$ . More important, the law of large numbers says that the sample mean must approach the population mean as the size of the sample grows. The value  $\bar{x} = 485$  therefore appears to be a reasonable estimate of the mean score  $\mu$  that all 486, 549 students would achieve if they took the test.

But how reliable is this estimate? A second sample of 500 students would surely not give a sample mean of 485 again. Unbiasedness says only that there is no systematic tendency to underestimate or overestimate the truth. Could we plausibly get a sample mean of 465 or 510 in repeated samples? *An estimate without an indication of its variability is of little value.*

## Statistical confidence



**unbiased estimator, p. 210**

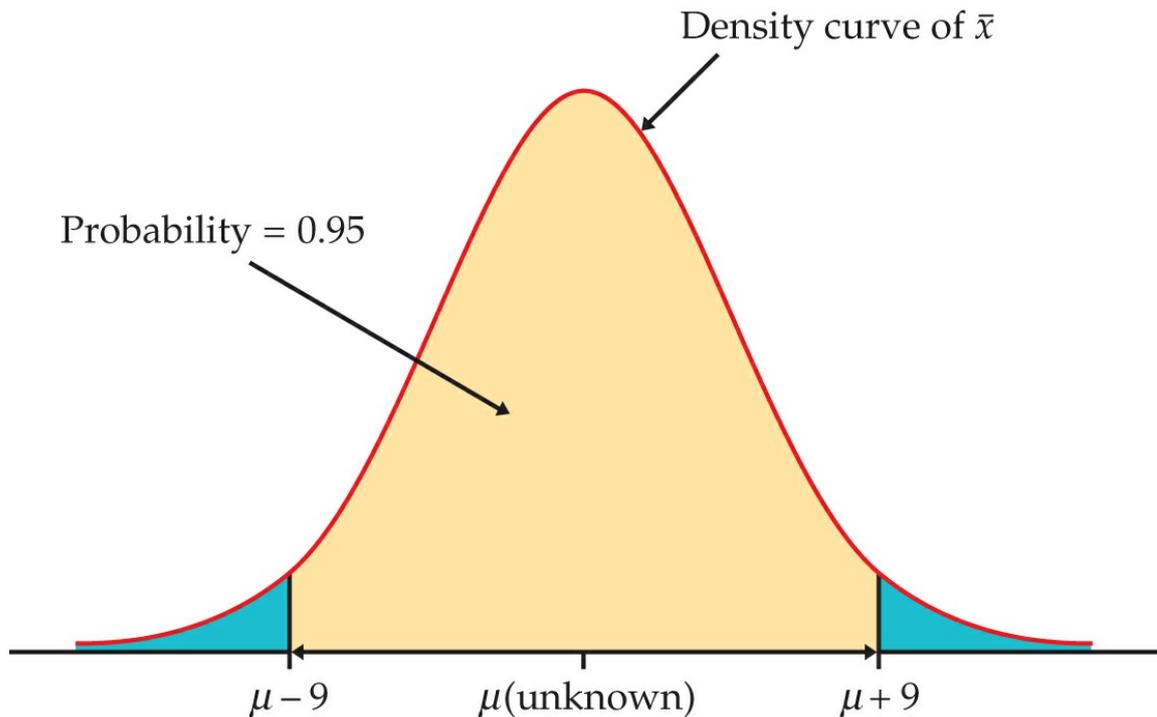
The unbiasedness of an estimator concerns the center of its sampling distribution, but questions about variation are answered by looking at its spread. We know that if the entire population of SATM scores has mean  $\mu$  and standard deviation  $\sigma$  then in repeated samples of size 500 the sample mean  $\bar{x}$  is approximately  $N(\mu, \sigma/\sqrt{500})$ . Let us suppose that we know that the standard deviation  $\sigma$  of SATM scores in our California population is  $\sigma = 100$ . (We will see in the next chapter how to proceed when  $\sigma$  is not known. For now, we are more interested in statistical reasoning than in details of realistic methods.) This means that in repeated sampling the sample mean  $\bar{x}$  has an approximately Normal distribution centered at the unknown population mean  $\mu$  and a standard deviation of

$$\sigma_{\bar{x}} = 100/\sqrt{500} = 4.5$$



**central limit theorem, p. 307**

Now we are ready to proceed. Consider this line of thought, which is illustrated by Figure 6.2:



**Figure 6.2**

Distribution of the sample mean for Example 6.3.  $\bar{x}$  lies within  $\pm 9$  points of  $\mu$  in 95% of all samples. This also means that  $\mu$  is within  $\pm 9$  points of  $\bar{x}$  in those samples.

- The 68–95–99.7 rule says that the probability is about 0.95 that  $\bar{x}$  will be within 9 points (that is, two standard deviations of  $\bar{x}$ ) of the population mean score  $\mu$ .
- To say that  $\bar{x}$  lies within 9 points of  $\mu$  is the same as saying that  $\mu$  is within 9 points of  $\bar{x}$ .
- So about 95% of all samples will contain the true  $\mu$  in the interval from  $\bar{x} - 9$  to  $\bar{x} + 9$ .

We have simply restated a fact about the sampling distribution of  $\bar{x}$ . *The language of statistical inference uses this fact about what would happen in the long run to express our confidence in the results of any one sample.* Our sample gave  $\bar{x} = 485$ . We say that we are *95% confident* that the unknown mean score for all California seniors lies between

$$\bar{x} - 9 = 485 - 9 = 476$$

and

$$\bar{x} + 9 = 485 + 9 = 494$$

Be sure you understand the grounds for our confidence. There are only two possibilities for our SRS:

1. The interval between 476 and 494 contains the true  $\mu$ .
2. The interval between 476 and 494 does not contain the true  $\mu$ .

We cannot know whether our sample is one of the 95% for which the interval  $\bar{x} \pm 9$

contains  $\mu$  or one of the unlucky 5% for which it does not contain  $\mu$ . The statement that we are 95% confident is shorthand for saying, “We arrived at these numbers by a method that gives correct results 95% of the time.”

## USE YOUR KNOWLEDGE

### 6.1 How much do you spend on lunch?

The average amount you spend on a lunch during the week is not known. Based on past experience, you are willing to assume that the standard deviation is \$2.40. If you take a random sample of 36 lunches, what is the value of the standard deviation of  $\bar{x}$ ?

### 6.2 Applying the 68–95–99.7 rule

In the setting of the previous exercise, the 68–95–99.7 rule says that the probability is about 0.95 that  $\bar{x}$  is within \$\_\_\_\_\_ of the population mean  $\mu$ . Fill in the blank.

### 6.3 Constructing a 95% confidence interval

In the setting of the previous two exercises, about 95% of all samples will capture the true mean in the interval  $\bar{x}$  plus or minus \$\_\_\_\_\_. Fill in the blank.

## Confidence intervals

In the setting of Example 6.3, the interval of numbers between the values  $\bar{x} \pm 9$  is called a *95% confidence interval* for  $\mu$ . Like most confidence intervals we will discuss, this one has the form

$$\text{estimate} \pm \text{margin of error}$$

The estimate ( $\bar{x}=485$  in this case) is our guess for the value of the unknown parameter. The **margin of error** (9 here) reflects how accurate we believe our guess is, based on the variability of the estimate, and how confident we are that the procedure will produce an interval that will contain the true population mean  $\mu$ .

margin of error

Figure 6.3 illustrates the behavior of 95% confidence intervals in repeated sampling from a Normal distribution with mean  $\mu$ . The center of each interval (marked by a dot) is at  $\bar{x}$  and varies from sample to sample. The sampling distribution of  $\bar{x}$  (also Normal) appears at the top of the figure to show the long-term pattern of this variation.

The 95% confidence intervals,  $\bar{x} \pm \text{margin of error}$ , from 25 SRSs appear below the sampling distribution. The arrows on either side of the dot ( $\bar{x}$ ) span the confidence interval. All except one of the 25 intervals contain the true value of  $\mu$ . In those intervals that contain  $\mu$  sometimes  $\mu$  is near the middle of the interval and sometimes it is closer to one of the ends. This again reflects the variation of  $\bar{x}$ . In practice, we don't know the value of  $\mu$ , but we have a method such that, in a very large number of samples, 95% of the confidence intervals will contain  $\mu$ .

Statisticians have constructed confidence intervals for many different parameters based on a variety of designs for data collection. We will meet a number of these in later chapters. Two important things about a confidence interval are common to all settings:

1. It is an interval of the form  $(a,b)$ , where  $a$  and  $b$  are numbers computed from the sample data.
2. It has a property called a confidence level that gives the probability of producing an interval that contains the unknown parameter.

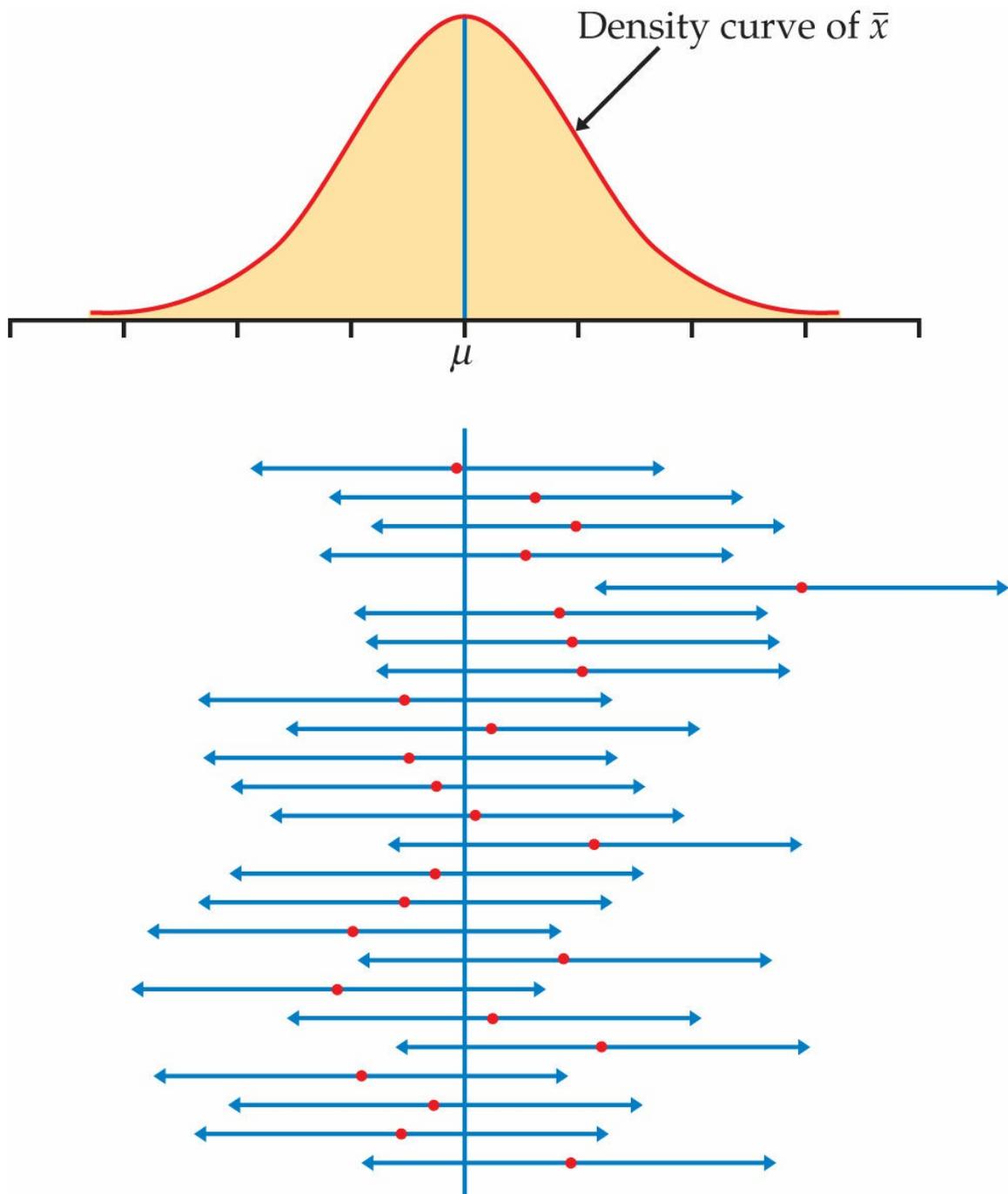
Users can choose the confidence level, but 95% is the standard for most situations. Occasionally, 90% or 99% is used. We will use  $C$  to stand for the confidence level in decimal form. For example, a 95% confidence level corresponds to  $C=0.95$ .

## CONFIDENCE INTERVAL

A level  $C$  **confidence interval** for a parameter is an interval computed from sample data by a method that has probability  $C$  of producing an interval containing the true value of the parameter.

With the *Confidence Interval* applet, you can construct diagrams similar to the one displayed in Figure 6.3. The only difference is that the applet displays the Normal population distribution at the top rather than the Normal sampling distribution of  $\bar{x}$ . You choose the confidence level  $C$ , the sample size  $n$ , and whether you want to generate 1 or 25 samples at a time. A running total (and percent) of the number of intervals that contain  $\mu$  is displayed so you can consider a larger number of samples.





**Figure 6.3**

Twenty-five samples from the same population gave these 95% confidence intervals. In the long run, 95% of all samples give an interval that covers  $\mu$ . The sampling distribution of  $\bar{x}$  is shown at the top.

When generating single samples, the data for the latest SRS are shown below the confidence interval. The spread in these data reflects the spread of the population distribution. This spread is assumed known, and it does not change with sample size. What does change, as you vary  $\bar{x}$  is the margin of error, since it reflects the uncertainty in the estimate of  $\mu$ . As you increase  $n$ , you'll find that the span of the confidence interval gets smaller and smaller.

## USE YOUR KNOWLEDGE

### 6.4 Generating a single confidence interval



Using the default settings in the *Confidence Interval* applet (95% confidence level and  $n=20$ ), click “Sample” to choose an SRS and display its confidence interval.

(a) Is the spread in the data, shown as yellow dots below the confidence interval, larger than the span of the confidence interval? Explain why this would typically be the case.

(b) For the same data set, you can compare the span of the confidence interval for different values of  $C$  by sliding the confidence level to a new value. For the SRS you generated in part (a), what happens to the span of the interval when you move  $C$  to 99%? What about 90%? Describe the relationship you find between the confidence level  $C$  and the span of the confidence interval.

### 6.5 80% confidence intervals



The idea of an 80% confidence interval is that the interval captures the true parameter value in 80% of all samples. That’s not high enough confidence for practical use, but 80% hits and 20% misses make it easy to see how a confidence interval behaves in repeated samples from the same population.

(a) Set the confidence level in the *Confidence Interval* applet to 80%. Click “Sample 25” to choose 25 SRSs and display their confidence intervals. How many of the 25 intervals contain the true mean  $\mu$ ? What proportion contain the true mean?

(b) We can’t determine whether a new SRS will result in an interval that contains  $\mu$  or not. The confidence level only tells us what percent will contain  $\mu$  in the long run. Click “Sample 25” again to get the confidence intervals from 50 SRSs. What proportion hit? Keep clicking “Sample

25” and record the proportion of hits among 100, 200, 300, 400, and 500 SRSs. As the number of samples increases, we expect the percent of captures to get closer to the confidence level, 80%. Do you find this pattern in your results?

## Confidence interval for a population mean

We will now construct a level  $C$  confidence interval for the mean  $\mu$  of a population when the data are an SRS of size  $n$ . The construction is based on the sampling distribution of the sample mean  $\bar{x}$ . This distribution is exactly  $N(\mu, \sigma/n)$  when the population has the  $N(\mu, \sigma)$  distribution. The central limit theorem says that this same sampling distribution is approximately correct for large samples whenever the population mean and standard deviation are  $\mu$  and  $\sigma$ . For now, we will assume we are in one of these two situations. We will discuss what we mean by “large sample” after we briefly study these intervals.

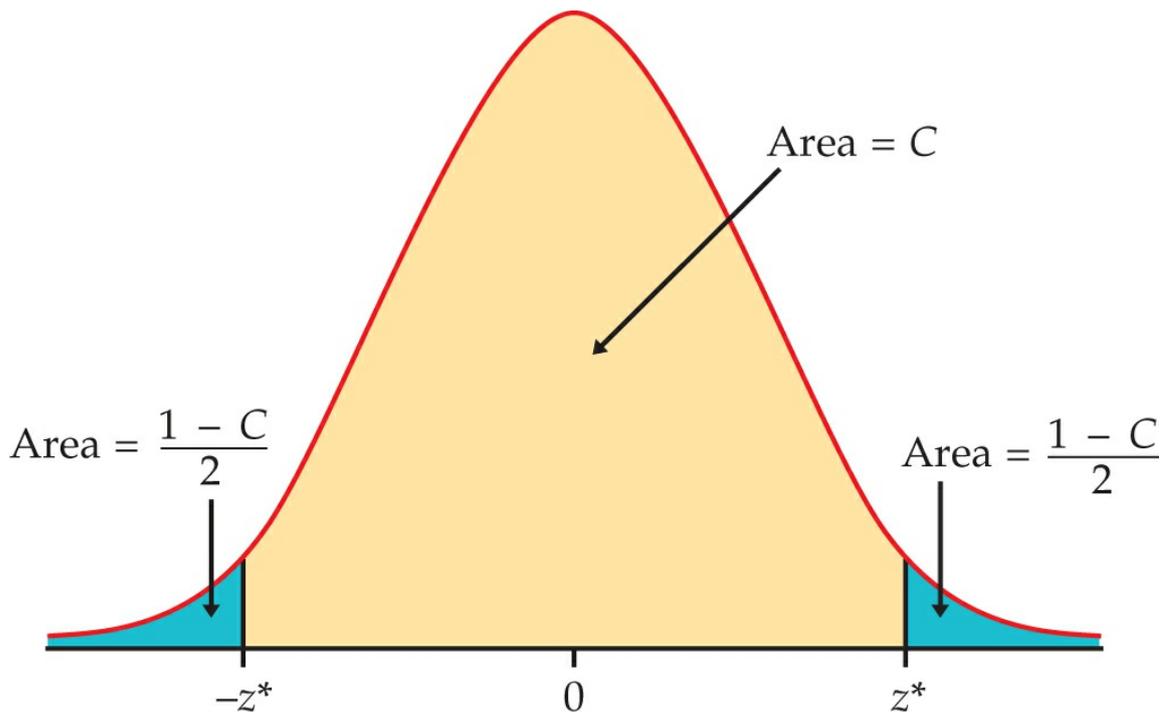


**central limit theorem, p. 307**

Our construction of a 95% confidence interval for the mean SATM score began by noting that any Normal distribution has probability about 0.95 within  $\pm 2$  standard deviations of its mean. To construct a level  $C$  confidence interval we first catch the central  $C$  area under a Normal curve. That is, we must find the number  $z^*$  such that any Normal distribution has probability  $C$  within  $\pm z^*$  standard deviations of its mean.

Because all Normal distributions have the same standardized form, we can obtain everything we need from the standard Normal curve. Figure 6.4 shows how  $C$  and  $z^*$  are related. Values of  $z^*$  for many choices of  $C$  appear in the row labeled  $z^*$  at the bottom of Table D. Here are the most important entries from that row:

$z^*$	1.645	1.960	2.576
$C$	90%	95%	99%



**Figure 6.4**

To construct a level  $C$  confidence interval, we must find the number  $z^*$ . The area between  $-z^*$  and  $z^*$  under the standard Normal curve is  $C$ .

Notice that for 95% confidence the value 2 obtained from the 68-95-99.7 rule is replaced with the more precise 1.96.

As Figure 6.4 reminds us, any Normal curve has probability  $C$  between the point  $z^*$  standard deviations below the mean and the point  $z^*$  standard deviations above the mean. The sample mean  $\bar{x}$  has the Normal distribution with mean  $\mu$  and standard deviation  $\sigma/n$ , so there is probability  $C$  that  $\bar{x}$  lies between

$$\mu - z^* \sigma/n \quad \text{and} \quad \mu + z^* \sigma/n$$

This is exactly the same as saying that the unknown population mean  $\mu$  lies between

$$\bar{x} - z^* \sigma/n \quad \text{and} \quad \bar{x} + z^* \sigma/n$$

That is, there is probability  $C$  that the interval  $\bar{x} \pm z^* \sigma/n$  contains  $\mu$ . This is our confidence interval. The estimate of the unknown  $\mu$  is  $\bar{x}$  and the margin of error is  $z^* \sigma/n$ .

## CONFIDENCE INTERVAL FOR A POPULATION MEAN

Choose an SRS of size  $n$  from a population having unknown mean  $\mu$  and known standard deviation  $\sigma$ . The **margin of error** for a level  $C$  confidence interval for  $\mu$  is

$$m = z^* \sigma/n$$

Here  $z^*$  is the value on the standard Normal curve with area  $C$  between the critical points  $-z^*$  and  $z^*$ . The level  $C$  **confidence interval** for  $\mu$  is

$$\bar{x} \pm m$$

The confidence level of this interval is exactly  $C$  when the population distribution is Normal and is approximately  $C$  when  $n$  is large in other cases.

## Example

### 6.4 Average credit card balance among college students

Starting in 2008, Sallie Mae, a major provider of education loans and savings programs, has conducted an annual study titled “How America Pays for College.” Unlike other studies on college funding, this study assesses all aspects of spending and borrowing, for both educational and noneducational purposes. In the 2012 survey, 1601 randomly selected individuals (817 parents of undergraduate students and 784 undergraduate students) were surveyed by telephone.<sup>3</sup>

Many of the survey questions focused on the undergraduate student, so the parents in the survey were responding for their children. Do you think we should combine responses across these two groups? Do you think your parents are fully aware of your spending and borrowing habits? The authors reported overall averages and percents in their report but did break things down by group in their data tables. For now, we will consider this a sample from one population, but we will revisit this issue later.

One survey question asked about the undergraduate’s current total outstanding balance on credit cards. Of the 1601 who were surveyed, only  $n=532$  provided an answer. *Nonresponse should always be considered as a source of bias.* In this case, the authors believed this nonresponse to be an ignorable source of bias and proceeded by treating the  $n=532$  sample as if it were a random sample. We will do the same.



The average credit card balance was \$755. The median balance was \$196, so this distribution is clearly skewed. Nevertheless, because the sample size is quite large, we can rely on the central limit theorem to assure us that the

confidence interval based on the Normal distribution will be a good approximation.

Let's compute an approximate 95% confidence interval for the true mean credit card balance among all undergraduates. We'll assume that the standard deviation for the population of credit card debts is \$1130. For 95% confidence, we see from Table D that  $z^*=1.960$ . The margin of error for the 95% confidence interval for  $\mu$  is therefore

$$\begin{aligned}m &= z^* \sigma n \\ &= 1.960(1130)(532) \\ &= 96.02\end{aligned}$$

We have computed the margin of error with more digits than we really need. Our mean is rounded to the nearest \$1, so we will do the same for the margin of error. Keeping additional digits would provide no additional useful information. Therefore, we will use  $m=96$ . The approximate 95% confidence interval is

$$\begin{aligned}x \pm m &= 755 \pm 96 \\ &= (659, 851)\end{aligned}$$

We are 95% confident that the average credit card debt among all undergraduates is between \$659 and \$851.

Suppose that the researchers who designed this study had used a different sample size. How would this affect the confidence interval? We can answer this question by changing the sample size in our calculations and assuming that the sample mean is the same.

## Example

### 6.5 How sample size affects the confidence interval

As in Example 6.4, the sample mean of the credit card debt is \$755 and the population standard deviation is \$1130. Suppose that the sample size is only 133 but still large enough for us to rely on the central limit theorem. In this case, the margin of error for 95% confidence is

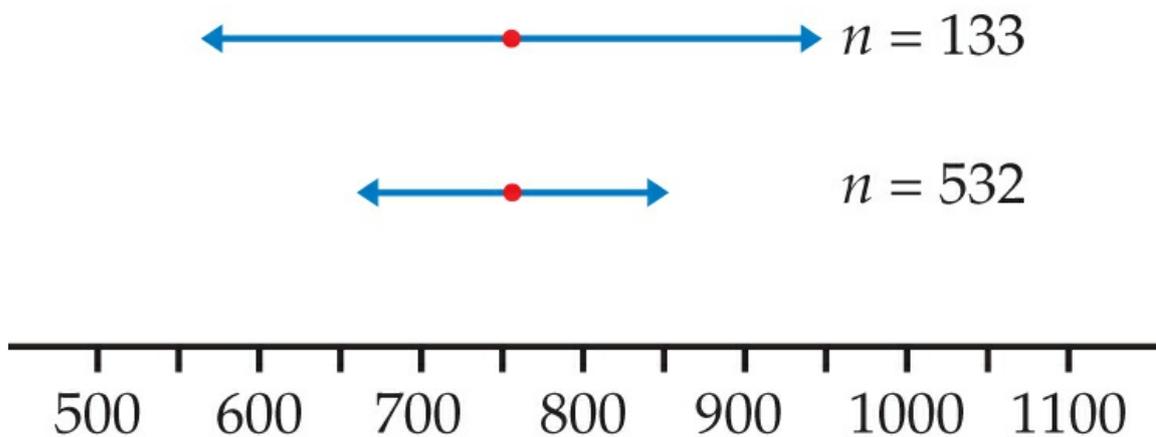
$$\begin{aligned}m &= z^* \sigma n \\ &= 1.960(1130)(133)\end{aligned}$$

$$=192.05$$

and the approximate 95% confidence interval is

$$\begin{aligned}x \pm m &= 755 \pm 192 \\ &(563, 947)\end{aligned}$$

Notice that the margin of error for this example is twice as large as the margin of error that we computed in Example 6.4. The only change that we made was to assume that the sample size is 133 rather than 532. This sample size is one-fourth of the original 532. Thus, we double the margin of error when we reduce the sample size to one-fourth of the original value. Figure 6.5 illustrates the effect in terms of the intervals.



**Figure 6.5**

Confidence intervals for  $n = 532$  and  $n = 133$ , for Example 6.4 and 6.5. A sample size 4 times as large results in a confidence interval that is half as wide.

## USE YOUR KNOWLEDGE

### 6.6 Average amount paid for college

Refer to Example 6.4 (page 360). The average annual amount the  $n = 1601$  families paid for college was \$20,902.<sup>4</sup> If the population standard deviation is \$7500, give the 95% confidence interval for  $\mu$  the average amount a family pays for a college undergraduate.

### 6.7 Changing the sample size

In the setting of the previous exercise, would the margin of error for

95% confidence be roughly doubled or halved if the sample size were raised to  $n = 6400$ ? Verify your answer by performing the calculations.

### 6.8 Changing the confidence level

In the setting of Exercise 6.6, would the margin of error for 99% confidence be larger or smaller? Verify your answer by performing the calculations.

The argument leading to the form of confidence intervals for the population mean  $\mu$  rested on the fact that the statistic  $\bar{x}$  used to estimate  $\mu$  has a Normal distribution. Because many sample estimates have Normal distributions (at least approximately), it is useful to notice that the confidence interval has the form

$$\text{estimate} \pm z^* \sigma_{\text{estimate}}$$

The estimate based on the sample is the center of the confidence interval. The margin of error is  $z^* \sigma_{\text{estimate}}$ . The desired confidence level determines  $z^*$  from Table D. The standard deviation of the estimate is found from knowledge of the sampling distribution in a particular case. When the estimate is  $\bar{x}$  from an SRS, the standard deviation of the estimate is  $\sigma_{\text{estimate}} = \sigma/n$ . We will return to this general form numerous times in the following chapters.

### How confidence intervals behave

The margin of error  $z^* \sigma/n$  for the mean of a Normal population illustrates several important properties that are shared by all confidence intervals in common use. The user chooses the confidence level, and the margin of error follows from this choice.

Both high confidence and a small margin of error are desirable characteristics of a confidence interval. High confidence says that our method almost always gives correct answers. A small margin of error says that we have pinned down the parameter quite precisely.

Suppose that in planning a study you calculate the margin of error and decide that it is too large. Here are your choices to reduce it:

- Use a lower level of confidence (smaller  $C$ ).
- Choose a larger sample size (larger  $n$ ).
- Reduce  $\sigma$ .

For most problems, you would choose a confidence level of 90%, 95%, or 99%, so  $z^*$  will be 1.645, 1.960, or 2.576, respectively. Figure 6.4 shows that  $z^*$  will be smaller for lower confidence (smaller  $C$ ). The bottom row of Table D also shows this. If  $n$  and  $\sigma$  are unchanged, a smaller  $z^*$  leads to a smaller margin of error.

## Example

### 6.6 How the confidence level affects the confidence interval

Suppose that for the student credit card data in Example 6.4 (page 360), we wanted 99% confidence. Table D tells us that for 99% confidence,  $z^*=2.576$ . The margin of error for 99% confidence based on 532 observations is

$$=2.5761130532$$

$$m=z^*\sigma n$$

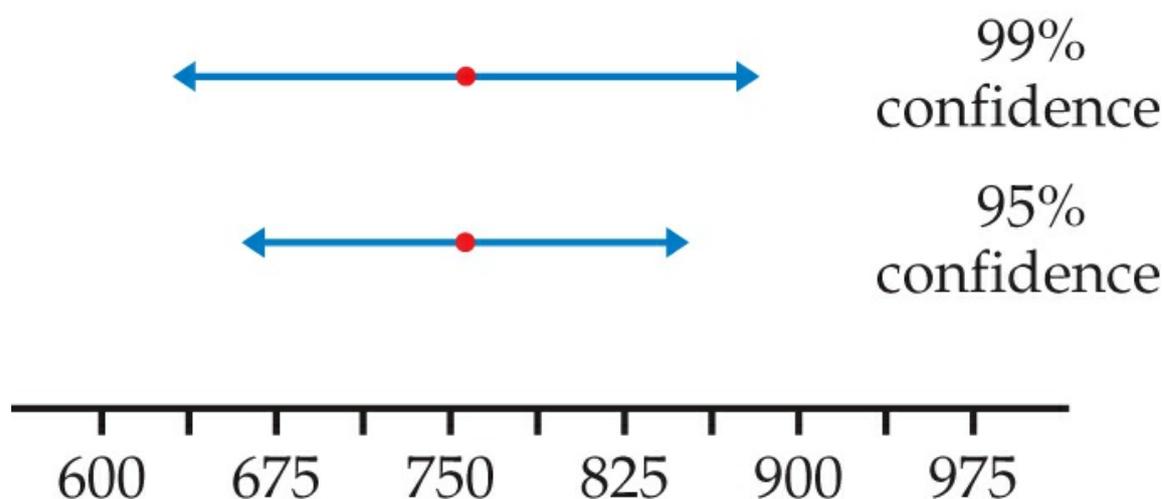
$$=126.20$$

and the 99% confidence interval is

$$\bar{x}\pm m=755\pm 126$$

$$(629,881)$$

Requiring 99%, rather than 95%, confidence has increased the margin of error from 96 to 126. Figure 6.6 compares the two intervals.



**Figure 6.6**

Confidence intervals for Examples 6.4 and 6.6. The larger the value of  $C$ , the wider the interval.

Similarly, choosing a larger the sample size  $n$  reduces the margin of error for any fixed confidence level. The square root in the formula implies that we must multiply the number of observations by 4 in order to cut the margin of error in half. Likewise, if we want to reduce the standard deviation of  $\bar{x}$  by a factor of 4, we must take a sample 16 times as large.

The standard deviation  $\sigma$  measures the variation in the population. You can think of the variation among individuals in the population as noise that obscures the average value  $\mu$ . It is harder to pin down the mean  $\mu$  of a highly variable

population; that is why the margin of error of a confidence interval increases with  $\sigma$ .

In practice, we can sometimes reduce  $\sigma$  by carefully controlling the measurement process. We also might change the mean of interest by restricting our attention to only part of a large population. Focusing on a subpopulation will often result in a smaller  $\sigma$ .

## Choosing the sample size

*A wise user of statistics never plans data collection without, at the same time, planning the inference.* You can arrange to have both high confidence and a small margin of error. The margin of error of the confidence interval for a population mean is

$$n=(z^*\sigma m)^2$$

Do notice once again that it is the size of the *sample* that determines the margin of error. The size of the *population* (as long as the population is much larger than the sample) does not influence the sample size we need.

To obtain a desired margin of error  $m$ , plug in the value of  $\sigma$  and the value of  $z^*$  for your desired confidence level, and solve for the sample size  $n$ . Here is the result.

### SAMPLE SIZE FOR DESIRED MARGIN OF ERROR

The confidence interval for a population mean will have a specified margin of error  $m$  when the sample size is

$$n=(z^*\sigma m)^2=(1.96\times 113050)^2=1962.14$$

This formula also does not account for collection costs. In practice, taking observations costs time and money. The required sample size may be impossibly expensive. In those situations, you might consider a larger margin of error and/or a lower confidence level to find a workable sample size.

## Example

### 6.7 How many undergraduates should we survey?

Suppose that we are planning a credit card use survey similar to the one described in Example 6.4. If we want the margin of error to be \$50 with 95% confidence, what sample size  $n$  do we need? For 95% confidence, Table D gives  $z^*=1.960$ . For  $\sigma$  we will use the value from the previous study, \$1130. If the margin of error is \$50, we have

$$n=(z^*\sigma/m)^2=(1.96\times 1130/50)^2=3065.84$$

Because 1962 measurements will give a slightly wider interval than desired and 1963 measurements a slightly narrower interval, we should choose  $n=1963$ . We need information from 1963 undergraduates to determine an estimate of mean debt with the desired margin of error.

It is always safe to round *up* to the next higher whole number when finding  $n$  because this will give us a smaller margin of error. The purpose of this calculation is to determine a sample size that is sufficient to provide useful results, but the determination of what is useful is a matter of judgment.

Would we need a much larger sample size to obtain a margin of error of \$40? Here is the calculation:

$$m=z^*\sigma/n$$

A sample of  $n=3066$  is much larger, and the costs of such a large sample may be prohibitive.

*Unfortunately, the actual number of usable observations is often less than what we plan at the beginning of a study.* This is particularly true of data collected in surveys but is an important consideration in most studies. Careful study designers often assume a nonresponse rate or dropout rate that specifies what proportion of the originally planned sample will fail to provide data. We use this information to calculate the sample size to be used at the start of the study.



For example, if in the preceding survey we expect only 25% of those contacted to respond, we would need to start with a sample size of  $4\times 1963=7852$  to obtain usable information from 1963 undergraduates and parents of undergraduates.

## USE YOUR KNOWLEDGE

### 6.9 Starting salaries

You are planning a survey of starting salaries for recent computer science majors. In the latest survey by the National Association of Colleges and Employers, the average starting salary was reported to be \$60,038.<sup>5</sup> If you assume that the standard deviation is \$4300, what sample size do you need to have a margin of error equal to \$500 with 95% confidence?

### 6.10 Changes in sample size

Suppose that in the setting of the previous exercise you have the resources to contact 400 recent graduates. If all respond, will your margin of error be larger or smaller than \$500? What if only 50% respond? Verify your answers by performing the calculations.

### Some cautions



We have already seen that small margins of error and high confidence can require large numbers of observations. You should also be keenly aware that *any formula for inference is correct only in specific circumstances*. If the government required statistical procedures to carry warning labels like those on drugs, most inference methods would have long labels. Our handy formula  $\bar{x} \pm z^* \sigma/n$  for estimating a population mean comes with the following list of warnings for the user:

- The data should be an SRS from the population. We are completely safe if we actually did a randomization and drew an SRS. We are not in great danger if the data can plausibly be thought of as independent observations from a population. That is the case in Examples 6.4 to 6.7, where we redefine our population to correspond to survey respondents.
- The formula is not correct for probability sampling designs more complex than an SRS. Correct methods for other designs are available. We will not discuss confidence intervals based on multistage or stratified samples (page 197). If you plan such samples, be sure that you (or your statistical consultant) know how to carry out the inference you desire.
- There is no correct method for inference from data haphazardly collected with bias of unknown size. Fancy formulas cannot rescue badly produced data.

 **LOOK BACK**  
resistant measure, p. 32

- Because  $\bar{x}$  is not a resistant measure, outliers can have a large effect on the confidence interval. *You should search for outliers and try to correct them or justify their removal before computing the interval.* If the outliers cannot be removed, ask your statistical consultant about procedures that are not sensitive to outliers.
- If the sample size is small and the population is not Normal, the true confidence level will be different from the value  $C$  used in computing the interval. *Prior to any calculations, examine your data carefully for skewness and other signs of non-Normality.* Remember though that the interval relies only on the distribution of  $\bar{x}$ , which even for quite small sample sizes is much closer to Normal than is the distribution of the individual observations. When  $n \geq 15$ , the confidence level is not greatly disturbed by non-Normal populations unless extreme outliers or quite strong skewness are present. Our debt data in Example 6.4 are clearly skewed, but because of the large sample size, we are confident that the distribution of the sample mean will be approximately Normal.
- The interval  $\bar{x} \pm z^* \sigma/n$  assumes that the standard deviation  $\sigma$  of the population is known. This unrealistic requirement renders the interval of little use in statistical practice. We will learn in the next chapter what to do when  $\sigma$  is unknown. If, however, the sample is large, the sample standard deviation  $s$  will be close to the unknown  $\sigma$ . The interval  $\bar{x} \pm z^* s/n$  is then an approximate confidence interval for  $\mu$ .

 **LOOK BACK**  
**standard deviation  $s$ , p. 42**

The most important caution concerning confidence intervals is a consequence of the first of these warnings. *The margin of error in a confidence interval covers only random sampling errors.* The margin of error is obtained from the sampling distribution and indicates how much error can be expected because of chance variation in randomized data production.



*Practical difficulties such as undercoverage and nonresponse in a sample survey cause additional errors. These errors can be larger than the random sampling error.* This often happens when the sample size is large (so that  $\sigma/n$  is small). Remember this unpleasant fact when reading the results of an opinion poll or other sample survey. The practical conduct of the survey influences the trustworthiness of its results in ways that are not included in the announced margin of error.

Every inference procedure that we will meet has its own list of warnings. Because many of the warnings are similar to those we have mentioned, we will not print the full warning label each time. It is easy to state (from the mathematics of

probability) conditions under which a method of inference is exactly correct. These conditions are *never* fully met in practice.

For example, no population is exactly Normal. *Deciding when a statistical procedure should be used in practice often requires judgment assisted by exploratory analysis of the data.* Mathematical facts are therefore only a part of statistics. The difference between statistics and mathematics can be stated thusly: mathematical theorems are true; statistical methods are often effective when used with skill.

Finally, you should understand what statistical confidence does not say. Based on our SRS in Example 6.3, we are 95% confident that the mean SATM score for the California students lies between 476 and 494. This says that this interval was calculated by a method that gives correct results in 95% of all possible samples. It does *not* say that the probability is 0.95 that the true mean falls between 476 and 494. *No randomness remains after we draw a particular sample and compute the interval.* The true mean either is or is not between 476 and 494. The probability calculations of standard statistical inference describe how often the *method*, not a particular sample, gives correct answers.

## USE YOUR KNOWLEDGE

### 6.11 Nonresponse in a survey

Let's revisit Example 6.4 (page 360). Of the 1601 participants in the survey, only 532 reported the undergraduate's outstanding credit card balance. For that example, we proceeded as if we had a random sample and calculated a margin of error at 95% confidence of \$96. Provide a couple of reasons why a survey respondent might not provide an estimate. Based on these reasons, do you think that this margin of error of \$96 is a good measure of the accuracy of the survey's results? Explain your answer.

## BEYOND THE BASICS

### The bootstrap

Confidence intervals are based on sampling distributions. In this section we have used the fact that the sampling distribution of  $\bar{x}$  is  $N(\mu, \sigma/\sqrt{n})$  when the

data are an SRS from an  $N(\mu, \sigma)$  population. If the data are not Normal, the central limit theorem tells us that this sampling distribution is still a reasonable approximation as long as the distribution of the data is not strongly skewed and there are no outliers. Even a fair amount of skewness can be tolerated when the sample size is large.

What if the population does not appear to be Normal and we have only a small sample? Then we do not know what the sampling distribution of  $\bar{x}$  looks like. The **bootstrap** is a procedure for approximating sampling distributions when theory cannot tell us their shape.<sup>6</sup>

### bootstrap

The basic idea is to act as if our sample were the population. We take many samples from it. Each of these is called a **resample**. We calculate the mean  $\bar{x}$  for each resample. We get different results from different resamples because we sample *with replacement*. An individual observation in the original sample can appear more than once in the resample.

### resample

For example, suppose that we have four measurements of a student's daily time spent online last month (in minutes):

190.5 109.0 95.5 137.0

one resample could be

109.0 95.5 137.0 109.0

with  $\bar{x}=112.625$ . Collect the  $\bar{x}$ 's from 1000 such resamples. Their distribution will be close to what we would get if we took 1000 samples from the entire population. We treat the distribution of  $\bar{x}$ 's from our 1000 resamples as if it were the sampling distribution. If we want a 95% confidence interval, for example, we could use the middle 95% of this distribution.

The bootstrap is practical only when you can use a computer to take 1000 or more samples quickly. It is an example of how the use of fast and easy computing is changing the way we do statistics. More details about the bootstrap can be found in Chapter 16.

## SECTION 6.1 Summary

The purpose of a **confidence interval** is to estimate an unknown parameter with an indication of how accurate the estimate is and of how confident we are that the result is correct.

Any confidence interval has two parts: an interval computed from the data and a confidence level. The interval often has the form

estimate  $\pm$  margin of error

The **confidence level** states the probability that the method will give a correct answer. That is, if you use 95% confidence intervals, in the long run 95% of your intervals will contain the true parameter value. When you apply the method once (that is, to a single sample), you do not know if your interval gave a correct answer (this happens 95% of the time) or not (this happens 5% of the time).

The **margin of error** for a level  $C$  confidence interval for the mean  $\mu$  of a Normal population with known standard deviation  $\sigma$ , based on an SRS of size  $n$ , is given by

$$n = (z^* \sigma / m)^2$$

Here  $z^*$  is obtained from the row labeled  $z^*$  at the bottom of Table D. The probability is  $C$  that a standard Normal random variable takes a value between  $-z^*$  and  $z^*$ . The confidence interval is

$$\bar{x} \pm m$$

If the population is not Normal and  $n$  is large, the confidence level of this interval is approximately correct.

Other things being equal, the margin of error of a confidence interval decreases as

- the confidence level  $C$  decreases,
- the sample size  $n$  increases, and
- the population standard deviation  $\sigma$  decreases.

The sample size  $n$  required to obtain a confidence interval of specified margin of error  $m$  for a population mean is

$$z = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard deviation of the estimate}}$$

where  $z^*$  is the critical point for the desired level of confidence.

A specific confidence interval formula is correct only under specific conditions. The most important conditions concern the method used to produce the data. Other factors such as the form of the population distribution may also be important. These conditions should be investigated *prior* to any calculations.

## SECTION 6.1 Exercises

*For Exercise 6.1 to 6.3, see page 356; for Exercises 6.4 and 6.5, see page 358; for Exercises 6.6 to 6.8, see page 362; for Exercises 6.9 and 6.10, see page 365; and for Exercise 6.11, see page 367.*

### 6.12 Margin of error and the confidence interval.

A stress level study based on a random sample of 49 undergraduates at your university reported a mean of 73 (on a 0 to 100 scale) with a margin of error of 8 for 95% confidence.

(a) Give the 95% confidence interval.

(b) If you wanted 99% confidence for the same study, would your margin of error be greater than, equal to, or less than 8? Explain your answer.

### 6.13 Changing the sample size.

Consider the setting of the previous exercise. Suppose that the sample mean is again 73 and the population standard deviation is 28. Make a diagram similar to Figure 6.5 (page 361) that illustrates the effect of sample size on the width of a 95% interval. Use the following sample sizes: 10, 20, 40, and 80. Summarize what the diagram shows.

### 6.14 Changing the confidence level.

Consider the setting of the previous two exercises. Suppose that the sample mean is still 73, the sample size is 49, and the population standard deviation is 28. Make a diagram similar to Figure 6.6 (page 363) that illustrates the effect of the confidence level on the width of the interval. Use 80%, 90%, 95%, and 99%. Summarize what the diagram shows.

### 6.15 Confidence interval mistakes and misunderstandings.

Suppose that 500 randomly selected alumni of the University of Okoboji were asked to rate the university's academic advising services on a 1 to 10 scale. The sample mean  $\bar{x}$  was found to be 8.6. Assume that the population standard deviation is known to be  $\sigma=2.2$ .

(a) Ima Bitlost computes the 95% confidence interval for the average satisfaction score as  $8.6 \pm 1.96(2.2)$ . What is her mistake?

(b) After correcting her mistake in part (a), she states, "I am 95% confident that the sample mean falls between 8.4 and 8.8." What is wrong with this statement?

(c) She quickly realizes her mistake in part (b) and instead states, "The probability that the true mean is between 8.4 and 8.8 is 0.95." What misinterpretation is she making now?

(d) Finally, in her defense for using the Normal distribution to determine the confidence interval she says, "Because the sample size is quite large, the population of alumni ratings will be approximately Normal." Explain to Ima her misunderstanding and correct this statement.

### 6.16 More confidence interval mistakes and misunderstandings.

Suppose that 100 randomly selected members of the Karaoke Channel were asked how much time they typically spend on the site during the week.<sup>7</sup> The sample mean  $\bar{x}$  was found to be 3.8 hours. Assume that the population standard deviation is known to be  $\sigma=2.9$

(a) Cary Oakey computes the 95% confidence interval for the average time on the site as  $3.8 \pm 1.96(2.9/100)$ . What is his mistake?

(b) He corrects this mistake and then states that "95% of the members spend between 3.23 and 4.37 hours a week on the site." What is wrong with his interpretation of this interval?

(c) The margin of error is slightly larger than half an hour. To reduce this to roughly 15 minutes, Cary says that the sample size needs to be doubled to 200. What is wrong with this statement?

## 6.17 The state of stress in the United States.

Since 2007, the American Psychological Association has supported an annual nationwide survey to examine stress across the United States.<sup>8</sup> A total of 340 Millennials (18- to 33-year-olds) were asked to indicate their average stress level (on a 10-point scale) during the past month. The mean score was 5.4. Assume that the population standard deviation is 2.3.

- Give the margin of error and find the 95% confidence interval for this sample.
- Repeat these calculations for a 99% confidence interval. How do the results compare with those in part (a)?

## 6.18 Inference based on integer values.

Refer to Exercise 6.17. The data for this study are integer values between 1 and 10. Explain why the confidence interval based on the Normal distribution should be a good approximation.

## 6.19 Mean TRAP in young women.

For many important processes that occur in the body, direct measurement of characteristics of the process is not possible. In many cases, however, we can measure a *biomarker*, a biochemical substance that is relatively easy to measure and is associated with the process of interest. Bone turnover is the net effect of two processes: the breaking down of old bone, called resorption, and the building of new bone, called formation. One biochemical measure of bone resorption is tartrate-resistant acid phosphatase (TRAP), which can be measured in blood. In a study of bone turnover in young women, serum TRAP was measured in 31 subjects.<sup>9</sup> The mean was 13.2 units per liter (U/l). Assume that the standard deviation is known to be 6.5 U/l. Give the margin of error and find a 95% confidence interval for the mean TRAP amount in young women represented by this sample.

## 6.20 Mean OC in young women.

Refer to the previous exercise. A biomarker for bone formation measured in the same study was osteocalcin (OC), measured in the blood. For the 31 subjects in the study, the mean was 33.4 nanograms per milliliter (ng/ml). Assume that the standard deviation is known to be 19.6 ng/ml. Report the 95% confidence interval.

## 6.21 Populations sampled and margins of error.

Consider the following two scenarios. (A) Take a simple random sample of 100 sophomore students at your college or university. (B) Take a simple random sample of 100 students at your college or university. For each of these samples you will record the amount spent on textbooks used for classes during the fall semester. Which sample should have the smaller margin of error? Explain your answer.



## 6.22 Average starting salary.

The National Association of Colleges and Employers (NACE) Fall Salary Survey shows that the current class of college graduates received an average starting-salary offer of \$44,259.<sup>10</sup> Your institution collected an SRS ( $n=400$ ) of its recent graduates and obtained a 95% confidence interval of (\$44,793, \$47,157). What can we conclude about the *difference* between the average starting salary of recent graduates at your institution and the overall NACE average? Write a short summary.

## 6.23 Consumption of sugar-sweetened beverages.

A recent study estimated that the U.S. per capita consumption of sugar-sweetened beverages among adults aged 20 to 34 years is 338 kilocalories per day (kcal/d).<sup>11</sup> Suppose that the population distribution is heavily skewed, with a standard deviation equal to 300 kcal/d. If you plan to take an SRS of 1000 young adults,

- (a) the 68–95–99.7 rule says that the probability is about 0.95 that  $\bar{x}$  is within \_\_\_\_\_ kcal/d of the population mean  $\mu$ . (Fill in the blank.)
- (b) about 95% of all samples will capture the true mean of kilocalories consumed per day in the interval  $\bar{x}$  plus or minus \_\_\_\_\_ kcal/d. (Fill in the blank.)

## 6.24 Apartment rental rates.

You want to rent an unfurnished two-bedroom apartment in Dallas next year. The mean monthly rent for a random sample of 10 apartments advertised in the local newspaper is \$1050. Assume that the monthly rents in Dallas follow a Normal distribution with a standard deviation of \$220. Find a 95% confidence interval for the mean monthly rent for unfurnished two-bedroom apartments available in Dallas.

## 6.25 More on apartment rental rates.

Refer to the previous exercise. Will the 95% confidence interval include approximately 95% of the rents for all unfurnished two-bedroom apartments in this area? Explain why or why not.



## 6.26 Inference based on skewed data.

The mean OC for the 31 subjects in Exercise 6.20 was 33.4 ng/ml. In our calculations, we assumed that the standard deviation was known to be 19.6 ng/ml. Use the 68–95–99.7 rule from Chapter 1 (page xx) to find the approximate bounds on the values of OC that would include these percents of the population. If the assumed standard deviation is correct, this distribution may be highly skewed. Why? (*Hint:* The measured values for a variable such as this are all positive.) Do you think that this skewness will invalidate the use of the Normal confidence interval in this case? Explain your answer.

## 6.27 Average hours per week listening to the radio.

The *Student Monitor* surveys 1200 undergraduates from four-year colleges and universities throughout the United States semiannually to understand trends among college students.<sup>12</sup> Recently, the *Student Monitor* reported that the average amount of time listening to the radio per week was 11.5 hours. Of the 1200 students surveyed, 83% said that they listened to the radio, so this collection of listening times has around 204 (17% $\times$ 1200) zeros. Assume that the standard deviation is 8.3 hours.

- (a) Give a 95% confidence interval for the mean time spent per week listening to the radio.
- (b) Is it true that 95% of the 1200 students reported weekly times that lie in the interval you found in part (a)? Explain your answer.
- (c) It appears that the population distribution has many zeros and is skewed to the right. Explain why the confidence interval based on the Normal distribution should nevertheless be a good approximation.

## 6.28 Average minutes per week listening to the radio.

Refer to the previous exercise.

- (a) Give the mean and standard deviation in minutes.
- (b) Calculate the 95% confidence interval in minutes from your answer to part (a).
- (c) Explain how you could have directly calculated this interval from the 95% interval that you calculated in the previous exercise.

## 6.29 Satisfied with your job?

Job satisfaction is one of four workplace measures that the Gallup-Healthways Well-Being Index tracks among U.S. workers. The question asked is “Are you satisfied or dissatisfied with your job or the work that you do?” In 2011, 87.5% responded that they were satisfied. Material provided with the results of the poll noted:

*Results are based on telephone interviews conducted as part of the Gallup-Healthways Well-Being Index survey Jan. 1–April 30, 2011, with a random sample of 61, 889 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia, selected using random-digit-dial sampling.*

*For results based on the total sample of national adults, one can say with 95% confidence that the maximum margin of sampling error is 1 percentage point.<sup>13</sup>*

The poll uses a complex multistage sample design, but the sample percent has approximately a Normal sampling distribution.

- (a) The announced poll result was  $87.5\% \pm 1\%$ . Can we be certain that the true population percent falls in this interval? Explain your answer.
- (b) Explain to someone who knows no statistics what the announced result  $87.5\% \pm 1\%$  means.
- (c) This confidence interval has the same form we have met earlier:

$$\text{estimate} \pm z^* \sigma_{\text{estimate}}$$

What is the standard deviation  $\sigma_{\text{estimate}}$  of the estimated percent?

- (d) Does the announced margin of error include errors due to practical problems such as nonresponse? Explain your answer.

## 6.30 Fuel efficiency.

Computers in some vehicles calculate various quantities related to performance. One of these is the fuel efficiency, or gas mileage, usually expressed as miles per gallon (mpg). For one vehicle equipped in this way, the miles per gallon were recorded each time the gas tank was filled, and the computer was then reset.<sup>14</sup> Here are the mpg values for a random sample of 20 of these records:  **MPG**

41.5	50.7	36.6	37.3	34.2	45.0	48.0	43.2	47.7	42.2
43.2	44.6	48.4	46.4	46.8	39.2	37.3	43.5	44.3	43.3

Suppose that the standard deviation is known to be  $\sigma = 3.5$  mpg.

- (a) What is  $\sigma_{\bar{x}}$  the standard deviation of  $\bar{x}$ ?
- (b) Examine the data for skewness and other signs of non-Normality. Show your plots and numerical summaries. Do you think it is reasonable to construct a confidence interval based on the Normal distribution? Explain your answer.

(c) Give a 95% confidence interval for  $\mu$ , the mean miles per gallon for this vehicle.

### 6.31 Fuel efficiency in metric units.

In the previous exercise you found an estimate with a margin of error for the average miles per gallon. Convert your estimate and margin of error to the metric units kilometers per liter (kpl). To change mpg to kpl, use the fact that 1 mile = 1.609 kilometers and 1 gallon = 3.785 liters.

### 6.32 How many “hits”?

The *Confidence Interval* applet lets you simulate large numbers of confidence intervals quickly. Select 95% confidence and then sample 50 intervals. Record the number of intervals that cover the true value (this appears in the “Hit” box in the applet). Press the “Reset” button and repeat 30 times. Make a stemplot of the results and find the mean. Describe the results. If you repeated this experiment very many times, what would you expect the average number of hits to be?

### 6.33 Required sample size for specified margin of error.

A new bone study is being planned that will measure the biomarker TRAP described in Exercise 6.19. Using the value of  $\sigma$  given there, 6.5 U/l, find the sample size required to provide an estimate of the mean TRAP with a margin of error of 1.5 U/l for 95% confidence.

### 6.34 Adjusting required sample size for dropouts.

Refer to the previous exercise. In similar previous studies, about 20% of the subjects drop out before the study is completed. Adjust your sample size requirement so that you will have enough subjects at the end of the study to meet the margin of error criterion.

### 6.35 Radio poll.

A college radio station invites listeners to enter a dispute about a proposed “pay as you throw” waste collection program. The station asks listeners to call in and state how much each 10 gallons of trash should cost. A total of 617 listeners call in. The station calculates the 95% confidence interval for the average fee desired by city residents to be \$1.03 to \$1.39. Is this result trustworthy? Explain your answer.

### 6.36 Accuracy of a laboratory scale.

To assess the accuracy of a laboratory scale, a standard weight known to weigh 10 grams is weighed repeatedly. The scale readings are Normally distributed with unknown mean (this mean is 10 grams if the scale has no bias). The standard deviation of the scale readings is known to be 0.0002 gram.

- The weight is measured five times. The mean result is 10.0023 grams. Give a 98% confidence interval for the mean of repeated measurements of the weight.
- How many measurements must be averaged to get a margin of error of  $\pm 0.0001$  with 98% confidence?

### 6.37 More than one confidence interval.

As we prepare to take a sample and compute a 95% confidence interval, we know that the probability that

the interval we compute will cover the parameter is 0.95. That's the meaning of 95% confidence. If we plan to use several such intervals, however, our confidence that *all* of them will give correct results is less than 95%. Suppose that we plan to take independent samples each month for five months and report a 95% confidence interval for each set of data.

(a) What is the probability that all five intervals will cover the true means? This probability (expressed as a percent) is our overall confidence level for the five simultaneous statements.

(b) What is the probability that at least four of the five intervals will cover the true means?

## 6.2 Tests of Significance

When you complete this section, you will be able to

- Outline the four steps common to all tests of significance.
- Formulate the null and alternative hypotheses of a significance test.
- Describe a common form for the test statistic in terms of the parameter estimate, its standard deviation, and the hypothesized value.
- Define what a  $P$ -value is and explain whether a small  $P$ -value provides evidence for or against the null hypothesis.
- Draw a conclusion from a test of significance based on the test's  $P$ -value and significance level  $\alpha$ .
- Describe the relationship between a level  $\alpha$  two-sided significance test for  $\mu$  and the  $1 - \alpha$  confidence interval.

The confidence interval is appropriate when our goal is to estimate population parameters. The second common type of inference is directed at a quite different goal: to assess the evidence provided by the data in favor of some claim about the population parameters.

### The reasoning of significance tests

A significance test is a formal procedure for comparing observed data with a hypothesis whose truth we want to assess. The hypothesis is a statement about the population parameters. The results of a test are expressed in terms of a probability that measures how well the data and the hypothesis agree. We use the following examples to illustrate these concepts.

#### Example

##### 6.8 Credit card debt by grade level

One purpose of Sallie Mae's annual study described in Example 6.4 (page 360) is to allow comparisons of different subgroups of undergraduates. For

example, the average outstanding credit card balance of freshmen is \$642, while the average outstanding balance for seniors is \$516. The difference of \$126 is fairly large, but we know that these numbers are estimates of the population means. If we took different samples, we would get different estimates.

Can we conclude from these data that the average outstanding balances of undergraduates in these two grade levels are different? One way to answer this question is to compute the probability of obtaining a difference as large or larger than the observed \$126 assuming that, in fact, there is no difference in the population means. This probability is 0.18. Because this probability is not particularly small, we conclude that observing a difference of \$126 is not very surprising when the population means are equal. The data do not provide enough evidence for us to conclude that the average outstanding credit card balances for freshmen and seniors differ.

Here is an example with a different conclusion.

## Example

### 6.9 Credit card debt by U.S. region

Sallie Mae's study also reports that the average outstanding balance among undergraduates in the South is \$771, while it is \$478 among undergraduates in the Midwest. Is the average balance among undergraduates in the South higher than the average balance among undergraduates in the Midwest? The observed difference is \$293, but as we learned in the previous example, an observed difference in means is not necessarily sufficient for us to conclude that the population means are different.

Again, we answer this question with a probability calculated under the assumption that there is *no difference in the population means*. The probability is 0.0002 of observing a difference in mean debt that is \$293 or more when there really is no difference. Because this probability is so small, we have sufficient evidence in the data to conclude that the average outstanding balance among undergraduates in the South is higher than the average balance among undergraduates in the Midwest.

What are the key steps in these examples?

- We started each with a question about the difference between two means. In Example 6.8, we compare freshmen with seniors. In Example 6.9, we compare

undergraduates in the South and Midwest. In both cases, we ask whether or not the data are compatible with “no difference,” that is, a difference of \$0.

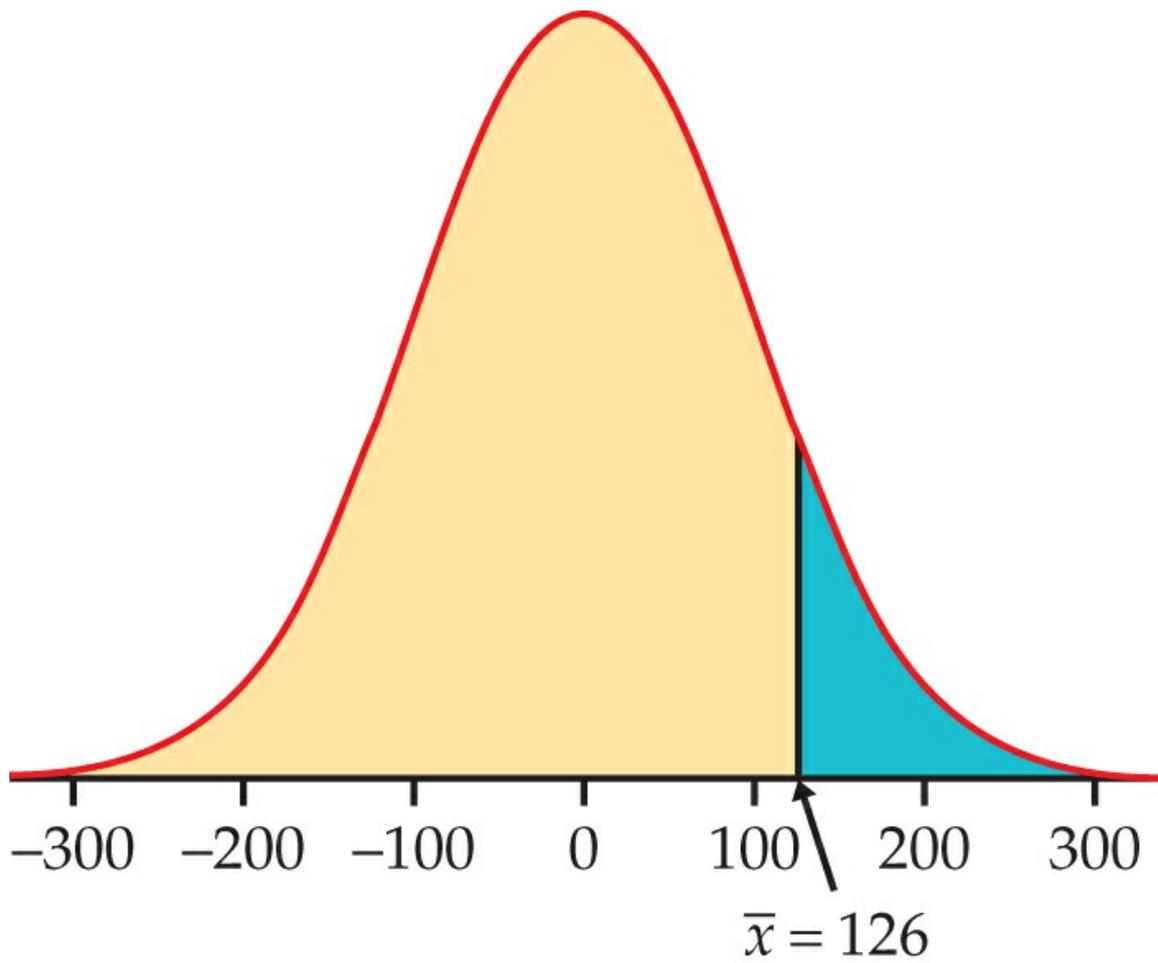
- Next we compared the difference given by the data, \$126 in the first case and \$293 in the second, with the value assumed in the question, \$0.
- The results of the comparisons are probabilities, 0.18 in the first case and 0.0002 in the second.

The 0.18 probability is not particularly small, so we have limited evidence to question the possibility that the true difference is zero. In the second case, however, the probability is very small. Something that happens with probability 0.0002 occurs only about 2 times out of 10,000. In this case we have two possible explanations:

1. We have observed something that is very unusual, or
2. The assumption that underlies the calculation, no difference in mean balance, is not true.

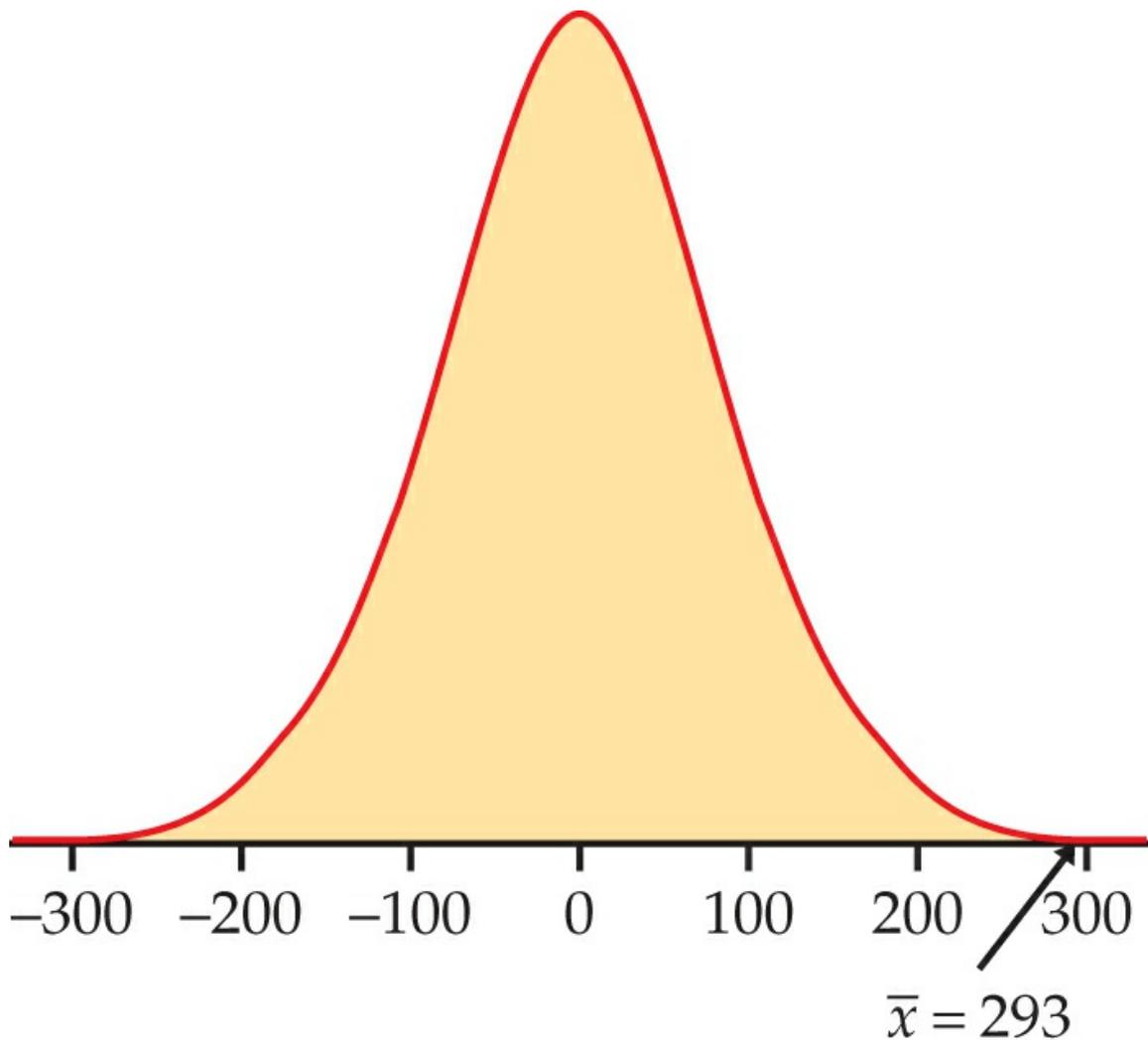
Because this probability is so small, we prefer the second conclusion: the average outstanding credit card balances for undergraduates in the South and for undergraduates in the Midwest are different, with the South balance higher than that of the Midwest.

The probabilities in Examples 6.8 and 6.9 are measures of the compatibility of the data (a difference in means of \$126 and \$293) with the *null hypothesis* that there is no difference in the population means. Figures 6.7 and 6.8 compare the two results graphically. For each a Normal curve centered at 0 is the sampling distribution. You can see from Figure 6.7 that we should not be particularly surprised to observe the difference \$126, but the difference \$293 in Figure 6.8 is clearly an unusual observation. We will now consider some of the formal aspects of significance testing.



**Figure 6.7**

Comparison of the sample mean in Example 6.8 with the null hypothesized value 0.



**Figure 6.8**

Comparison of the sample mean in Example 6.9 with the null hypothesized value 0.

Generic curve with area to the left and right of absolute value of  $z$  shaded."Generic curve with area to the left of  $z$  shaded."Generic curve with area to the right of  $z$  shaded."Generic curve with area to the left and right of absolute value of  $z$  shaded."Generic curve with area to the left of  $z$  shaded."Generic curve with area to the right of  $z$  shaded."

### Stating hypotheses

In Example 6.8 and 6.9, we asked whether the difference in the observed means is reasonable if, in fact, there is no difference in the population means. To answer this, we begin by supposing that the statement following the “if” in the previous sentence is true. In other words, we suppose that the true difference is \$0. We then ask whether the data provide evidence against the supposition we have made. If so, we have evidence in favor of an effect (the means are different) we are seeking. Often, the first step in a test of significance is to state a claim that we will try to

find evidence *against*.

## NULL HYPOTHESIS

The statement being tested in a test of significance is called the **null hypothesis**. The test of significance is designed to assess the strength of the evidence against the null hypothesis. Usually the null hypothesis is a statement of “no effect” or “no difference.”

We abbreviate “null hypothesis” as  $H_0$ . A null hypothesis is a statement about the population parameters. For example, our null hypothesis for Example 6.8 is

$H_0$ :there is no difference in the population means

Note that the null hypothesis refers to the *population* means for all undergraduates, including those for whom we do not have data.

### alternative hypothesis

It is convenient also to give a name to the statement we hope or suspect is true instead of  $H_0$ . This is called the **alternative hypothesis** and is abbreviated as  $H_a$ . In Example 6.8, the alternative hypothesis states that the means are different. We write this as

$H_a$ :the population means are not the same

*Hypotheses always refer to some populations or a model, not to a particular outcome. For this reason, we must state  $H_0$  and  $H_a$  in terms of population parameters.*



Because  $H_a$  expresses the effect that we hope to find evidence *for*, we will sometimes begin with  $H_a$  and then set up  $H_0$  as the statement that the hoped-for effect is not present. Stating  $H_a$ , however, is often the more difficult task. It is not always clear, in particular, whether  $H_a$  should be **one-sided** or **two-sided**, which refers to whether a parameter differs from its null hypothesis value in a specific direction or in either direction.

### one-sided or two-sided alternatives

The alternative hypothesis should express the hopes or suspicions we bring to

the data. *It is cheating to first look at the data and then frame  $H_a$  to fit what the data show.* If you do not have a specific direction firmly in mind in advance, you must use a two-sided alternative. Moreover, some users of statistics argue that we should always use a two-sided alternative.



## USE YOUR KNOWLEDGE

### 6.38 Food court survey

The food court closest to your dormitory has been redesigned. A survey is planned to assess whether or not students think that the new design is an improvement. It will contain 8 questions; a seven-point scale will be used for the answers, with scores less than 4 favoring the previous food court and scores greater than 4 favoring the new design (to varying degrees). The average of these 8 questions will be used as the student's opinion. State the null and alternative hypotheses you would use for examining whether or not the new design is viewed as an improvement.

### 6.39 DXA scanners

A dual-energy X-ray absorptiometry (DXA) scanner is used to measure bone mineral density for people who may be at risk for osteoporosis. One company believes that its scanner is not giving accurate readings. To assess this, the company uses an object called a "phantom" that has known mineral density  $\mu=1.4$  grams per square centimeter. The company scans the phantom 10 times and compares the sample mean reading  $\bar{x}$  with the theoretical mean  $\mu$  using a significance test. State the null and alternative hypotheses for this test.

## Test statistics

We will learn the form of significance tests in a number of common situations. Here are some principles that apply to most tests and that help in understanding these tests:

- The test is based on a statistic that estimates the parameter that appears in the

hypotheses. Usually this is the same estimate we would use in a confidence interval for the parameter. When  $H_0$  is true, we expect the estimate to take a value near the parameter value specified by  $H_0$ . We call this specified value the hypothesized value.

- Values of the estimate far from the hypothesized value give evidence against  $H_0$ . The alternative hypothesis determines which directions count against  $H_0$ .
- To assess how far the estimate is from the hypothesized value, standardize the estimate. In many common situations the test statistic has the form

$$z = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard deviation of the estimate}}$$

A **test statistic** measures compatibility between the null hypothesis and the data. We use it for the probability calculation that we need for our test of significance. It is a random variable with a distribution that we know.

### test statistic

Let's return to our comparison of credit card balances among freshmen and seniors and specify the hypotheses as well as calculate the test statistic.

## Example

### 6.10 Average credit card balances of freshmen and seniors: the hypotheses



In Example 6.8, the hypotheses are stated in terms of the difference in average outstanding credit card balance between freshmen and seniors:

$H_0$ : there is no difference in the population means

$H_a$ : there is a difference in the population means

Because  $H_a$  is two-sided, large values of both positive and negative differences count as evidence against the null hypothesis.

We can also state the null hypothesis as  $H_0$ : the true mean difference is 0. This statement makes it more clear that hypothesized value for this comparison of credit card balances is 0.

## Example

### 6.11 Average credit card balances of freshmen and seniors: the test statistic

In Example 6.8, the estimate of the difference is \$126. Using methods that we will discuss in detail later, we can determine that the standard deviation of the estimate is \$95. For this problem the test statistic is

$$z = \frac{126 - 0}{95} = 1.33$$

For our data,

$$z=126-095=1.33$$

We have observed a sample estimate that is about one and one-third standard deviations away from the hypothesized value of the parameter.

Because the sample sizes are sufficiently large for us to conclude that the distribution of the sample estimate is approximately Normal, the standardized test statistic  $z$  will have approximately the  $N(0,1)$  distribution. We will use facts about the Normal distribution in what follows.



**Normal distribution, p. 58**

## ***P*-values**

If all test statistics were Normal, we could base our conclusions on the value of the  $z$  test statistic. In fact, the Supreme Court of the United States has said that “two or three standard deviations” ( $z=2$  or  $3$ ) is its criterion for rejecting  $H_0$  (see Exercise 6.44 on page 381), and this is the criterion used in most applications involving the law. But because not all test statistics are Normal, we use the language of probability to express the meaning of a test statistic.

A test of significance finds the probability of getting an outcome *as extreme or more extreme than the actually observed outcome*. “Extreme” means “far from what we would expect if  $H_0$  were true.” The direction or directions that count as “far from what we would expect” are determined by  $H_a$  and  $H_0$ .

## ***P*-VALUE**

The probability, assuming  $H_0$  is true, that the test statistic would take a value as extreme or more extreme than that actually observed is called the ***P*-value** of the test. The smaller the *P*-value, the stronger the evidence against  $H_0$  provided by the data.

The key to calculating the *P*-value is the sampling distribution of the test statistic. For the problems we consider in this chapter, we need only the standard Normal distribution for the test statistic  $z$ .

In Example 6.8 we want to know if the average outstanding credit card balance for freshmen differs from the average balance for seniors. The difference we calculated based on our sample is \$126, which corresponds to 1.33 standard deviations away from zero—that is,  $z=1.33$ . Because we are using a two-sided

alternative for this problem, the evidence against  $H_0$  is measured by the probability that we observe a value of  $Z$  as extreme or more extreme than 1.33 in either direction.

## Example

### 6.12 Average credit card balances of freshmen and seniors: the $P$ -value

In Example 6.11 we found that the test statistic for testing

$H_0$ :the true mean difference is 0

versus

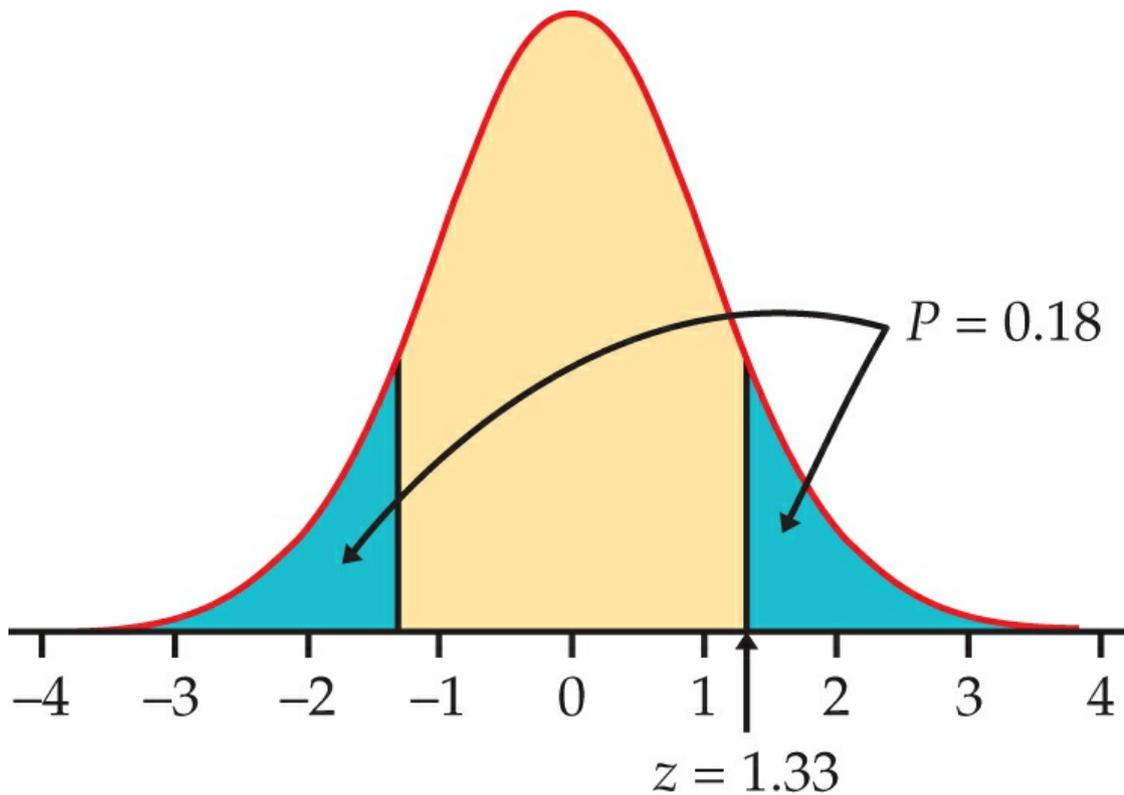
$H_a$ :there is a difference in the population means

is

$$z = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard deviation of the estimate}}$$

If  $H_0$  is true, then  $z$  is a single observation from the standard Normal,  $N(0,1)$  distribution. Figure 6.9 illustrates this calculation. The  $P$ -value is the probability of observing a value of  $Z$  at least as extreme as the one that we observed,  $z=1.33$ . From Table A, our table of standard Normal probabilities, we find

$$P(Z \geq 1.33) = 1 - 0.9082 = 0.0918$$



**Figure 6.9**

The  $P$ -value for Example 6.12. The  $P$ -value is the probability (when  $H_0$  is true) that  $\bar{x}$  takes a value as extreme or more extreme than the actual observed value,  $z = 1.33$ . Because the alternative hypothesis is two-sided, we use both tails of the distribution.

The probability for being extreme in the negative direction is the same:

$$P(Z \leq -1.33) = 0.0918$$

So the  $P$ -value is

$$P = 2P(Z \geq 1.33) = 2(0.0918) = 0.1836$$

This is the value that we reported on page 372. There is an 18% chance of observing a difference as extreme as the \$126 in our sample if the true population difference is zero. This  $P$ -value tells us that our outcome is not particularly extreme. In other words, the data do not provide substantial evidence for us to doubt the validity of the null hypothesis.

## USE YOUR KNOWLEDGE

### 6.40 Normal curve and the $P$ -value

A test statistic for a two-sided significance test for a population mean is

$z=2.31$ . Sketch a standard Normal curve and mark this value of  $z$  on it. Find the  $P$ -value and shade the appropriate areas under the curve to illustrate your calculations.

#### 6.41 More on the Normal curve and the $P$ -value

A test statistic for a two-sided significance test for a population mean is  $z=-1.81$ . Sketch a standard Normal curve and mark this value of  $z$  on it. Find the  $P$ -value and shade the appropriate areas under the curve to illustrate your calculations.

### Statistical significance

We started our discussion of the reasoning of significance tests with the statement of null and alternative hypotheses. We then learned that a test statistic is the tool used to examine the compatibility of the observed data with the null hypothesis. Finally, we translated the test statistic into a  $P$ -value to quantify the evidence against  $H_0$ . One important final step is needed: to state our conclusion.

We can compare the  $P$ -value we calculated with a fixed value that we regard as decisive. This amounts to announcing in advance how much evidence against  $H_0$  we will require to reject  $H_0$ . The decisive value is called the **significance level**. It is commonly denoted by  $\alpha$  (the Greek letter alpha). If we choose  $\alpha=0.05$ , we are requiring that the data give evidence against  $H_0$  so strong that it would happen no more than 5% of the time (1 time in 20) when  $H_0$  is true. If we choose  $\alpha=0.01$ , we are insisting on stronger evidence against  $H_0$ , evidence so strong that it would appear only 1% of the time (1 time in 100) if  $H_0$  is in fact true.

#### significance level

### STATISTICAL SIGNIFICANCE

If the  $P$ -value is as small or smaller than  $\alpha$  we say that the data are **statistically significant at level  $\alpha$** .

*“Significant” in the statistical sense does not mean “important.”* The original meaning of the word is “signifying something.” In statistics the term is used to indicate only that the evidence against the null hypothesis has reached the standard set by  $\alpha$ . For example, significance at level 0.01 is often expressed by the statement “The results were significant ( $P<0.01$ ).” Here  $P$  stands for the  $P$ -value. The  $P$ -value is more informative than a statement of significance because we can then

assess significance at any level we choose. For example, a result with  $P=0.03$  is significant at the  $\alpha=0.05$  level but is not significant at the  $\alpha=0.01$  level. We discuss this in more detail at the end of this section.



## Example

### 6.13 Average outstanding credit card balances of freshmen and seniors: the conclusion

In Example 6.12 we found that the  $P$ -value is

$$P=2P(Z\geq 1.33)=2(0.0918)=0.1836$$

There is an 18% chance of observing a difference as extreme as the \$126 in our sample if the true population difference is zero. Because this  $P$ -value is larger than the  $\alpha=0.05$  significance level, we conclude that our test result is not significant. We could report the result as “the data fail to provide evidence that would cause us to conclude that there is a difference in average outstanding balances between freshmen and seniors ( $z=1.33$ ,  $P=0.18$ ).”

This statement does not mean that we conclude that the null hypothesis is true, only that the level of evidence we require to reject the null hypothesis is not met. Our criminal court system follows a similar procedure in which a defendant is presumed innocent ( $H_0$ ) until proven guilty. If the level of evidence presented is not strong enough for the jury to find the defendant guilty beyond a reasonable doubt, the defendant is acquitted. Acquittal does not imply innocence, only that the degree of evidence was not strong enough to prove guilt.

If the  $P$ -value is small, we reject the null hypothesis. Here is the conclusion for our second example.

## Example

## 6.14 Credit card debt by U.S. region: the conclusion

In Example 6.9 we found that the difference in debt between undergraduates in the South and in the Midwest was \$293. Since the cost of living is higher in the South than in the Midwest,<sup>15</sup> we had a prior expectation that the outstanding balance would be higher for undergraduates in the South. It is appropriate to use a one-sided alternative in this situation. So, our hypotheses are

$H_0$ : the true mean difference is 0

versus

$H_a$ : the difference between the average outstanding balance of undergraduates in the South and the Midwest is positive

The standard deviation is \$82.5 (again, we defer details regarding this calculation), and the test statistic is

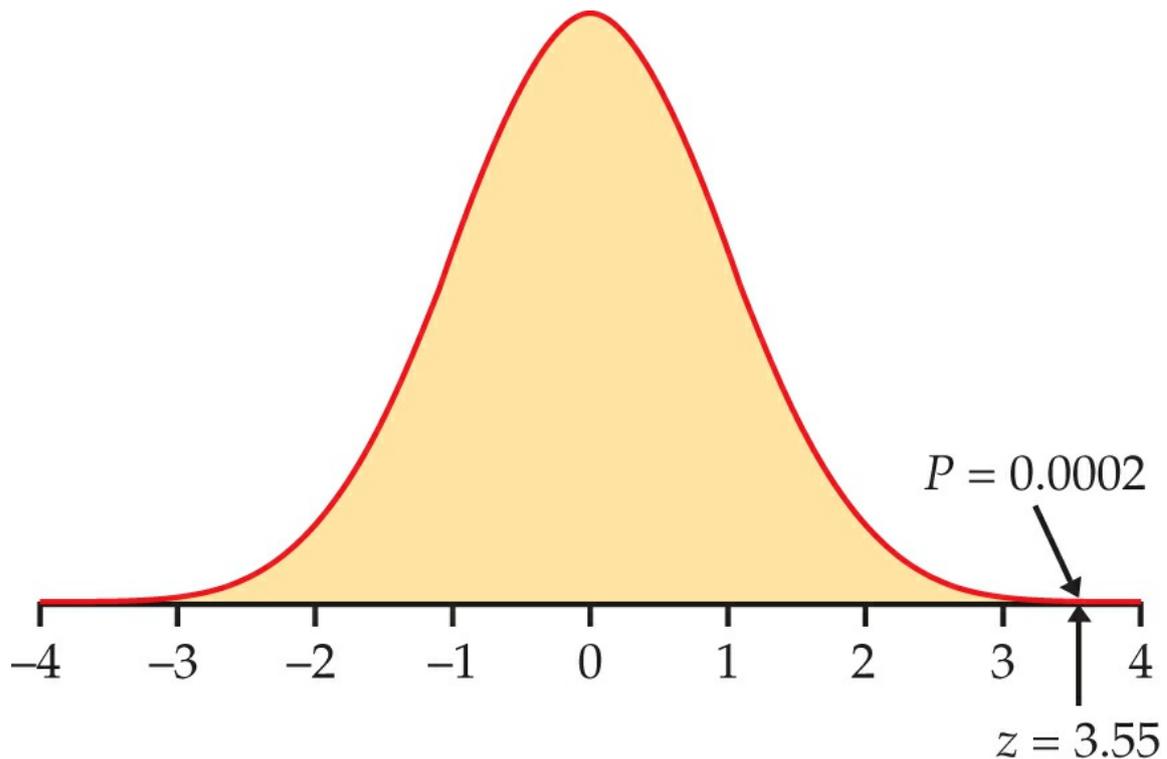
$$z = \frac{293 - 0}{82.5}$$

$$\begin{aligned} z &= \frac{\text{estimate} - \text{hypothesized value}}{\text{standard deviation of the estimate}} \\ &= 3.55 \end{aligned}$$

Because only positive differences in credit card debt count against the null hypothesis, the one-sided alternative leads to the calculation of the  $P$ -value using the upper tail of the Normal distribution. The  $P$ -value is

$$\begin{aligned} P &= P(Z \geq 3.55) \\ &= 0.0002 \end{aligned}$$

The calculation is illustrated in Figure 6.10. There is about a 2-in-10,000 chance of observing a difference as large or larger than the \$293 in our sample if the true population difference is zero. This  $P$ -value tells us that our outcome is extremely rare. We conclude that the null hypothesis must be false. Since the observed difference is positive, here is one way to report the result: “The data clearly show that the mean credit card debt for undergraduates in the South is larger than the mean credit card debt for undergraduates in the Midwest ( $z=3.55$ ,  $P<0.001$ ).”



**Figure 6.10**

The  $P$ -value for Example 6.14. The  $P$ -value is the probability (when  $H_0$  is true) that  $\bar{x}$  takes a value as extreme or more extreme than the actual observed value,  $z = 3.55$ . We look at only the right tail because we are considering the one-sided ( $>$ ) alternative.

Note that the calculated  $P$ -value for this example is 0.0002 but we reported the result as  $P < 0.001$ . The value 0.001, 1 in 1000, is sufficiently small to force a clear rejection of  $H_0$ . Standard practice is to report very small  $P$ -values as simply less than 0.001.

## USE YOUR KNOWLEDGE

### 6.42 Finding significant $z$ -scores

Consider a two-sided significance test for a population mean.

- Sketch a Normal curve similar to that shown in Figure 6.9 (page 378), but find the value  $z$  such that  $P = 0.05$ .
- Based on your curve from part (a), what values of the  $z$  statistic are statistically significant at the  $\alpha = 0.05$  level?

### 6.43 More on finding significant z-scores

Consider a one-sided significance test for a population mean, where the alternative is “greater than.”

(a) Sketch a Normal curve similar to that shown in Figure 6.10 but find the value  $z$  such that  $P=0.05$ .

(b) Based on your curve from part (a), what values of the  $z$  statistic are statistically significant at the  $\alpha=0.05$  level?

### 6.44 The Supreme Court speaks

The Supreme Court has said that  $z$ -scores beyond 2 or 3 are generally convincing statistical evidence. For a two-sided test, what significance level corresponds to  $z=2$ ? To  $z=3$ ?

A test of significance is a process for assessing the significance of the evidence provided by data against a null hypothesis. The four steps common to all tests of significance are as follows:

1. State the *null hypothesis*  $H_0$  and the *alternative hypothesis*  $H_a$ . The test is designed to assess the strength of the evidence against  $H_0$ ;  $H_a$  is the statement that we will accept if the evidence enables us to reject  $H_0$ .
2. Calculate the value of the *test statistic* on which the test will be based. This statistic usually measures how far the data are from  $H_0$ .
3. Find the *P-value* for the observed data. This is the probability, calculated assuming that  $H_0$  is true, that the test statistic will weigh against  $H_0$  at least as strongly as it does for these data.
4. State a conclusion. One way to do this is to choose a *significance level*  $\alpha$ , how much evidence against  $H_0$  you regard as decisive. If the *P-value* is less than or equal to  $\alpha$ , you conclude that the alternative hypothesis is true; if it is greater than  $\alpha$ , you conclude that the data do not provide sufficient evidence to reject the null hypothesis. Your conclusion is a sentence or two that summarizes what you have found by using a test of significance.

We will learn the details of many tests of significance in the following chapters. The proper test statistic is determined by the hypotheses and the data collection design. We use computer software or a calculator to find its numerical value and the *P-value*. The computer will not formulate your hypotheses for you, however. Nor will it decide if significance testing is appropriate or help you to interpret the

*P*-value that it presents to you. These steps require judgment based on a sound understanding of this type of inference.

## Tests for a population mean

Our discussion has focused on the reasoning of statistical tests, and we have outlined the key ideas for one type of procedure. Our examples focused on the comparison of two population means. Here is a summary for a test about one population mean.

We want to test the hypothesis that a parameter has a specified value. This is the null hypothesis. For a test of a population mean  $\mu$ , the null hypothesis is

$$H_0: \text{the true population mean is equal to } \mu_0$$

which often is expressed as

$$H_0: \mu = \mu_0$$

where  $\mu_0$  is the hypothesized value of  $\mu$  that we would like to examine.

The test is based on data summarized as an estimate of the parameter. For a population mean this is the sample mean  $\bar{x}$ . Our test statistic measures the difference between the sample estimate and the hypothesized parameter in terms of standard deviations of the test statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Recall from Chapter 5 that the standard deviation of  $\bar{x}$  is  $\sigma/\sqrt{n}$ . Therefore, the test statistic is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$



**distribution of sample mean, p. 307**

Again recall from Chapter 5 that, if the population is Normal, then  $\bar{x}$  will be Normal and  $z$  will have the standard Normal distribution when  $H_0$  is true. By the central limit theorem, both distributions will be approximately Normal when the sample size is large even if the population is not Normal. We'll assume that we're in one of these two settings for now.



**central limit theorem, p. 307**

Suppose that we have calculated a test statistic  $z=1.7$ . If the alternative is one-sided on the high side, then the *P*-value is the probability that a standard Normal random variable  $Z$  takes a value as large or larger than the observed 1.7. That is,

$$P = P(Z \geq 1.7)$$

$$\begin{aligned}
&=1-P(Z<1.7) \\
&=1-0.9554 \\
&=0.0446
\end{aligned}$$

Similar reasoning applies when the alternative hypothesis states that the true  $\mu$  lies below the hypothesized  $\mu_0$  (one-sided). When  $H_a$  states that  $\mu$  is simply unequal to  $\mu_0$  (two-sided), values of  $z$  away from zero in either direction count against the null hypothesis. The  $P$ -value is the probability that a standard Normal  $Z$  is at least as far from zero as the observed  $z$ . Again, if the test statistic is  $z=1.7$  the two-sided  $P$ -value is the probability that  $Z\leq-1.7$  or  $Z\geq 1.7$ . Because the standard Normal distribution is symmetric, we calculate this probability by finding  $P(Z\geq 1.7)$  and *doubling* it:

$$\begin{aligned}
P(Z\leq-1.7 \text{ or } Z\geq 1.7) &=2P(Z\geq 1.7) \\
&=2(1-0.9554)=0.0892
\end{aligned}$$

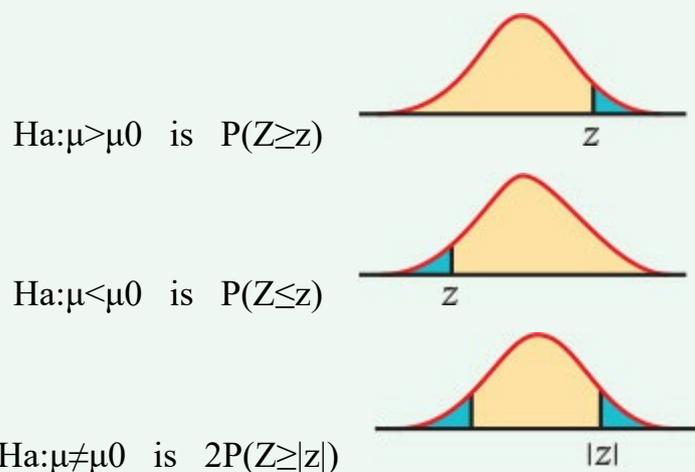
We would make exactly the same calculation if we observed  $z=-1.7$ . It is the absolute value  $|z|$  that matters, not whether  $z$  is positive or negative. Here is a statement of the test in general terms.

### **$z$ TEST FOR A POPULATION MEAN**

To test the hypothesis  $H_0:\mu=\mu_0$  based on an SRS of size  $n$  from a population with unknown mean  $\mu$  and known standard deviation  $\sigma$ , compute the **test statistic**

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{262 - 286}{155/\sqrt{100}}$$

In terms of a standard Normal random variable  $Z$ , the  $P$ -value for a test of  $H_0$  against



These  $P$ -values are exact if the population distribution is Normal and are approximately correct for large  $n$  in other cases.

---

## **Example**

### **6.15 Energy intake from sugar-sweetened beverages**



Consumption of sugar-sweetened beverages (SSBs) has been positively associated with weight gain and obesity and negatively associated with the intake of important micronutrients. One study used data from the National Health and Nutrition Examination Survey (NHANES) to estimate SSB

consumption among adolescents (aged 12 to 19 years). More than 7500 individuals provided data for this study.<sup>16</sup> The mean consumption was 286 calories per day.

You survey 100 students at your large university and find the average consumption of SSBs per day to be 262 calories. Is there evidence that the average calories per day from SSBs at your university differs from this large U.S. survey average?

The null hypothesis is “no difference” from the published mean  $\mu_0=286$ . The alternative is two-sided because you did not have a particular direction in mind before examining the data. So the hypotheses about the unknown mean  $\mu$  of the students at your university are

$$H_0:\mu=286$$

$$H_a:\mu\neq 286$$

As usual in this chapter, we make the unrealistic assumption that the population standard deviation is known. In this case we’ll assume that  $\sigma=155$  calories. The  $z$  test requires that the 100 students in the sample are an SRS from the population of students at your university. We will assume that the students in the sample were selected in a proper random manner. We’ll also assume that  $n = 100$  is sufficiently large that we can rely on the central limit theorem to assure us that the  $P$ -value based on the Normal distribution will be a good approximation.

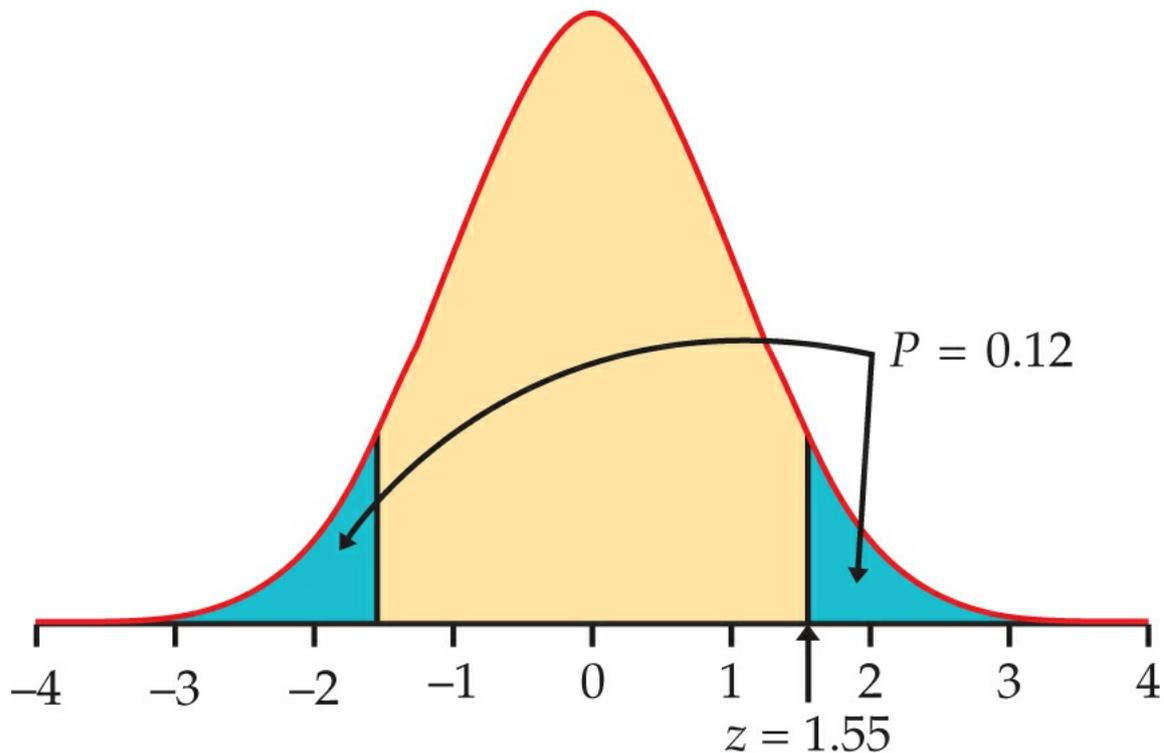
We compute the test statistic:

$$\begin{aligned} z &= \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{262 - 286}{155/\sqrt{100}} \\ &= -1.55 \end{aligned}$$

Figure 6.11 illustrates the  $P$ -value, which is the probability that a standard Normal variable  $Z$  takes a value at least 1.55 away from zero. From Table A we find that this probability is

$$P = 2P(Z \geq 1.55) = 2(1 - 0.9394) = 0.1212$$

That is, more than 12% of the time an SRS of size 100 from the students at your university would have a mean consumption from SSBs at least as far from 286 as that of this sample if the population mean were 286. The observed  $\bar{x}=262$  is therefore not strong evidence that the student population mean at your university differs from that of the large population of adolescents.



**Figure 6.11**

Sketch of the  $P$ -value calculation for the two-sided test in Example 6.15. The test statistic is  $z = -1.55$ .

The data in Example 6.15 do *not* establish that the mean consumption  $\mu$  for the students at your university is 286 calories. We sought evidence that  $\mu$  differed from 286 and failed to find convincing evidence. That is all we can say. No doubt the mean amount at your university is not exactly equal to 286 calories. A large enough sample would give evidence of the difference, even if it is very small.

Tests of significance assess the evidence *against*  $H_0$ . If the evidence is strong, we can confidently reject  $H_0$  in favor of the alternative. *Failing to find evidence against  $H_0$  means only that the data are consistent with  $H_0$ , not that we have clear evidence that  $H_0$  is true.*

## Example

### 6.16 Significance test of the mean SATM score

In a discussion of SAT Mathematics (SATM) scores, someone comments: “Because only a select minority of California high school students take the test, the scores overestimate the ability of typical high school seniors. I think

that if all seniors took the test, the mean score would be no more than 475.” You do not agree with this claim and decide to use the SRS of 500 seniors from Example 6.3 (page 354) to assess the degree of evidence against it. Those 500 seniors had a mean SATM score of  $\bar{x}=485$  Is this strong enough evidence to conclude that this person’s claim is wrong?

Because the claim states that the mean is “no more than 475,” the alternative hypothesis is one-sided. The hypotheses are

$$H_0:\mu=475$$

$$H_a:\mu>475$$

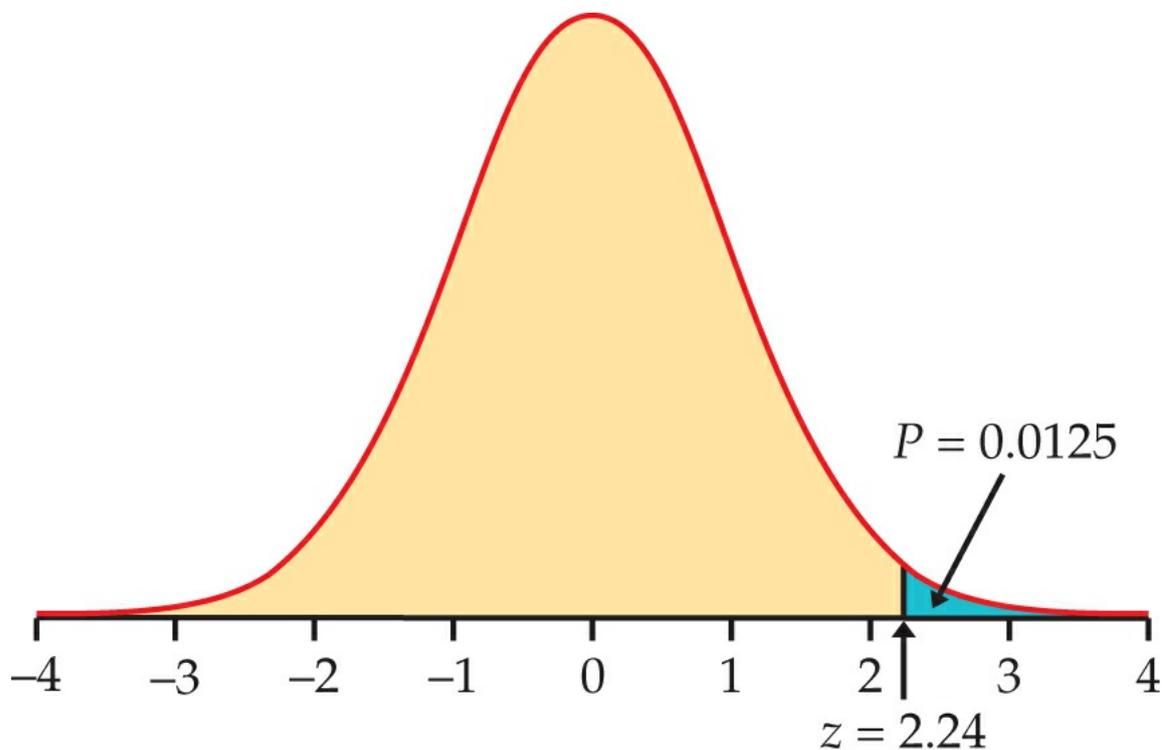
As we did in the discussion following Example 6.3, we assume that  $\sigma=100$ . The  $z$  statistic is

$$\begin{aligned}\bar{x} &= 6.79 + 6.13 + 7.173 = 6.70 \\ &= 2.24\end{aligned}$$

Because  $H_a$  is one-sided on the high side, large values of  $z$  count against  $H_0$ . From Table A, we find that the  $P$ -value is

$$P=P(Z\geq 2.24)=1-0.9875=0.0125$$

Figure 6.12 illustrates this  $P$ -value. A mean score as large as that observed would occur roughly 12 times in 1000 samples if the population mean were 475. This is convincing evidence that the mean SATM score for all California high school seniors is higher than 475. You can confidently tell this person that his or her claim is incorrect.



### Figure 6.12

Sketch of the  $P$ -value calculation for the one-sided test in Example 6.16. The test statistic is  $z = 2.24$ .

## USE YOUR KNOWLEDGE

### 6.45 Computing the test statistic and $P$ -value

You will perform a significance test of  $H_0: \mu = 25$  based on an SRS of  $n = 36$ . Assume that  $\sigma = 8$ .

- (a) If  $\bar{x} = 27.5$  what is the test statistic  $z$ ?
- (b) What is the  $P$ -value if  $H_a: \mu > 25$ ?
- (c) What is the  $P$ -value if  $H_a: \mu \neq 25$ ?

### 6.46 Testing a random number generator

Statistical software often has a “random number generator” that is supposed to produce numbers uniformly distributed between 0 and 1. If this is true, the numbers generated come from a population with  $\mu = 0.5$ . A command to generate 100 random numbers gives outcomes with mean  $\bar{x} = 0.478$  and  $s = 0.296$ . Because the sample is reasonably large, take the population standard deviation also to be  $\sigma = 0.296$ . Do we have evidence that the mean of all numbers produced by this software is not 0.5?

## Two-sided significance tests and confidence intervals

Recall the basic idea of a confidence interval, discussed in Section 6.1. We constructed an interval that would include the true value of  $\mu$  with a specified probability  $C$ . Suppose that we use a 95% confidence interval ( $C = 0.95$ ). Then the values of  $\mu_0$  that are not in our interval would seem to be incompatible with the data. This sounds like a significance test with  $\alpha = 0.05$  (or 5%) as our standard for drawing a conclusion. The following examples demonstrate that this is correct.

## Example

### 6.17 Water quality testing





The Deely Laboratory is a drinking-water testing and analysis service. One of the common contaminants it tests for is lead. Lead enters drinking water through corrosion of plumbing materials, such as lead pipes, fixtures, and solder. The service knows that their analysis procedure is unbiased but not perfectly precise, so the laboratory analyzes each water sample three times and reports the mean result. The repeated measurements follow a Normal distribution quite closely. The standard deviation of this distribution is a property of the analytic procedure and is known to be  $\sigma=0.25$  parts per billion (ppb).

The Deely Laboratory has been asked by the university to evaluate a claim that the drinking water in the Student Union has a lead concentration of 6 ppb, well below the Environmental Protection Agency's action level of 15 ppb. Since the true concentration of the sample is the mean  $\mu$  of the population of repeated analyses, the hypotheses are

$$H_0: \mu=6$$

$$H_0: \mu=6$$

The lab chooses the 1% level of significance,  $\alpha=0.01$ .

Three analyses of one specimen give concentrations

$$6.79 \quad 6.13 \quad 7.17$$

The sample mean of these readings is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{6.70 - 6.00}{0.25/\sqrt{3}} = 4.83$$

The test statistic is

$$\bar{x} \pm z^* \sigma/\sqrt{n} = 6.70 \pm 2.576(0.25/\sqrt{3})$$

Because the alternative is two-sided, the  $P$ -value is

$$P = 2P(Z \geq 4.83)$$

We cannot find this probability in Table A. The largest value of  $z$  in that table is 3.49. All that we can say from Table A is that  $P$  is less than  $2P(Z \geq 3.49) = 2(1 - 0.9998) = 0.0004$ . If we use the bottom row of Table D, we find that the largest value of  $z^*$  is 3.291, corresponding to a  $P$ -value of  $1 - 0.999 = 0.001$ . Software or a calculator could be used to give an accurate value of the  $P$ -value. However, because the  $P$ -value is clearly less than the lab's standard of 1%, we reject  $H_0$ . Because  $\bar{x}$  is larger than 6.00, we can conclude that the true concentration of lead is higher than the university's claim.

We can compute a 99% confidence interval for the same data to get a likely range for the actual mean concentration  $\mu$ .

## Example

### 6.18 99% confidence interval for the mean concentration

The 99% confidence interval for  $\mu$  in Example 6.17 is

$$\begin{aligned} z &= \bar{x} - \mu_0 \sigma / n \\ &= 6.70 \pm 0.37 \\ &= (6.33, 7.07) \end{aligned}$$

The hypothesized value  $\mu_0 = 6.00$  in Example 6.17 falls outside the confidence interval we computed in Example 6.18. In other words, it is in the region we are 99% confident that  $\mu$  is *not* in. Thus, we can reject

$$H_0: \mu = 6.0$$

at the 1% significance level. On the other hand, we cannot reject

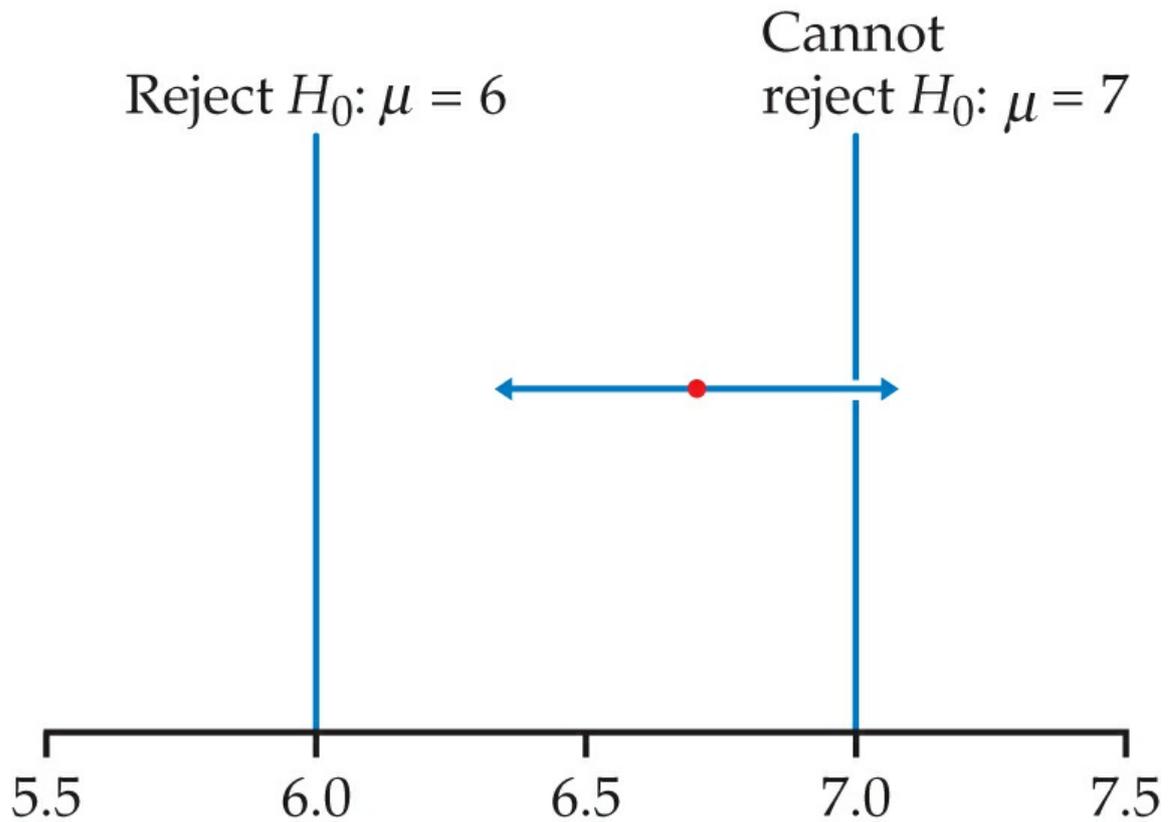
$$H_0: \mu = 7.0$$

at the 1% level in favor of the two-sided alternative  $H_a: \mu \neq 7.0$ , because 7.0 lies inside the 99% confidence interval for  $\mu$ . Figure 6.13 illustrates both cases.

The calculation in Example 6.17 for a 1% significance test is very similar to the calculation for a 99% confidence interval. In fact, a two-sided test at significance level  $\alpha$  can be carried out directly from a confidence interval with confidence level  $C = 1 - \alpha$ .

## TWO-SIDED SIGNIFICANCE TESTS AND CONFIDENCE INTERVALS

A level  $\alpha$  two-sided significance test rejects a hypothesis  $H_0: \mu = \mu_0$  exactly when the value  $\mu_0$  falls outside a level  $1 - \alpha$  confidence interval for  $\mu$ .



**Figure 6.13**

The link between two-sided significance tests and confidence intervals. For the study described in Example 6.17 and 6.18, values of  $\mu$  falling outside a 99% confidence interval can be rejected at the 1% significance level; values falling inside the interval cannot be rejected.

## USE YOUR KNOWLEDGE

### 6.47 Two-sided significance tests and confidence intervals

The  $P$ -value for a two-sided test of the null hypothesis  $H_0: \mu = 30$  is 0.041.

- Does the 95% confidence interval include the value 30? Explain.
- Does the 99% confidence interval include the value 30? Explain.

### 6.48 More on two-sided tests and confidence intervals

A 95% confidence interval for a population mean is (23,48).

- Can you reject the null hypothesis that  $\mu = 50$  against the two-sided alternative at the 5% significance level? Explain.

(b) Can you reject the null hypothesis that  $\mu=45$  against the two-sided alternative at the 5% significance level? Explain.

## The $P$ -value versus a statement of significance

The observed result in Example 6.17 was  $z=4.83$ . The conclusion that this result is significant at the 1% level does not tell the whole story. The observed  $z$  is far beyond the  $z$  corresponding to 1%, and the evidence against  $H_0$  is far stronger than 1% significance suggests. The actual  $P$ -value

$$2P(Z \geq 4.83) = 0.0000014$$

gives a better sense of how strong the evidence is. *The  $P$ -value is the smallest level  $\alpha$  at which the data are significant.* Knowing the  $P$ -value allows us to assess significance at any level.

### Example

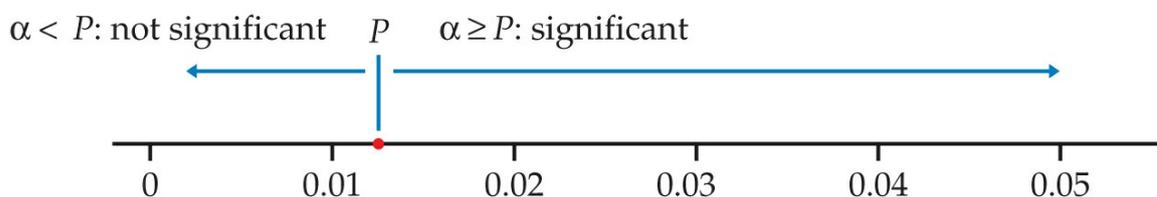
#### 6.19 Test of the mean SATM score: significance

In Example 6.16, we tested the hypotheses

$$H_0: \mu = 475$$

$$H_a: \mu > 475$$

concerning the mean SAT Mathematics score  $\mu$  of California high school seniors. The test had the  $P$ -value  $P=0.0125$ . This result is significant at the  $\alpha=0.05$  level because  $0.0125 \leq 0.05$ . It is not significant at the  $\alpha=0.01$  level, because the  $P$ -value is larger than 0.01. See Figure 6.14.



**Figure 6.14**

Link between the  $P$ -value and the significance level  $\alpha$ . An outcome with  $P$ -value  $P$  is significant at all levels  $\alpha$  at or above  $P$  and is not significant at smaller levels  $\alpha$ .

A  $P$ -value is more informative than a reject-or-not finding at a fixed significance level. But assessing significance at a fixed level  $\alpha$  is easier, because no

probability calculation is required. You need only look up a number in a table. A value  $z^*$  with a specified area to its right under the standard Normal curve is called a **critical value** of the standard Normal distribution. Because the practice of statistics almost always employs computer software or a calculator that calculates  $P$ -values automatically, the use of tables of critical values is becoming outdated. We include the usual tables of critical values (such as Table D) at the end of the book for learning purposes and to rescue students without good computing facilities. The tables can be used directly to carry out fixed  $\alpha$  tests. They also allow us to approximate  $P$ -values quickly without a probability calculation. The following example illustrates the use of Table D to find an approximate  $P$ -value.

### critical value

## Example

### 6.20 Debt levels of freshmen and seniors: assessing significance

In Example 6.11 (page 376) we found the test statistic  $z=1.33$  for testing the null hypothesis that there was no difference in the mean outstanding credit card balance between freshmen and seniors. The alternative was two-sided. Under the null hypothesis,  $z$  has a standard Normal distribution, and from the last row in Table D we can see that there is a 95% chance that  $z$  is between  $\pm 1.96$ . Therefore, we reject  $H_0$  in favor of  $H_a$  whenever  $z$  is outside this range. Since our calculated value is 1.33, we are within the range and we do not reject the null hypothesis at the 5% level of significance.

## USE YOUR KNOWLEDGE

### 6.49 $P$ -value and significance level

The  $P$ -value for a significance test is 0.021.

- Do you reject the null hypothesis at level  $\alpha=0.05$ ?
- Do you reject the null hypothesis at level  $\alpha=0.01$ ?
- Explain how you determined your answers to parts (a) and (b).

### 6.50 More on $P$ -value and significance level

The  $P$ -value for a significance test is 0.072.

- (a) Do you reject the null hypothesis at level  $\alpha=0.05$ ?
- (b) Do you reject the null hypothesis at level  $\alpha=0.01$ ?
- (c) Explain how you determined your answers to parts (a) and (b).

### 6.51 One-sided and two-sided $P$ -values

The  $P$ -value for a two-sided significance test is 0.062.

- (a) State the  $P$ -values for the two one-sided tests.
- (b) What additional information do you need to properly assign these  $P$ -values to the  $>$  and  $<$  (one-sided) alternatives?

## Section 6.2 Summary

A **test of significance** is intended to assess the evidence provided by data against a **null hypothesis  $H_0$**  in favor of an **alternative hypothesis  $H_a$** .

The hypotheses are stated in terms of population parameters. Usually  $H_0$  is a statement that no effect or no difference is present, and  $H_a$  says that there is an effect or difference, in a specific direction (**one-sided alternative**) or in either direction (**two-sided alternative**).

The test is based on a **test statistic**. The  **$P$ -value** is the probability, computed assuming that  $H_0$  is true, that the test statistic will take a value at least as extreme as that actually observed. Small  $P$ -values indicate strong evidence against  $H_0$ . Calculating  $P$ -values requires knowledge of the sampling distribution of the test statistic when  $H_0$  is true.

If the  $P$ -value is as small or smaller than a specified value  $\alpha$  the data are **statistically significant** at significance level  $\alpha$ .

Significance tests for the hypothesis  $H_0:\mu=\mu_0$  concerning the unknown mean  $\mu$  of a population are based on the  **$z$  statistic**:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{2.02 - 2.00}{0.02/\sqrt{25}} = 2.5 \geq 1.645$$

The  $z$  test assumes an SRS of size  $n$ , known population standard deviation  $\sigma$ , and either a Normal population or a large sample.  $P$ -values are computed from the Normal distribution (Table A). Fixed  $\alpha$  tests use the table of **standard Normal critical values** (Table D).

## SECTION 6.2 Exercises

For Exercises 6.38 and 6.39, see pages 375; for Exercises 6.40 and 6.41, see pages

378; for Exercises 6.42 to 6.44, see pages 381; for Exercises 6.45 and 6.46, see page 385–386; for Exercise 6.47 and 6.48, see page 388; and for Exercise 6.49 to 6.51, see page 389–390.

### 6.52 What's wrong?

Here are several situations where there is an incorrect application of the ideas presented in this section. Write a short paragraph explaining what is wrong in each situation and why it is wrong.

- (a) A researcher tests the following null hypothesis:  $H_0: \bar{x} = 23$ .
- (b) A random sample of size 30 is taken from a population that is assumed to have a standard deviation of 5. The standard deviation of the sample mean is  $5/30$ .
- (c) A study with  $\bar{x} = 45$  reports statistical significance for  $H_a: \mu > 50$ .
- (d) A researcher tests the hypothesis  $H_0: \mu = 350$  and concludes that the population mean is equal to 350.

### 6.53 What's wrong?

Here are several situations where there is an incorrect application of the ideas presented in this section. Write a short paragraph explaining what is wrong in each situation and why it is wrong.

- (a) A significance test rejected the null hypothesis that the sample mean is equal to 500.
- (b) A test preparation company wants to test that the average score of their students on the ACT is better than the national average score of 21.2. They state their null hypothesis to be  $H_0: \mu > 21.2$ .
- (c) A study summary says that the results are statistically significant and the  $P$ -value is 0.98.
- (d) The  $z$  test statistic is equal to 0.018. Because this is less than  $\alpha = 0.05$ , the null hypothesis was rejected.

### 6.54 Determining hypotheses.

State the appropriate null hypothesis  $H_0$  and alternative hypothesis  $H_a$  in each of the following cases.

- (a) A 2010 study reported that 88% of students owned a cell phone. You plan to take an SRS of students to see if the percent has increased.
- (b) The examinations in a large freshman chemistry class are scaled after grading so that the mean score is 75. The professor thinks that students who attend early-morning recitation sections will have a higher mean score than the class as a whole. Her students in these sections this semester can be considered a sample from the population of all students who might attend an early-morning section, so she compares their mean score with 75.
- (c) The student newspaper at your college recently changed the format of its opinion page. You want to test whether students find the change an improvement. You take a random sample of students and select those who regularly read the newspaper. They are asked to indicate their opinions on the changes using a five-point scale:  $-2$  if the new format is much worse than the old,  $-1$  if the new format is somewhat worse than the old,  $0$  if the new format is the same as the old,  $+1$  if the new format is somewhat better than the old, and  $+2$  if the new format is much better than the old.

### 6.55 More on determining hypotheses.

State the null hypothesis  $H_0$  and the alternative hypothesis  $H_a$  in each case. Be sure to identify the parameters that you use to state the hypotheses.

(a) A university gives credit in first-year calculus to students who pass a placement test. The mathematics department wants to know if students who get credit in this way differ in their success with second-year calculus. Scores in second-year calculus are scaled so the average each year is equivalent to a 77. This year 21 students who took second-year calculus passed the placement test.

(b) Experiments on learning in animals sometimes measure how long it takes a mouse to find its way through a maze. The mean time is 20 seconds for one particular maze. A researcher thinks that playing rap music will cause the mice to complete the maze more slowly. She measures how long each of 12 mice takes with the rap music as a stimulus.

(c) The average square footage of one-bedroom apartments in a new student-housing development is advertised to be 880 square feet. A student group thinks that the apartments are smaller than advertised. They hire an engineer to measure a sample of apartments to test their suspicion.

### 6.56 Even more on determining hypotheses.

In each of the following situations, state an appropriate null hypothesis  $H_0$  and alternative hypothesis  $H_a$ . Be sure to identify the parameters that you use to state the hypotheses. (We have not yet learned how to test these hypotheses.)

(a) A sociologist asks a large sample of high school students which television channel they like best. She suspects that a higher percent of males than of females will name MTV as their favorite channel.

(b) An education researcher randomly divides sixth-grade students into two groups for physical education class. He teaches both groups basketball skills, using the same methods of instruction in both classes. He encourages Group A with compliments and other positive behavior but acts cool and neutral toward Group B. He hopes to show that positive teacher attitudes result in a higher mean score on a test of basketball skills than do neutral attitudes.

(c) An education researcher believes that among college students there is a negative correlation between time spent at social network sites and self-esteem, measured on a 0 to 100 scale. To test this, she gathers social-networking information and self-esteem data from a sample of students at your college.

### 6.57 Translating research questions into hypotheses.

Translate each of the following research questions into appropriate  $H_0$  and  $H_a$ .

(a) U.S. Census Bureau data show that the mean household income in the area served by a shopping mall is \$42,800 per year. A market research firm questions shoppers at the mall to find out whether the mean household income of mall shoppers is higher than that of the general population.

(b) Last year, your online registration technicians took an average of 0.4 hours to respond to trouble calls from students trying to register. Do this year's data show a different average response time?

### 6.58 Computing the $P$ -value.

A test of the null hypothesis  $H_0: \mu = \mu_0$  gives test statistic  $z = 1.77$ .

(a) What is the  $P$ -value if the alternative is  $H_a: \mu > \mu_0$ ?

(b) What is the  $P$ -value if the alternative is  $H_a: \mu < \mu_0$ ?

(c) What is the  $P$ -value if the alternative is  $H_a:\mu\neq\mu_0$ ?

### 6.59 More on computing the $P$ -value.

A test of the null hypothesis  $H_0:\mu=\mu_0$  gives test statistic  $z=-1.69$ .

(a) What is the  $P$ -value if the alternative is  $H_a:\mu>\mu_0$ ?

(b) What is the  $P$ -value if the alternative is  $H_a:\mu<\mu_0$ ?

(c) What is the  $P$ -value if the alternative is  $H_a:\mu\neq\mu_0$ ?

### 6.60 Timing of food intake and weight loss.

A study found that a large group of late lunch eaters lost less weight over a 20-week observation period than a large group of early lunch eaters ( $P=0.002$ ).<sup>17</sup> Explain what this  $P=0.002$  means in a way that could be understood by someone who has not studied statistics.

### 6.61 Peer pressure and choice of major.

A study followed a cohort of students entering a business/economics program.<sup>18</sup> All students followed a common track during the first three semesters and then chose to specialize in either business or economics. Through a series of surveys, the researchers were able to classify roughly 50% of the students as either peer driven (ignored abilities and chose major to follow peers) or ability driven (ignored peers and chose major based on ability). When looking at entry wages after graduation, the researchers conclude that a peer-driven student can expect an average wage that is 13% less than that of an ability-driven student. The report states that the significance level is  $P=0.09$ . Can you be confident of the researchers' conclusion regarding the wage decrease? Explain your answer.

### 6.62 Symbol of wealth in ancient China?

Every society has its own symbols of wealth and prestige. In ancient China, it appears that owning pigs was such a symbol. Evidence comes from examining burial sites. If the skulls of sacrificed pigs tend to appear along with expensive ornaments, that suggests that the pigs, like the ornaments, signal the wealth and prestige of the person buried. A study of burials from around 3500 B.C. concluded that "there are striking differences in grave goods between burials with pig skulls and burials without them... . A test indicates that the two samples of total artifacts are significantly different at the 0.01 level."<sup>19</sup> Explain clearly why "significantly different at the 0.01 level" gives good reason to think that there really is a systematic difference between burials that contain pig skulls and those that lack them.

### 6.63 Alcohol awareness among college students.

A study of alcohol awareness among college students reported a higher awareness for students enrolled in a health and safety class than for those enrolled in a statistics class.<sup>20</sup> The difference is described as being statistically significant. Explain what this means in simple terms and offer an explanation for why the health and safety students had a higher mean score.

### 6.64 Change in eighth-grade average mathematics score.

A report based on the 2011 National Assessment of Educational Progress (NAEP)<sup>21</sup> states that the average score on their mathematics test for eighth-grade students attending public schools is significantly higher than in 2009. The report also states that the average score for eighth-grade students attending private schools is not significantly different from the average score in 2009. A footnote states that comparisons are determined by two-sided statistical tests with 0.05 as the level of significance. Explain what this footnote means in language understandable to someone who knows no statistics. Do not use the word “significance” in your answer.

### **6.65 More on change in eighth-grade average mathematics score.**

Refer to the previous exercise. On the basis of the NAEP study, a friend who works for the school newspaper wants to report that between 2009 and 2011 the average mathematics score improved for students attending public schools but stayed the same for students attending private schools. Do you agree with this statement? Explain your answer.

### **6.66 Background television in homes of U.S. children.**

In one study, U.S. parents were surveyed to determine the amount of background television their children were exposed to. A total of  $n=1454$  families with one child between the ages of 8 months and 8 years participated.<sup>22</sup> For those families in which the caregiver had a high school degree or less, the child was exposed to an average of 313.0 minutes of background television per day. For those families in which the caregiver had some college or a college degree, the child was exposed to an average of 218.8 minutes per day. These average times were reported to be significantly different, with  $P<0.05$ . The actual  $P$ -value is 0.003. Explain why the actual  $P$ -value is more informative than the statement of significance at the 0.05 level.

### **6.67 Sleep quality and elevated blood pressure.**

A study looked at  $n=238$  adolescents, all free of severe illness.<sup>23</sup> Subjects wore a wrist actigraph, which allowed the researchers to estimate sleep patterns. Those subjects classified as having low sleep efficiency had an average systolic blood pressure that was 5.8 millimeters of mercury (mm Hg) higher than that of other adolescents. The standard deviation of this difference is 1.4 mm Hg. Based on these results, test whether this difference is significant at the 0.01 level.



### **6.68 Are the pine trees randomly distributed from north to south?**

In Example 6.1 (page 352) we looked at the distribution of longleaf pine trees in the Wade Tract. One way to formulate hypotheses about whether or not the trees are randomly distributed in the tract is to examine the average location in the north–south direction. The values range from 0 to 200, so if the trees are uniformly distributed in this direction, any difference from the middle value (100) should be due to chance variation. The sample mean for the 584 trees in the tract is 99.74. A theoretical calculation based on the assumption that the trees are uniformly distributed gives a standard deviation of 58. Carefully state the null and alternative hypotheses in terms of this variable. Note that this requires that you translate the research question about the random distribution of the trees into specific statements about the mean of a probability distribution. Test your hypotheses, report your results, and write a short summary of what you have found.



### **6.69 Are the pine trees randomly distributed from east to west?**

Answer the questions in the previous exercise for the east–west direction, for which the sample mean is 113.8.

## 6.70 Who is the author?

Statistics can help decide the authorship of literary works. Sonnets by a certain Elizabethan poet are known to contain an average of  $\mu=8.9$  new words (words not used in the poet's other works). The standard deviation of the number of new words is  $\sigma=2.5$ . Now a manuscript with six new sonnets has come to light, and scholars are debating whether it is the poet's work. The new sonnets contain an average of  $\bar{x}=10.2$  words not used in the poet's known works. We expect poems by another author to contain more new words, so to see if we have evidence that the new sonnets are not by our poet we test

$$H_0:\mu=8.9$$

$$H_a:\mu>8.9$$

Give the  $z$  test statistic and its  $P$ -value. What do you conclude about the authorship of the new poems?

## 6.71 Attitudes toward school.

The Survey of Study Habits and Attitudes (SSHA) is a psychological test that measures the motivation, attitude toward school, and study habits of students. Scores range from 0 to 200. The mean score for U.S. college students is about 115, and the standard deviation is about 30. A teacher who suspects that older students have better attitudes toward school gives the SSHA to 25 students who are at least 30 years of age. Their mean score is  $\bar{x}=127.8$ .

(a) Assuming that  $\sigma=30$  for the population of older students, carry out a test of

$$H_0:\mu=115$$

$$H_a:\mu>115$$

Report the  $P$ -value of your test, and state your conclusion clearly.

(b) Your test in part (a) required two important assumptions in addition to the assumption that the value of  $\sigma$  is known. What are they? Which of these assumptions is most important to the validity of your conclusion in part (a)?

## 6.72 Nutritional intake among Canadian high-performance athletes.

Since previous studies have reported that elite athletes are often deficient in their nutritional intake (for example, total calories, carbohydrates, protein), a group of researchers decided to evaluate Canadian high-performance athletes.<sup>24</sup> A total of  $n=324$  athletes from eight Canadian sports centers participated in the study. One reported finding was that the average caloric intake among the  $n=201$  women was 2403.7 kilocalories per day (kcal/d). The recommended amount is 2811.5 kcal/d. Is there evidence that female Canadian athletes are deficient in caloric intake?

(a) State the appropriate  $H_0$  and  $H_a$  to test this.

(b) Assuming a standard deviation of 880 kcal/d, carry out the test. Give the  $P$ -value, and then interpret the result in plain language.

## 6.73 Are the measurements similar?

Refer to Exercise 6.30 (page 371). In addition to the computer's calculations of miles per gallon, the driver also recorded the miles per gallon by dividing the miles driven by the number of gallons at each fill-up. The following data are the differences between the computer's and the driver's calculations for that

random sample of 20 records. The driver wants to determine if these calculations are different. Assume that the standard deviation of a difference is  $\sigma=30$ .  **MPGDIFF**

5.0	6.5	-0.6	1.7	3.7	4.5	8.0	2.2	4.9	3.0
4.4	0.1	3.0	1.1	1.1	5.0	2.1	3.7	-0.6	-4.2

- (a) State the appropriate  $H_0$  and  $H_a$  to test this suspicion.
- (b) Carry out the test. Give the  $P$ -value, and then interpret the result in plain language.

### 6.74 Adjusting for the cost of living.

In Example 6.9 (page 373), we compared the average credit card balance between undergraduates in the Midwest and the South. In testing the difference, we considered a one-sided test because the cost of living is higher in the South (Example 6.14). Assuming that \$1 in the Midwest is worth about \$1.09 in the South, test whether there is a difference between the average balances in the two regions using South dollars. For simplicity, assume that the standard deviation is unchanged.

### 6.75 Nicotine content in cigarettes.

According to data from the Tobacco Institute Testing Laboratory, Camel Lights king size cigarettes contain an average of 0.61 milligrams of nicotine. An advocacy group commissions an independent test to see if the mean nicotine content is higher than the industry laboratory claims.

- (a) What are  $H_0$  and  $H_a$ ?
- (b) Suppose that the test statistic is  $z=1.72$ . Is this result significant at the 5% level?
- (c) Is the result significant at the 1% level?

### 6.76 Impact of $\bar{x}$ on significance.

The *Statistical Significance* applet illustrates statistical tests with a fixed level of significance for Normally distributed data with known standard deviation. Open the applet and keep the default settings for the null ( $\mu=0$ ) and the alternative ( $\mu>0$ ) hypotheses, the sample size ( $n=10$ ), the standard deviation ( $\sigma=1$ ), and the significance level ( $\alpha=0.05$ ). In the “I have data, and the observed  $\bar{x}$  is  $\bar{x} =$ ” box enter the value 1. Is the difference between  $\bar{x}$  and  $\mu_0$  significant at the 5% level? Repeat for  $\bar{x}$  equal to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. Make a table giving  $\bar{x}$  and the results of the significance tests. What do you conclude?

### 6.77 Effect of changing $\alpha$ on significance.

Repeat the previous exercise with significance level  $\alpha = 0.01$ . How does the choice of  $\alpha$  affect which values of  $\bar{x}$  are far enough away from  $\mu_0$  to be statistically significant?

### 6.78 Changing to a two-sided alternative.

Repeat the previous exercise but with the two-sided alternative hypothesis. How does this change affect which values of  $\bar{x}$  are far enough away from  $\mu_0$  to be statistically significant at the 0.01 level?

### 6.79 Changing the sample size.

Refer to Exercise 6.76. Suppose that you increase the sample size  $n$  from 10 to 40. Again make a table giving  $\bar{x}$  and the results of the significance tests at the 0.05 significance level. What do you conclude?

### 6.80 Impact of $\bar{x}$ on the $P$ -value.

We can also study the  $P$ -value using the *Statistical Significance* applet. Reset the applet to the default settings for the null ( $\mu=0$ ) and the alternative ( $\mu>0$ ) hypotheses, the sample size ( $n = 10$ ), the standard deviation ( $\sigma=1$ ), and the significance level ( $\alpha=0.05$ ). In the “I have data, and the observed  $\bar{x}$  is  $\bar{x} =$ ” box enter the value 1. What is the  $P$ -value? It is shown at the top of the blue vertical line. Repeat for  $\bar{x}$  equal to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. Make a table giving  $\bar{x}$  and  $P$ -values. How does the  $P$ -value change as  $\bar{x}$  moves farther away from  $\mu_0$ ?

### 6.81 Changing to a two-sided alternative, continued.

Repeat the previous exercise but with the two-sided alternative hypothesis. How does this change affect the  $P$ -values associated with each  $\bar{x}$ ? Explain why the  $P$ -values change in this way.

### 6.82 Other changes and the $P$ -value.

Refer to the previous exercise.

- What happens to the  $P$ -values when you change the significance level  $\alpha$  to 0.01? Explain the result.
- What happens to the  $P$ -values when you change the sample size  $n$  from 10 to 40? Explain the result.

## 6.83 Understanding levels of significance.

Explain in plain language why a significance test that is significant at the 1% level must always be significant at the 5% level.

## 6.84 More on understanding levels of significance.

You are told that a significance test is significant at the 5% level. From this information can you determine whether or not it is significant at the 1% level? Explain your answer.

## 6.85 Test statistic and levels of significance.

Consider a significance test for a null hypothesis versus a two-sided alternative. Give a value of  $z$  that will give a result significant at the 1% level but not at the 0.5% level.

## 6.86 Using Table D to find a $P$ -value.

You have performed a two-sided test of significance and obtained a value of  $z=2.31$ . Use Table D to find the approximate  $P$ -value for this test.

### **6.87 More on using Table D to find a $P$ -value.**

You have performed a one-sided test of significance and obtained a value of  $z=0.54$ . Use Table D to find the approximate  $P$ -value for this test when the alternative is greater than.

### **6.88 Using Table A and Table D to find a $P$ -value.**

Consider a significance test for a null hypothesis versus a two-sided alternative. Between what values from Table D does the  $P$ -value for an outcome  $z=1.88$  lie? Calculate the  $P$ -value using Table A, and verify that it lies between the values you found from Table D.

### **6.89 More on using Table A and Table D to find a $P$ -value.**

Refer to the previous exercise. Find the  $P$ -value for  $z=-1.88$ .

## 6.3 Use and Abuse of Tests

When you complete this section, you will be able to

- Explain why it is important to report the  $P$ -value and not just report whether the result is statistically significant or not.
- Discriminate between practical (or scientific) significance and statistical significance.
- Identify poorly designed studies where formal statistical inference is suspect.
- Understand the problems with searching solely for statistical significance, whether through the investigation of multiple tests or by identifying and testing using the same data set.

Carrying out a test of significance is often quite simple, especially if the  $P$ -value is given effortlessly by a computer. Using tests wisely is not so simple. Each test is valid only in certain circumstances, with properly produced data being particularly important.

The  $z$  test, for example, should bear the same warning label that was attached in Section 6.1 to the corresponding confidence interval (page 365). Similar warnings accompany the other tests that we will learn. There are additional caveats that concern tests more than confidence intervals, enough to warrant this separate section. Some hesitation about the unthinking use of significance tests is a sign of statistical maturity.

The reasoning of significance tests has appealed to researchers in many fields, so that tests are widely used to report research results. In this setting  $H_a$  is a “research hypothesis” asserting that some effect or difference is present. The null hypothesis  $H_0$  says that there is no effect or no difference. A low  $P$ -value represents good evidence that the research hypothesis is true. Here are some comments on the use of significance tests, with emphasis on their use in reporting scientific research.

### Choosing a level of significance



The spirit of a test of significance is to give a clear statement of the degree of

evidence provided by the sample against the null hypothesis. The  $P$ -value does this. It is common practice to report  $P$ -values and to describe results as statistically significant whenever  $P \leq 0.05$ . *However, there is no sharp border between “significant” and “not significant,” only increasingly strong evidence as the  $P$ -value decreases.* Having both the  $P$ -value and the statement that we reject or fail to reject  $H_0$  allows us to draw better conclusions from our data.

## Example

### 6.21 Information provided by the $P$ -value

Suppose that the test statistic for a two-sided significance test for a population mean is  $z=1.95$ . From Table A we can calculate the  $P$ -value. It is

$$P=2[1-P(Z \leq 1.95)]=2(1-0.9744)=0.0512$$

We have failed to meet the standard of evidence for  $\alpha=0.05$ . However, with the information provided by the  $P$ -value, we can see that the result just barely missed the standard. If the effect in question is interesting and potentially important, we might want to design another study with a larger sample to investigate it further.

Here is another example where the  $P$ -value provides useful information beyond that provided by the statement that we reject or fail to reject the null hypothesis.

## Example

### 6.22 More on information provided by the $P$ -value

We have a test statistic of  $z=-4.66$  for a two-sided significance test on a population mean. Software tells us that the  $P$ -value is 0.000003. This means that there are 3 chances in 1, 000,000 of observing a sample mean this far or farther away from the null hypothesized value of  $\mu$ . This kind of event is virtually impossible if the null hypothesis is true. There is no ambiguity in the result; we can clearly reject the null hypothesis.

We frequently report small  $P$ -values such as that in the previous example as  $P < 0.001$ . This corresponds to a chance of 1 in 1000 and is sufficiently small to lead us to a clear rejection of the null hypothesis.

One reason for the common use of  $\alpha = 0.05$  is the great influence of Sir R. A. Fisher, the inventor of formal statistical methods for analyzing experimental data. Here is his opinion on choosing a level of significance: “A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance.”<sup>25</sup>

## What statistical significance does not mean

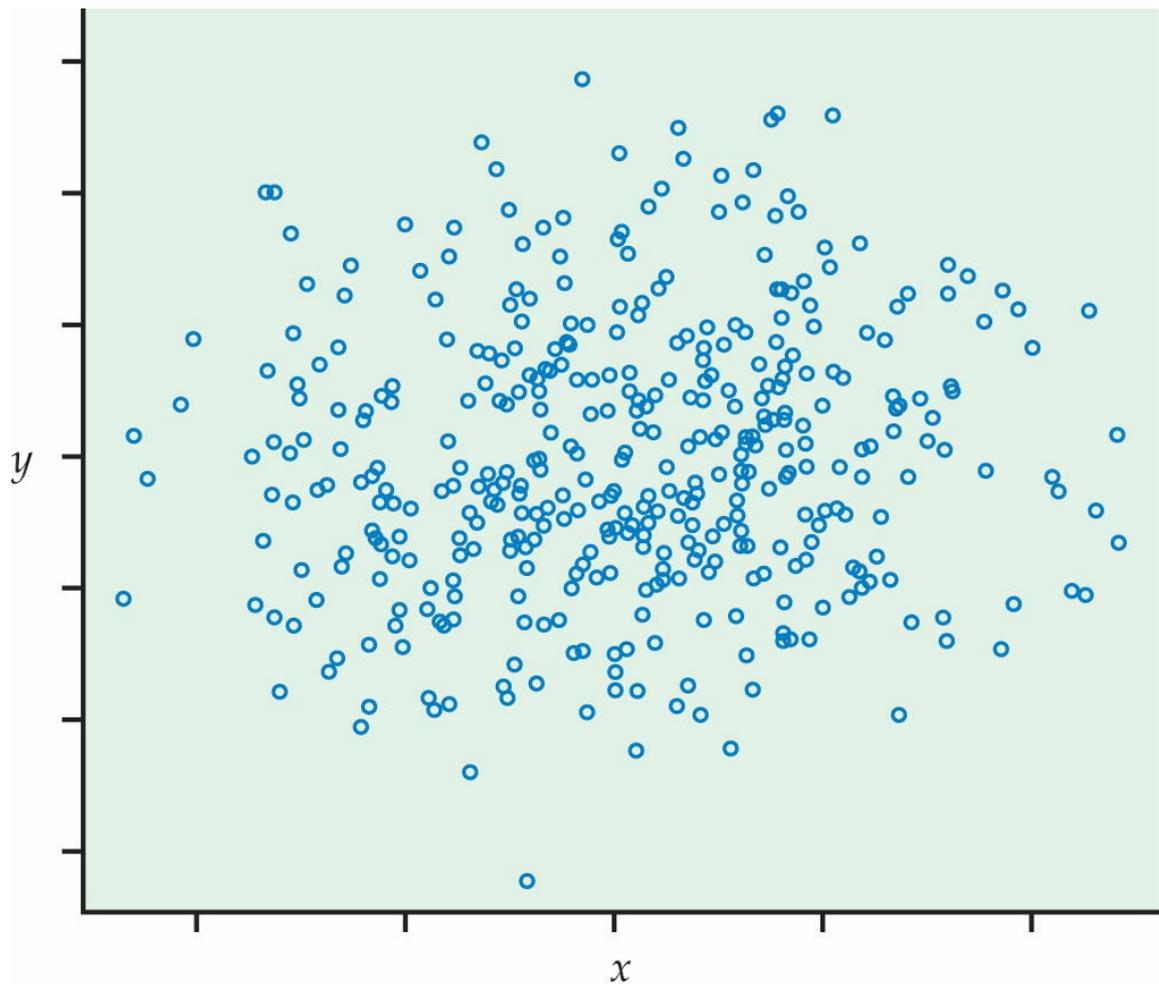


When a null hypothesis (“no effect” or “no difference”) can be rejected at the usual level  $\alpha = 0.05$  there is good evidence that an effect is present. That effect, however, can be extremely small. *When large samples are available, even tiny deviations from the null hypothesis will be significant.*

### Example

#### 6.23 It's significant but is it important?

Suppose that we are testing the hypothesis of no correlation between two variables. With 400 observations, an observed correlation of only  $r = 0.1$  is significant evidence at the  $\alpha = 0.05$  level that the correlation in the population is not zero. Figure 6.15 is an example of 400  $(x, y)$  pairs that have an observed correlation of 0.10. The low significance level does *not* mean that there is a strong association, only that there is strong evidence of some association. The proportion of the variability in one of the variables explained by the other is  $r^2 = 0.01$  or 1%.



**Figure 6.15**

Scatterplot of  $n = 400$  observations with an observed correlation of 0.10. There is not a strong association between the two variables even though there is significant evidence ( $P < 0.05$ ) that the population correlation is not zero.



For practical purposes, we might well decide to ignore this association. *Statistical significance is not the same as practical significance.* Statistical significance rarely tells us about the importance of the experimental results. This depends on the context of the experiment.



The remedy for attaching too much importance to statistical significance is to pay attention to the actual experimental results as well as to the  $P$ -value. Plot your data and examine them carefully. Beware of outliers. *The foolish user of statistics*

who feeds the data to a computer without exploratory analysis will often be embarrassed. It is usually wise to give a confidence interval for the parameter in which you are interested. Confidence intervals are not used as often as they should be, while tests of significance are overused.

## USE YOUR KNOWLEDGE

### 6.90 Is it significant?

More than 200,000 people worldwide take the GMAT examination each year when they apply for MBA programs. Their scores vary Normally with mean about  $\mu=525$  and standard deviation about  $\sigma=100$ . One hundred students go through a rigorous training program designed to raise their GMAT scores. Test the following hypotheses about the training program

$$H_0: \mu=525$$

$$H_a: \mu>525$$

in each of the following situations.

- (a) The students' average score is  $\bar{x}=541.4$ . Is this result significant at the 5% level?
- (b) Now suppose that the average score is  $\bar{x}=541.5$ . Is this result significant at the 5% level?
- (c) Explain how you would reconcile this difference in significance, especially if any increase greater than 15 points is considered a success.

## Don't ignore lack of significance

There is a tendency to conclude that there is no effect whenever a  $P$ -value fails to attain the usual 5% standard. A provocative editorial in the *British Medical Journal* entitled "Absence of Evidence Is Not Evidence of Absence" deals with this issue.<sup>26</sup> Here is one of the examples they cite.

### Example

## 6.24 Interventions to reduce HIV-1 transmission

A randomized trial of interventions for reducing transmission of HIV-1 reported an incident rate ratio of 1.00, meaning that the intervention group and the control group both had the same rate of HIV-1 infection. The 95% confidence interval was reported as 0.63 to 1.58.<sup>27</sup> The editorial notes that a summary of these results that says the intervention has no effect on HIV-1 infection is misleading. The confidence interval indicates that the intervention may be capable of achieving a 37% decrease in infection; it might also be harmful and produce a 58% increase in infection. Clearly, more data are needed to distinguish between these possibilities.

The situation can be worse. Research in some fields has rarely been published unless significance at the 0.05 level is attained.

### Example

## 6.25 Journal survey of reported significance results

A survey of four journals published by the American Psychological Association showed that of 294 articles using statistical tests, only 8 reported results that did not attain the 5% significance level.<sup>28</sup> It is very unlikely that these were the only 8 studies of scientific merit that did not attain significance at the 0.05 level. Manuscripts describing other studies were likely rejected because of a lack of statistical significance or never submitted in the first place due to the expectation of rejection.

In some areas of research, small effects that are detectable only with large sample sizes can be of great practical significance. Data accumulated from a large number of patients taking a new drug may be needed before we can conclude that there are life-threatening consequences for a small number of people.

On the other hand, sometimes a meaningful result is not found significant.

### Example

## 6.26 A meaningful but statistically insignificant result

A sample of size 10 gave a correlation of  $r=0.5$  between two variables. The  $P$ -value is 0.102 for a two-sided significance test. In many situations, a correlation this large would be interesting and worthy of additional study. When it takes a lot of effort (say, in terms of time or money) to obtain samples, researchers often use small studies like these as pilot projects to gain interest from various funding sources. With financial support, a larger, more powerful study can then be run.



*Another important aspect of planning a study is to verify that the test you plan to use does have high probability of detecting an effect of the size you hope to find. This probability is the *power* of the test. Power calculations are discussed in Section 6.4.*

## Statistical inference is not valid for all sets of data



design of experiments, p. 175



In Chapter 3, we learned that badly designed surveys or experiments often produce invalid results. *Formal statistical inference cannot correct basic flaws in the design.*

### Example

#### 6.27 English vocabulary and studying a foreign language

There is no doubt that there is a significant difference in English vocabulary scores between high school seniors who have studied a foreign language and those who have not. But because the effect of actually studying a language is confounded with the differences between students who choose language study

and those who do not, this statistical significance is hard to interpret. The most plausible explanation is that students who were already good at English chose to study another language. A randomized comparative experiment would isolate the actual effect of language study and so make significance meaningful. However, such an experiment probably could not be done.



Tests of significance and confidence intervals are based on the laws of probability. Randomization in sampling or experimentation ensures that these laws apply. But we must often analyze data that do not arise from randomized samples or experiments. *To apply statistical inference to such data, we must have confidence in a probability model for the data.* The diameters of successive holes bored in auto engine blocks, for example, may behave like independent observations from a Normal distribution. We can check this probability model by examining the data. If the Normal distribution model appears correct, we can apply the methods of this chapter to do inference about the process mean diameter  $\mu$ .

## USE YOUR KNOWLEDGE

### 6.91 Home security systems

A recent TV advertisement for home security systems said that homes without an alarm system are three times more likely to be broken into. Suppose that this conclusion was obtained by examining an SRS of police records of break-ins and determining whether the percent of homes with alarm systems was significantly smaller than 50%. Explain why the significance of this study is suspect and propose an alternative study that would help clarify the importance of an alarm system.

## Beware of searching for significance



Statistical significance is an outcome much desired by researchers. It means (or ought to mean) that you have found an effect that you were looking for. *The*

*reasoning behind statistical significance works well if you decide what effect you are seeking, design an experiment or sample to search for it, and use a test of significance to weigh the evidence you get.* But because a successful search for a new scientific phenomenon often ends with statistical significance, it is all too tempting to make significance itself the object of the search. There are several ways to do this, none of them acceptable in polite scientific society.

## Example

### 6.28 Genomics studies

In genomics experiments, it is common to assess the differences in expression for tens of thousands of genes. If each of these genes was examined separately and statistical significance declared for all that had  $P$ -values that pass the 0.05 standard, we would have quite a mess. In the absence of any real biological effects, we would expect that, by chance alone, approximately 5% of these tests will show statistical significance. Much research in genomics is directed toward appropriate ways to deal with this situation.<sup>29</sup>

We do not mean that searching data for suggestive patterns is not proper scientific work. It certainly is. Many important discoveries have been made by accident rather than by design. Exploratory analysis of data is an essential part of statistics. We do mean that the usual reasoning of statistical inference does not apply when the search for a pattern is successful. *You cannot legitimately test a hypothesis on the same data that first suggested that hypothesis.* The remedy is clear. Once you have a hypothesis, design a study to search specifically for the effect you now think is there. If the result of this study is statistically significant, you have real evidence.



## Section 6.3 Summary

$P$ -values are more informative than the reject-or-not result of a level  $\mu$  test. Beware of placing too much weight on traditional values of  $\mu$ , such as  $\alpha=0.05$ .

Very small effects can be highly significant (small  $P$ ), especially when a test is based on a large sample. A statistically significant effect need not be practically important. Plot the data to display the effect you are seeking, and use confidence

intervals to estimate the actual values of parameters.

On the other hand, lack of significance does not imply that  $H_0$  is true, especially when the test has a low probability of detecting an effect.

Significance tests are not always valid. Faulty data collection, outliers in the data, and testing a hypothesis on the same data that suggested the hypothesis can invalidate a test. Many tests run at once will probably produce some significant results by chance alone, even if all the null hypotheses are true.

## SECTION 6.3 Exercises

For Exercise 6.90, see page 397; and for Exercise 6.91, see page 399.

### 6.92 A role as a statistical consultant.

You are the statistical expert for a graduate student planning her PhD research. After you carefully present the mechanics of significance testing, she suggests using  $\alpha=0.20$  for the study because she would be more likely to obtain statistically significant results and she *really* needs significant results to graduate. Explain in simple terms why this would not be a good use of statistical methods.

### 6.93 What do you know?

A research report described two results that both achieved statistical significance at the 5% level. The  $P$ -value for the first is 0.048; for the second it is 0.0002. Do the  $P$ -values add any useful information beyond that conveyed by the statement that both results are statistically significant? Write a short paragraph explaining your views on this question.

### 6.94 Selective publication based on results.

In addition to statistical significance, selective publication can also be due to the observed outcome. A recent review of 74 studies of antidepressant agents found 38 studies with positive results and 36 studies with negative or questionable results. All but 1 of the 38 positive studies were published. Of the remaining 36, 22 were not published, and 11 were published in such a way as to convey a positive outcome.<sup>30</sup> Describe how this selective reporting can have adverse consequences on health care.

### 6.95 What a test of significance can answer.

Explain whether a test of significance can answer each of the following questions.

- (a) Is the sample or experiment properly designed?
- (b) Is the observed effect compatible with the null hypothesis?
- (c) Is the observed effect important?

### 6.96 Vitamin C and colds.

In a study to investigate whether vitamin C will prevent colds, 400 subjects are assigned at random to one of two groups. The experimental group takes a vitamin C tablet daily, while the control group takes a

placebo. At the end of the experiment, the researchers calculate the difference between the percents of subjects in the two groups who were free of colds. This difference is statistically significant ( $P=0.03$ ) in favor of the vitamin C group. Can we conclude that vitamin C has a strong effect in preventing colds? Explain your answer.

### 6.97 How far do rich parents take us?

How much education children get is strongly associated with the wealth and social status of their parents, termed “socioeconomic status,” or SES. The SES of parents, however, has little influence on whether children who have graduated from college continue their education. One study looked at whether college graduates took the graduate admissions tests for business, law, and other graduate programs. The effects of the parents’ SES on taking the LSAT test for law school were “both statistically insignificant and small.”

- (a) What does “statistically insignificant” mean?
- (b) Why is it important that the effects were small in size as well as statistically insignificant?

### 6.98 Do you agree?

State whether or not you agree with each of the following statements and provide a short summary of the reasons for your answers.

- (a) If the  $P$ -value is larger than 0.05, the null hypothesis is true.
- (b) Practical significance is not the same as statistical significance.
- (c) We can perform a statistical analysis using any set of data.
- (d) If you find an interesting pattern in a set of data, it is appropriate to then use a significance test to determine its significance.
- (e) It’s always better to use a significance level of  $\alpha=0.05$  than to use  $\alpha=0.01$  because it is easier to find statistical significance.

### 6.99 Practical significance and sample size.

Every user of statistics should understand the distinction between statistical significance and practical importance. A sufficiently large sample will declare very small effects statistically significant. Consider the study of elite female Canadian athletes in Exercise 6.72 (page 393). Female athletes were consuming an average of 2403.7 kcal/d with a standard deviation of 880 kcal/d. Suppose that a nutritionist is brought in to implement a new health program for these athletes. This program should increase mean caloric intake but not change the standard deviation. Given the standard deviation and how calorie deficient these athletes are, a change in the mean of 50 kcal/d to 2453.7 is of little importance. However, with a large enough sample, this change can be significant. To see this, calculate the  $P$ -value for the test of

$$H_0: \mu = 2403.7$$

$$H_a: \mu > 2403.7$$

in each of the following situations:

- (a) A sample of 100 athletes; their average caloric intake is  $\bar{x}=2453.7$ .
- (b) A sample of 500 athletes; their average caloric intake is  $\bar{x}=2453.7$ .
- (c) A sample of 2500 athletes; their average caloric intake is  $\bar{x}=2453.7$ .

### 6.100 Statistical versus practical significance.

A study with 7500 subjects reported a result that was statistically significant at the 5% level. Explain why this result might not be particularly important.

### 6.101 More on statistical versus practical significance.

A study with 14 subjects reported a result that failed to achieve statistical significance at the 5% level. The  $P$ -value was 0.051. Write a short summary of how you would interpret these findings.



### 6.102 Find journal articles.

Find two journal articles that report results with statistical analyses. For each article, summarize how the results are reported and write a critique of the presentation. Be sure to include details regarding use of significance testing at a particular level of significance,  $P$ -values, and confidence intervals.

### 6.103 Create an example of your own.

For each of the following cases, provide an example and an explanation as to why it is appropriate.

- (a) A set of data or an experiment for which statistical inference is not valid.
- (b) A set of data or an experiment for which statistical inference is valid.



### 6.104 Predicting success of trainees.

What distinguishes managerial trainees who eventually become executives from those who, after expensive training, don't succeed and leave the company? We have abundant data on past trainees—data on their personalities and goals, their college preparation and performance, even their family backgrounds and their hobbies. Statistical software makes it easy to perform dozens of significance tests on these dozens of variables to see which ones best predict later success. We find that future executives are significantly more likely than washouts to have an urban or suburban upbringing and an undergraduate degree in a technical field.

Explain clearly why using these “significant” variables to select future trainees is not wise. Then suggest a follow-up study using this year's trainees as subjects that should clarify the importance of the variables identified by the first study.

### 6.105 Searching for significance.

Give an example of a situation where searching for significance would lead to misleading conclusions.

### 6.106 More on searching for significance.

You perform 1000 significance tests using  $\alpha=0.05$ . Assuming that all null hypotheses are true, about how many of the test results would you expect to be statistically significant? Explain how you obtained your answer.

### 6.107 Interpreting a very small $P$ -value.

Assume that you are performing a large number of significance tests. Let  $n$  be the number of these tests. How large would  $n$  need to be for you to expect about one  $P$ -value to be 0.00001 or smaller? Use this information to write an explanation of how to interpret a result that has  $P=0.00001$  in this setting.

### 6.108 An adjustment for multiple tests.

One way to deal with the problem of misleading  $P$ -values when performing more than one significance test is to adjust the criterion you use for statistical significance. The **Bonferroni procedure** does this in a simple way. If you perform two tests and want to use the  $\alpha=5\%$  significance level, you would require a  $P$ -value of  $0.05/2=0.025$  to declare either one of the tests significant. In general, if you perform  $k$  tests and want protection at level  $\mu$ , use  $\alpha/k$  as your cutoff for statistical significance. You perform six tests and obtain individual  $P$ -values of 0.083, 0.032, 0.246, 0.003, 0.010, and  $<0.001$ . Which of these are statistically significant using the Bonferroni procedure with  $\alpha=0.05$ ?

### 6.109 Significance using the Bonferroni procedure.

Refer to the previous exercise. A researcher has performed 12 tests of significance and wants to apply the Bonferroni procedure with  $\alpha=0.05$ . The calculated  $P$ -values are 0.041, 0.569, 0.050, 0.416, 0.002, 0.006, 0.286, 0.021, 0.888, 0.010,  $<0.002$ , and 0.533. Which of these tests reject their null hypotheses with this procedure?

## 6.4 Power and Inference as a Decision

When you complete this section, you will be able to

- Define what is meant by the power of a test.
- Determine the power of a test to detect an alternative for a given sample size  $n$ .
- Describe the two types of possible errors when performing a test that focuses on deciding between two hypotheses.
- Relate the two errors to the significance level and power of the test.

Although we prefer to use  $P$ -values rather than the reject-or-not view of the level  $\mu$  significance test, the latter view is very important for planning studies and for understanding statistical decision theory. We will discuss these two topics in this section.

### Power

Level  $\mu$  significance tests are closely related to confidence intervals—in fact, we saw that a two-sided test can be carried out directly from a confidence interval. The significance level, like the confidence level, says how reliable the method is in repeated use. If we use 5% significance tests repeatedly when  $H_0$  is in fact true, we will be wrong (the test will reject  $H_0$ ) 5% of the time and right (the test will fail to reject  $H_0$ ) 95% of the time.

The ability of a test to detect that  $H_0$  is false is measured by the probability that the test will reject  $H_0$  when an alternative is true. The higher this probability is, the more sensitive the test is.

### POWER

The probability that a level  $\mu$  significance test will reject  $H_0$  when a particular alternative value of the parameter is true is called the **power** of the test to detect that alternative.

## Example

### 6.29 The power of the TBBMC significance test

Can a six-month exercise program increase the total body bone mineral content (TBBMC) of young women? A team of researchers is planning a study to examine this question. Based on the results of a previous study, they are willing to assume that  $\sigma=2$  for the percent change in TBBMC over the six-month period. They also believe that a change in TBBMC of 1% is important, so they would like to have a reasonable chance of detecting a change this large or larger. Is 25 subjects a large enough sample for this project?

We will answer this question by calculating the power of the significance test that will be used to evaluate the data to be collected. The calculation consists of three steps:

1. State  $H_0$ ,  $H_a$  (the particular alternative we want to detect), and the significance level  $\mu$ .
2. Find the values of  $\bar{x}$  that will lead us to reject  $H_0$ .
3. Calculate the probability of observing these values of  $\bar{x}$  when the alternative is true.

**Step 1.** The null hypothesis is that the exercise program has no effect on TBBMC. In other words, the mean percent change is zero. The alternative is that exercise is beneficial; that is, the mean change is positive. Formally, we have

$$H_0:\mu=0$$

$$H_a:\mu>0$$

The alternative of interest is  $\mu=1\%$  increase in TBBMC. A 5% test of significance will be used.

**Step 2.** The  $z$  test rejects  $H_0$  at the  $\alpha=0.05$  level whenever

$$\bar{x} \geq 1.645225$$

Be sure you understand why we use 1.645. Rewrite this in terms of  $\bar{x}$ :

$$P(\bar{x} \geq 0.658 \text{ when } \mu=1) = P(\bar{x} - \mu\sigma/n \geq 0.658 - 12/25)$$

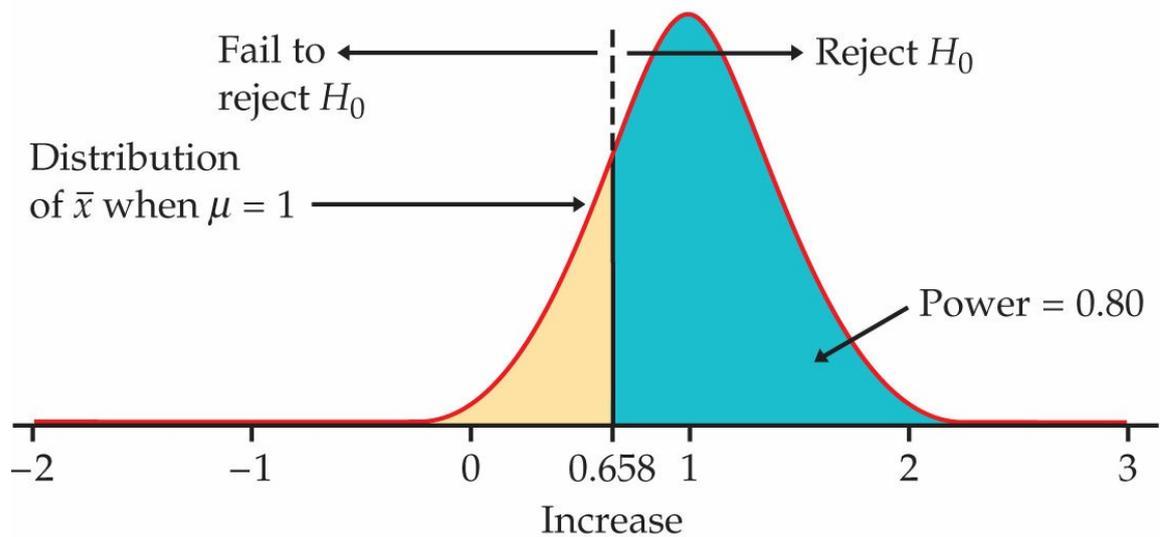
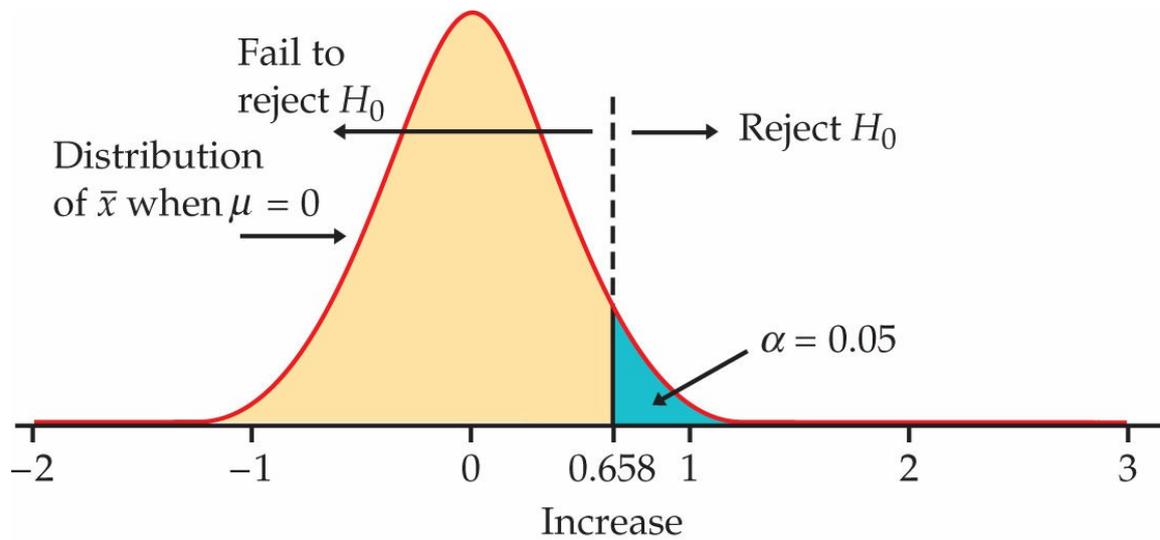
$$\bar{x} \geq 0.658$$

Because the significance level is  $\alpha=0.05$  this event has probability 0.05 of occurring *when the population mean  $\mu$  is 0*.

**Step 3.** The power to detect the alternative  $\mu=1\%$  is the probability that  $H_0$  will be rejected *when in fact  $\mu=1\%$* . We calculate this probability by standardizing  $\bar{x}$ ,

using the value  $\mu=1$ , the population standard deviation  $\sigma=2$ , and the sample size  $n=25$ . The power is

$$z = \frac{\bar{x} - 6.00}{2/\sqrt{25}}$$
$$= P(Z \geq -0.855) = 0.80$$



**Figure 6.16**

The sampling distributions of  $\bar{x}$  when  $\mu = 0$  and when  $\mu = 1$ . The power is the probability that the test rejects  $H_0$  when the alternative is true.

Figure 6.16 illustrates the power with the sampling distribution of  $\bar{x}$  when  $\mu=1$ . This significance test rejects the null hypothesis that exercise has no effect on TBBMC 80% of the time if the true effect of exercise is a 1% increase in TBBMC. If the true effect of exercise is a greater percent increase, the test will have greater power; it will reject with a higher probability.

Here is another example of a power calculation, this time for a two-sided  $z$  test.

## Example

### 6.30 Power of the lead concentration test

Example 6.17 (page 386) presented a test of

$$H_0: \mu = 6.0$$

$$H_a: \mu \neq 6.0$$

at the 1% level of significance. What is the power of this test against the specific alternative  $\mu = 6.5$ ?

The test rejects  $H_0$  when  $|z| \geq 2.576$ . The test statistic is

$$P(\bar{x} \geq 6.37) = P(\bar{x} - \mu\sigma/n \geq 6.37 - 6.500.25/3)$$

Some arithmetic shows that the test rejects when either of the following is true:

$$z \geq 2.576 \quad (\text{in other words, } \bar{x} \geq 6.37)$$

$$z \leq -2.576 \quad (\text{in other words, } \bar{x} \leq 5.63)$$

These are disjoint events, so the power is the sum of their probabilities, *computed assuming that the alternative  $\mu = 6.5$  is true*. We find that

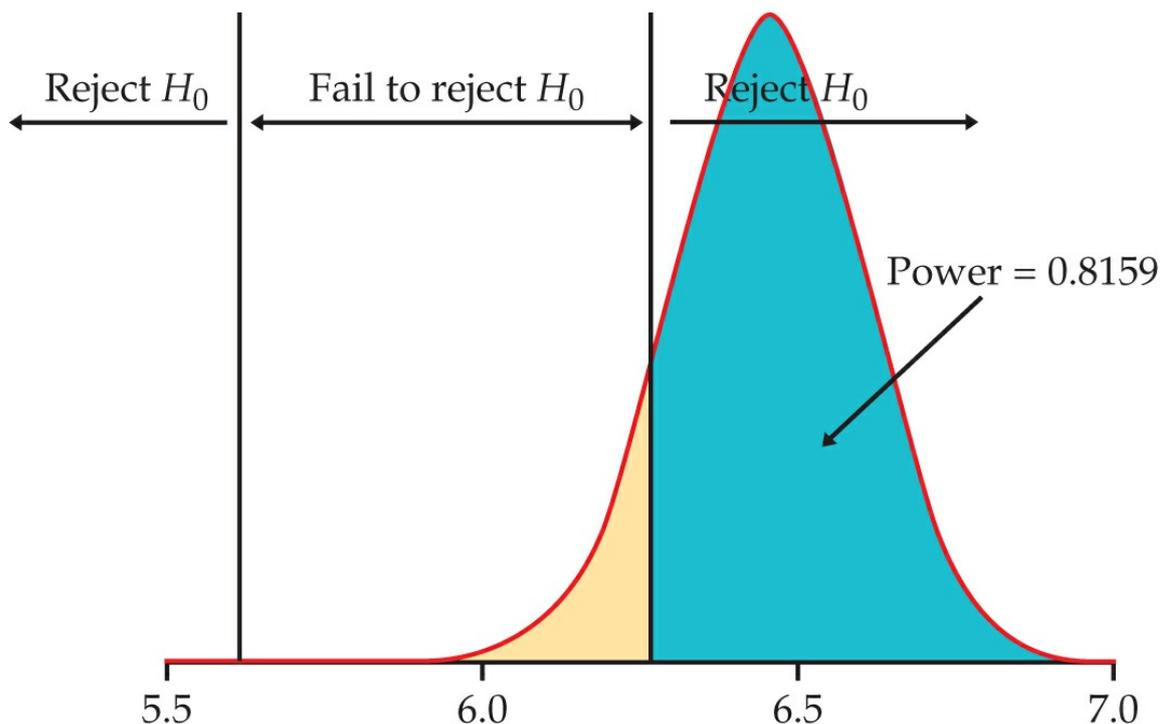
$$P(\bar{x} \geq 5.63) = P(\bar{x} - \mu\sigma/n \geq 5.63 - 6.500.25/3)$$

$$= P(Z \geq -0.90) = 0.8159$$

$$z = \bar{x} - 220.01/5$$

$$= P(Z \leq -6.03) \approx 0$$

Figure 6.17 illustrates this calculation. Because the power is about 0.82, we are quite confident that the test will reject  $H_0$  when this alternative is true.



### Figure 6.17

The power for Example 6.30. Unlike Figure 6.16, only the sampling distribution under the alternative is shown.

High power is desirable. Along with 95% confidence intervals and 5% significance tests, 80% power is becoming a standard. Many U.S. government agencies that provide research funds require that the sample size for the funded studies be sufficient to detect important results 80% of the time using a 5% test of significance.

### Increasing the power

Suppose that you have performed a power calculation and found that the power is too small. What can you do to increase it? Here are four ways:

- Increase  $\alpha$  A 5% test of significance will have a greater chance of rejecting the alternative than a 1% test because the strength of evidence required for rejection is less.
- Consider a particular alternative that is farther away from  $\mu_0$ . Values of  $\mu$  that are in  $H_a$  but lie close to the hypothesized value  $\mu_0$  are harder to detect (lower power) than values of  $\mu$  that are far from  $\mu_0$ .
- Increase the sample size. More data will provide more information about  $\bar{x}$  so we have a better chance of distinguishing values of  $\mu$ .
- Decrease  $\sigma$ . This has the same effect as increasing the sample size: more information about  $\mu$ . Improving the measurement process and restricting attention to a subpopulation are possible ways to decrease  $\sigma$ .

Power calculations are important in planning studies. Using a significance test with low power makes it unlikely that you will find a significant effect even if the truth is far from the null hypothesis. A null hypothesis that is, in fact, false can become widely believed if repeated attempts to find evidence against it fail because of low power. The following example illustrates this point.

### Example

#### 6.31 Are stock markets efficient?

The “efficient market hypothesis” for the time series of stock prices says that future stock prices (when adjusted for inflation) show only random variation.

No information available now will help us predict stock prices in the future, because the efficient working of the market has already incorporated all available information in the present price. Many studies have tested the claim that one or another kind of information is helpful. In these studies, the efficient market hypothesis is  $H_0$ , and the claim that prediction is possible is  $H_a$ . Almost all the studies have failed to find good evidence against  $H_0$ . As a result, the efficient market hypothesis is quite popular. But an examination of the significance tests employed finds that the power is generally low. Failure to reject  $H_0$  when using tests of low power is not evidence that  $H_0$  is true. As one expert says, “The widespread impression that there is strong evidence for market efficiency may be due just to a lack of appreciation of the low power of many statistical tests.”<sup>31</sup>

## Inference as decision

We have presented tests of significance as methods for assessing the strength of evidence against the null hypothesis. This assessment is made by the  $P$ -value, which is a probability computed under the assumption that  $H_0$  is true. The alternative hypothesis (the statement we seek evidence for) enters the test only to help us see what outcomes count against the null hypothesis.

There is another way to think about these issues. Sometimes we are really concerned about making a decision or choosing an action based on our evaluation of the data. **Acceptance sampling** is one such circumstance. A producer of bearings and a skateboard manufacturer agree that each carload lot of bearings shall meet certain quality standards. When a carload arrives, the manufacturer chooses a sample of bearings to be inspected. On the basis of the sample outcome, the manufacturer will either accept or reject the carload. Let’s examine how the idea of inference as a decision changes the reasoning used in tests of significance.

### acceptance sampling

## Two types of error

Tests of significance concentrate on  $H_0$ , the null hypothesis. If a decision is called for, however, there is no reason to single out  $H_0$ . There are simply two hypotheses, and we must accept one and reject the other. It is convenient to call the two hypotheses  $H_0$  and  $H_a$ , but  $H_0$  no longer has the special status (the statement we try to find evidence against) that it had in tests of significance. In the acceptance sampling problem, we must decide between

$H_0$ :the lot of bearings meets standards

$H_a$ :the lot does not meet standards

on the basis of a sample of bearings.

We hope that our decision will be correct, but sometimes it will be wrong. There are two types of incorrect decisions. We can accept a bad lot of bearings, or we can reject a good lot. Accepting a bad lot injures the consumer, while rejecting a good lot hurts the producer. To help distinguish these two types of error, we give them specific names.

## TYPE I AND TYPE II ERRORS

If we reject  $H_0$  (accept  $H_a$ ) when in fact  $H_0$  is true, this is a **Type I error**. If we accept  $H_0$  (reject  $H_a$ ) when in fact  $H_a$  is true, this is a **Type II error**.

		Truth about the population	
		$H_0$ true	$H_a$ true
Decision based on sample	Reject $H_0$	Type I error	Correct decision
	Accept $H_0$	Correct decision	Type II error

**Figure 6.18**

The two types of error in testing hypotheses.

The possibilities are summed up in Figure 6.18. If  $H_0$  is true, our decision either is correct (if we accept  $H_0$ ) or is a Type I error. If  $H_a$  is true, our decision either is correct or is a Type II error. Only one error is possible at one time. Figure 6.19 applies these ideas to the acceptance sampling example.

		Truth about the lot	
		Does meet standards	Does not meet standards
Decision based on sample	Reject the lot	Type I error	Correct decision
	Accept the lot	Correct decision	Type II error

**Figure 6.19**

The two types of error in the acceptance sampling setting.

## Error probabilities

Any rule for making decisions is assessed in terms of the probabilities of the two types of error. This is in keeping with the idea that statistical inference is based on probability. We cannot (short of inspecting the whole lot) guarantee that good lots of bearings will never be rejected and bad lots never be accepted. But by random sampling and the laws of probability, we can say what the probabilities of both kinds of error are.

Significance tests with fixed level  $\alpha$  give a rule for making decisions because the test either rejects  $H_0$  or fails to reject it. If we adopt the decision-making way of thought, failing to reject  $H_0$  means deciding that  $H_0$  is true. We can then describe the performance of a test by the probabilities of Type I and Type II errors.

### Example

#### 6.32 Outer diameter of a skateboard bearing



The mean outer diameter of a skateboard bearing is supposed to be 22.000 millimeters (mm). The outer diameters vary Normally with standard deviation  $\sigma=0.010$  mm. When a lot of the bearings arrives, the skateboard manufacturer takes an SRS of 5 bearings from the lot and measures their outer diameters. The manufacturer rejects the bearings if the sample mean diameter is significantly different from 22 mm at the 5% significance level.

This is a test of the hypotheses

$$H_0: \mu=22$$

$$H_a: \mu \neq 22$$

To carry out the test, the manufacturer computes the  $z$  statistic:

$$z = \frac{\bar{x} - 22}{0.010/\sqrt{5}}$$

and rejects  $H_0$  if

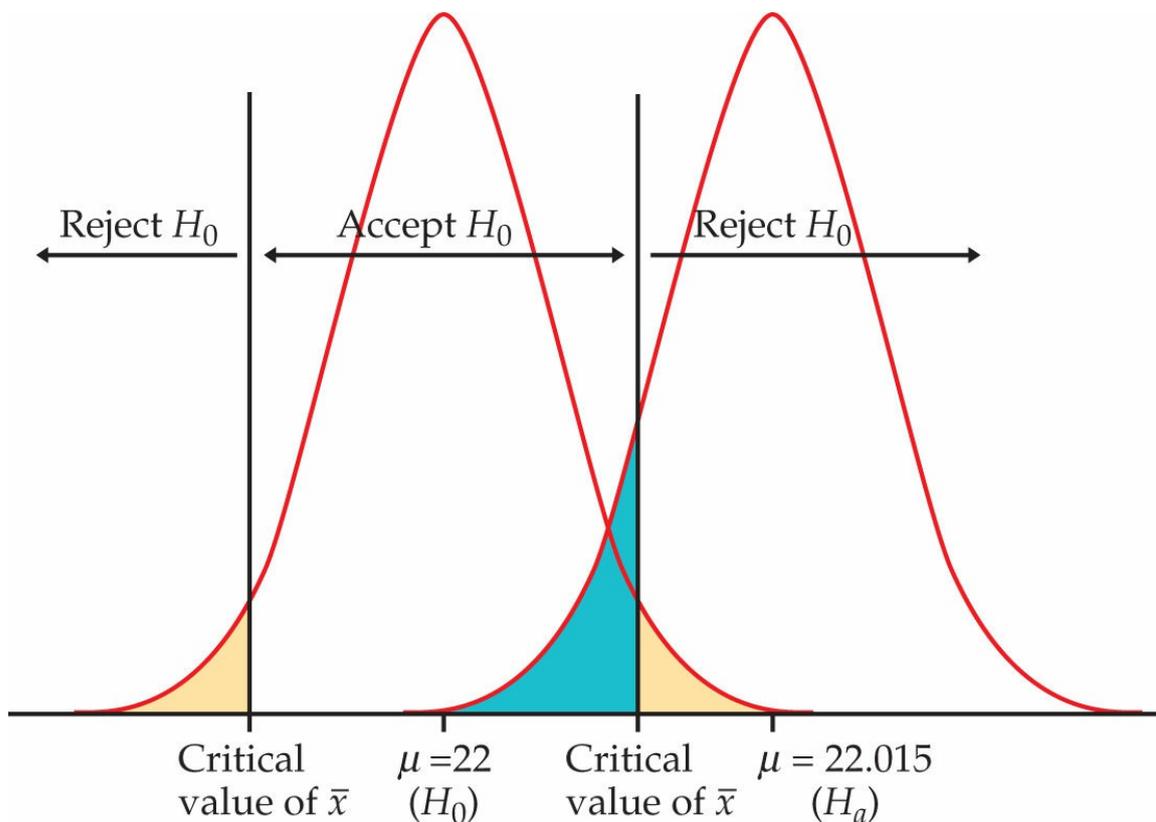
$$z < -1.96 \quad \text{or} \quad z > 1.96$$

A Type I error is to reject  $H_0$  when in fact  $\mu=22$ .

What about Type II errors? Because there are many values of  $\mu$  in  $H_a$ , we will concentrate on one value. The producer and the manufacturer agree that a lot of bearings with mean 0.015 mm away from the desired mean 22.000 should be rejected. So a particular Type II error is to accept  $H_0$  when in fact  $\mu=22.015$ .

Figure 6.20 shows how the two probabilities of error are obtained from the two sampling distributions of  $\bar{x}$  for  $\mu=22$  and for  $\mu=22.015$ . When  $\mu=22$ ,  $H_0$  is true and to reject  $H_0$  is a Type I error. When  $\mu=22.015$ , accepting  $H_0$  is a Type

II error. We will now calculate these error probabilities.



**Figure 6.20**

The two error probabilities for Example 6.32. The probability of a Type I error (yellow area) is the probability of rejecting  $H_0: \mu = 22$  when in fact  $\mu = 22$ . The probability of a Type II error (blue area) is the probability of accepting  $H_0$  when in fact  $\mu = 22.015$ .

The probability of a Type I error is the probability of rejecting  $H_0$  when it is really true. In Example 6.32, this is the probability that  $|z| \geq 1.96$  when  $\mu = 22$ . But this is exactly the significance level of the test. The critical value 1.96 was chosen to make this probability 0.05, so we do not have to compute it again. The definition of “significant at level 0.05” is that sample outcomes this extreme will occur with probability 0.05 when  $H_0$  is true.

## SIGNIFICANCE AND TYPE I ERROR

The significance level  $\alpha$  of any fixed level test is the probability of a Type I error. That is,  $\alpha$  is the probability that the test will reject the null hypothesis  $H_0$  when  $H_a$  is in fact true.

The probability of a Type II error for the particular alternative  $\mu = 22.015$  in Example 6.32 is the probability that the test will fail to reject  $H_0$  when  $\mu$  has this

alternative value. The *power* of the test to detect the alternative  $\mu=22.015$  is just the probability that the test *does* reject  $H_0$ . By following the method of Example 6.30, we can calculate that the power is about 0.92. The probability of a Type II error is therefore  $1-0.92$  or 0.08.

## POWER AND TYPE II ERROR

The power of a fixed level test to detect a particular alternative is 1 minus the probability of a Type II error for that alternative.

The two types of error and their probabilities give another interpretation of the significance level and power of a test. The distinction between tests of significance and tests as rules for deciding between two hypotheses does not lie in the calculations *but in the reasoning that motivates the calculations*. In a test of significance we focus on a single hypothesis ( $H_0$ ) and a single probability (the  $P$ -value). The goal is to measure the strength of the sample evidence against  $H_0$ . Calculations of power are done to check the sensitivity of the test. If we cannot reject  $H_0$ , we conclude only that there is not sufficient evidence against  $H_0$ , not that  $H_0$  is actually true. If the same inference problem is thought of as a decision problem, we focus on two hypotheses and give a rule for deciding between them based on the sample evidence. We therefore must focus equally on two probabilities, the probabilities of the two types of error. We must choose one hypothesis and cannot abstain on grounds of insufficient evidence.

### The common practice of testing hypotheses

Such a clear distinction between the two ways of thinking is helpful for understanding. In practice, the two approaches often merge. We continued to call one of the hypotheses in a decision problem  $H_0$ . The common practice of *testing hypotheses* mixes the reasoning of significance tests and decision rules as follows:

1. State  $H_0$  and  $H_a$  just as in a test of significance.
2. Think of the problem as a decision problem, so that the probabilities of Type I and Type II errors are relevant.
3. Because of Step 1, Type I errors are more serious. So choose an  $\alpha$  (significance level) and consider only tests with probability of a Type I error no greater than  $\alpha$ .
4. Among these tests, select one that makes the probability of a Type II error as small as possible (that is, power as large as possible). If this probability is too large, you will have to take a larger sample to reduce the chance of an error.

Testing hypotheses may seem to be a hybrid approach. It was, historically, the effective beginning of decision-oriented ideas in statistics. An impressive mathematical theory of hypothesis testing was developed between 1928 and 1938 by Jerzy Neyman and Egon Pearson. The decision-making approach came later (1940s). Because decision theory in its pure form leaves you with two error probabilities and no simple rule on how to balance them, it has been used less often than either tests of significance or tests of hypotheses. Decision ideas have been applied in testing problems mainly by way of the Neyman-Pearson hypothesis-testing theory. That theory asks you first to choose  $\alpha$ , and the influence of Fisher has often led users of hypothesis testing comfortably back to  $\alpha=0.05$  or  $\alpha=0.01$ . Fisher, who was exceedingly argumentative, violently attacked the Neyman-Pearson decision-oriented ideas, and the argument still continues.

## Section 6.4 Summary

The **power** of a significance test measures its ability to detect an alternative hypothesis. The power to detect a specific alternative is calculated as the probability that the test will reject  $H_0$  when that alternative is true. This calculation requires knowledge of the sampling distribution of the test statistic under the alternative hypothesis. Increasing the size of the sample increases the power when the significance level remains fixed.

An alternative to significance testing regards  $H_0$  and  $H_a$  as two statements of equal status that we must decide between. This **decision theory** point of view regards statistical inference in general as giving rules for making decisions in the presence of uncertainty.

In the case of testing  $H_0$  versus  $H_a$ , decision analysis chooses a decision rule on the basis of the probabilities of two types of error. A **Type I error** occurs if  $H_0$  is rejected when it is in fact true. A **Type II error** occurs if  $H_0$  is accepted when in fact  $H_a$  is true.

In a fixed level  $\alpha$  significance test, the significance level  $\alpha$  is the probability of a Type I error, and the power to detect a specific alternative is 1 minus the probability of a Type II error for that alternative.

## SECTION 6.4 Exercises

### 6.110 Make a recommendation.

Your manager has asked you to review a research proposal that includes a section on sample size justification. A careful reading of this section indicates that the power is 28% for detecting an effect that would be considered important. Write a short report for your manager explaining what this means and make a recommendation on whether or not this study should be run.

### 6.111 Explain power and sample size.

Two studies are identical in all respects except for the sample sizes. Consider the power versus a particular sample size. Will the study with the larger sample size have more power or less power than the one with the smaller sample size? Explain your answer in terms that could be understood by someone with very little knowledge of statistics.

### 6.112 Power for a different alternative.

The power for a two-sided test of the null hypothesis  $\mu=0$  versus the alternative  $\mu=4$  is 0.83. What is the power versus the alternative  $\mu=-4$ ? Explain your answer.

### 6.113 More on the power for a different alternative.

A one-sided test of the null hypothesis  $\mu=20$  versus the alternative  $\mu=30$  has power equal to 0.6. Will the power for the alternative  $\mu=40$  be higher or lower than 0.6? Draw a picture and use this to explain your answer.

### 6.114 Effect of changing the alternative $\mu$ on power.

The *Statistical Power* applet illustrates a power calculation similar to that in Figure 6.16 (page 404). Open the applet and keep the default settings for the null ( $\mu=0$ ) and the alternative ( $\mu>0$ ) hypotheses, the sample size ( $n=10$ ), the standard deviation ( $\sigma=1$ ), and the significance level ( $\alpha=0.05$ ). In the “alt  $\mu$  =” box enter the value 1. What is the power? Repeat for alternative  $\mu$  equal to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. Make a table giving  $\mu$  and the power. What do you conclude?

### 6.115 Other changes and the effect on power.

Refer to the previous exercise. For each of the following changes, explain what happens to the power for each alternative  $\mu$  in the table.

- (a) Change to the two-sided alternative.
- (b) Increase  $\sigma$  to 2.
- (c) Increase  $n$  from 10 to 20.

### 6.116 Power of the random north–south distribution of trees test.

In Exercise 6.68 (page 392) you performed a two-sided significance test of the null hypothesis that the average north–south location of the longleaf pine trees sampled in the Wade Tract was  $\mu = 100$ . There were 584 trees in the sample and the standard deviation was assumed to be 58. The sample mean in that analysis was  $\bar{x}=99.74$ . Use the *Statistical Power* applet to compute the power for the alternative  $\mu=99$  using a two-sided test at the 5% level of significance.

### 6.117 Power of the random east–west distribution of trees test.

Refer to the previous exercise. Note that in the east–west direction, the average location was 113.8. Use the *Statistical Power* applet to find the power for the alternative  $\mu = 110$ .

### 6.118 Planning another test to compare consumption.

Example 6.15 (page 383) gives a test of a hypothesis about the mean consumption of sugar- sweetened beverages at your university based on a sample of size  $n=100$ . The hypotheses are

$$H_0: \mu=286$$

$$H_a: \mu \neq 286$$

While the result was not statistically significant, it did provide some evidence that the mean was smaller than 286. Thus, you plan to recruit another sample of students from your university but this time use a one-sided alternative. You were thinking of surveying  $n=100$  students but now wonder if this sample size gives adequate power to detect a decrease of 15 calories per day to  $\mu=271$ .

- Given  $\alpha=0.05$ , for what values of  $z$  will you reject the null hypothesis?
- Using  $\sigma=155$  and  $\mu=286$  for what values of  $\bar{x}$  will you reject  $H_0$ ?
- Using  $\sigma=155$  and  $\mu=271$ , what is the probability that  $\bar{x}$  will fall in the region defined in part (b)?
- Will a sample size of  $n=100$  give you adequate power? Or do you need to find ways to increase the power? Explain your answer.
- Use the *Statistical Power* applet to determine the sample size  $n$  that gives you power near 0.80.

### 6.119 Power of the mean SATM score test.

Example 6.16 (page 384) gives a test of a hypothesis about the SATM scores of California high school students based on an SRS of 500 students. The hypotheses are

$$H_0: \mu=475$$

$$H_a: \mu > 475$$

Assume that the population standard deviation is  $\sigma=100$ . The test rejects  $H_0$  at the 1% level of significance when  $z \geq 2.326$  where

Is this test sufficiently sensitive to usually detect an increase of 10 points in the population mean SATM score? Answer this question by calculating the power of the test to detect the alternative  $\mu=485$ .



### 6.120 Choose the appropriate distribution.

You must decide which of two discrete distributions a random variable  $X$  has. We will call the distributions  $p_0$  and  $p_1$ . Here are the probabilities they assign to the values  $x$  of  $X$ :

$x$	0	1	2	3	4	5	6
$p_0$	0.1	0.1	0.2	0.1	0.1	0.1	0.3
$p_1$	0.2	0.2	0.2	0.1	0.1	0.1	0.1

You have a single observation on  $X$  and wish to test

$$H_0: p_0 \text{ is correct}$$

$$H_a: p_1 \text{ is correct}$$

One possible decision procedure is to reject  $H_0$  only if  $X \leq 2$ .

- Find the probability of a Type I error, that is, the probability that you reject  $H_0$  when  $p_0$  is the correct distribution.

(b) Find the probability of a Type II error.

### **6.121 Computer-assisted career guidance systems.**

A wide variety of computer-assisted career guidance systems have been developed over the last decade. These programs use factors such as student interests, aptitude, skills, personality, and family history to recommend a career path. For simplicity, suppose that a program recommends a high school graduate either to go to college or to join the workforce.

(a) What are the two hypotheses and the two types of error that the program can make?

(b) The program can be adjusted to decrease one error probability at the cost of an increase in the other error probability. Which error probability would you choose to make smaller, and why? (This is a matter of judgment. There is no single correct answer.)

## CHAPTER 6 Exercises

### 6.122 Telemarketing wages.

An advertisement in the student newspaper asks you to consider working for a telemarketing company. The ad states, “Earn between \$500 and \$1000 per week.” Do you think that the ad is describing a confidence interval? Explain your answer.

### 6.123 Exercise and statistics exams.

A study examined whether light exercise performed an hour before the final exam in statistics affects how students perform on the exam. The  $P$ -value was given as 0.27.

- (a) State null and alternative hypotheses that could be used for this study. (*Note:* There is more than one correct answer.)
- (b) Do you reject the null hypothesis? State your conclusion in plain language.
- (c) What other facts about the study would you like to know for a proper interpretation of the results?

### 6.124 Stress by occupation.

As part of a study on the impact of job stress on smoking, researchers used data from the Health and Retirement Study (HRS) to collect information on 3825 ever-smoker individuals who were 50 to 64 years of age.<sup>32</sup> An ever-smoker is someone who was a smoker at some time in his or her life. The HRS is a biennial survey, thus providing the researchers with 17,043 person-year observations. One of the questions on the survey asked a participant how much he or she agrees or disagrees with the statement “My job involves a lot of stress.” The answers were coded as a 1 if a participant “strongly agreed” and 0 otherwise. The following table summarizes these responses by occupation.

Occupation	$\hat{p}$	$n$
Professional	0.23	2447
Managerial	0.22	2552
Administrative	0.17	2309
Sales	0.15	1811
Mechanical	0.12	1979
Service	0.13	2592
Operator	0.12	2782
Farm	0.08	571

- (a) Because these responses are binary, use the formula for the standard deviation of a sample proportion (page 330) and construct 95% confidence intervals for each occupation.
- (b) Summarize the results. Do there appear to be certain groups of occupations with similar stress levels?

(c) A friend questions the use of the standard deviation formula in part (a). Refer back to the binomial setting (page 322). What might your friend be concerned with?

### 6.125 Food selection by children in school cafeterias.

A group of researchers examined whether children’s food selection in a school cafeteria met the standards set by the School Meals Initiative. They measured food selection and food intake of 2049 fourth- through sixth-grade students in 33 schools over a 3-day period using digital photography. The following table summarizes some of the food intake measurements.<sup>33</sup>

	Boys		Girls	
	$n = 852$		$n = 1197$	
Food intake	Mean	St. Dev.	Mean	St. Dev.
Energy (kilojoules)	2448	717	2170	693
Protein (g)	24.5	7.5	22.1	7.7
Calcium (mg)	324.1	130.6	265.0	128.9

Given the large sample sizes, we can assume that the sample standard deviations are the population standard deviations.

- (a) Compute 95% confidence intervals for all three intake measures for the boys.
- (b) Compute 95% confidence intervals for all three intake measures for the girls.
- (c) In the next chapter, we will describe the confidence interval for the difference between two means. For now, let’s compare the boy and girl confidence intervals for each food intake measure. Do you think these pairs of intervals provide strong evidence against the null hypothesis that the boys and girls consume, on average, the same amount? Explain your answer.

### 6.126 Coverage percent of 95% confidence interval.

For this exercise you will use the *Confidence Interval* applet. Set the confidence level at 95% and click the “Sample” button 10 times to simulate 10 confidence intervals. Record the percent hit. Simulate another 10 intervals by clicking another 10 times (do not click the “Reset” button). Record the percent hit for your 20 intervals. Repeat the process of simulating 10 additional intervals and recording the results until you have a total of 200 intervals. Plot your results and write a summary of what you have found.



### 6.127 Coverage percent of 90% confidence interval.

Refer to the previous exercise. Do the simulations and report the results for 90% confidence.



### 6.128 Effect of sample size on significance.

You are testing the null hypothesis that  $\mu=0$  versus the alternative  $\mu>0$  using  $\alpha=0.05$ . Assume that  $\sigma=14$ . Suppose that  $\bar{x}=4$  and  $n=10$ . Calculate the test statistic and its  $P$ -value. Repeat, assuming the same value of  $\bar{x}$  but with  $n=20$ . Do the same for sample sizes of 30, 40, and 50. Plot the values of the test statistic versus the sample size. Do the same for the  $P$ -values. Summarize what this demonstration shows about the effect of the sample size on significance testing.

### 6.129 Blood phosphorus level in dialysis patients.

Patients with chronic kidney failure may be treated by dialysis, in which a machine removes toxic wastes from the blood, a function normally performed by the kidneys. Kidney failure and dialysis can cause other changes, such as retention of phosphorus, that must be corrected by changes in diet. A study of the nutrition of dialysis patients measured the level of phosphorus in the blood of several patients on six occasions. Here are the data for one patient (in milligrams of phosphorus per deciliter of blood):<sup>34</sup>

5.4 5.2 4.5 4.9 5.7 6.3

The measurements are separated in time and can be considered an SRS of the patient's blood phosphorus level. Assume that this level varies Normally with  $\sigma=0.9$  mg/dl.  PMGDL

- (a) Give a 95% confidence interval for the mean blood phosphorus level.
- (b) The normal range of phosphorus in the blood is considered to be 2.6 to 4.8 mg/dl. Is there strong evidence that this patient has a mean phosphorus level that exceeds 4.8?

### 6.130 Cellulose content in alfalfa hay.

An agronomist examines the cellulose content of a variety of alfalfa hay. Suppose that the cellulose content in the population has standard deviation  $\sigma=8$  milligrams per gram (mg/g). A sample of 15 cuttings has mean cellulose content  $\bar{x}=145$  mg/g.

- (a) Give a 90% confidence interval for the mean cellulose content in the population.
- (b) A previous study claimed that the mean cellulose content was  $\mu=140$  mg/g, but the agronomist believes that the mean is higher than that figure. State  $H_0$  and  $H_a$  and carry out a significance test to see if the new data support this belief.
- (c) The statistical procedures used in parts (a) and (b) are valid when several assumptions are met. What are these assumptions?

### 6.131 Odor threshold of future wine experts.

Many food products contain small quantities of substances that would give an undesirable taste or smell if they are present in large amounts. An example is the "off-odors" caused by sulfur compounds in wine. Oenologists (wine experts) have determined the odor threshold, the lowest concentration of a compound that the human nose can detect. For example, the odor threshold for dimethyl sulfide (DMS) is given in the oenology literature as 25 micrograms per liter of wine ( $\mu\text{g/l}$ ). Untrained noses may be less sensitive, however. Here are the DMS odor thresholds for 10 beginning students of oenology:

31 31 43 36 23 34 32 30 20 24

Assume (this is not realistic) that the standard deviation of the odor threshold for untrained noses is known to be  $\sigma=7$   $\mu\text{g/l}$ .  ODOR

- (a) Make a stemplot to verify that the distribution is roughly symmetric with no outliers. (A Normal quantile plot confirms that there are no systematic departures from Normality.)
- (b) Give a 95% confidence interval for the mean DMS odor threshold among all beginning oenology students.

(c) Are you convinced that the mean odor threshold for beginning students is higher than the published threshold,  $25 \mu\text{g/l}$ ? Carry out a significance test to justify your answer.

### 6.132 Where do you buy?

Consumers can purchase nonprescription medications at food stores, mass merchandise stores such as Target and Wal-Mart, or pharmacies. About 45% of consumers make such purchases at pharmacies. What accounts for the popularity of pharmacies, which often charge higher prices?

A study examined consumers' perceptions of overall performance of the three types of stores, using a long questionnaire that asked about such things as "neat and attractive store," "knowledgeable staff," and "assistance in choosing among various types of nonprescription medication." A performance score was based on 27 such questions. The subjects were 201 people chosen at random from the Indianapolis telephone directory. Here are the means and standard deviations of the performance scores for the sample:<sup>35</sup>

Store type	$\bar{x}$	$s$
Food stores	18.67	24.95
Mass merchandisers	32.38	33.37
Pharmacies	48.60	35.62

We do not know the population standard deviations, but a sample standard deviation  $s$  from so large a sample is usually close to  $\sigma$ . Use  $s$  in place of the unknown  $\sigma$  in this exercise.

- What population do you think the authors of the study want to draw conclusions about? What population are you certain they can draw conclusions about?
- Give 95% confidence intervals for the mean performance for each type of store.
- Based on these confidence intervals, are you convinced that consumers think that pharmacies offer higher performance than the other types of stores? (In Chapter 12, we will study a statistical method for comparing the means of several groups.)

### 6.133 CEO pay.

A study of the pay of corporate chief executive officers (CEOs) examined the increase in cash compensation of the CEOs of 104 companies, adjusted for inflation, in a recent year. The mean increase in real compensation was  $\bar{x}=6.9\%$ , and the standard deviation of the increases was  $s=55\%$ . Is this good evidence that the mean real compensation  $\mu$  of all CEOs increased that year? The hypotheses are

$$H_0: \mu=0 \text{ (no increase)}$$

$$H_a: \mu>0 \text{ (an increase)}$$

Because the sample size is large, the sample  $s$  is close to the population  $\sigma$ , so take  $\sigma=55\%$ .

- Sketch the Normal curve for the sampling distribution of  $\bar{x}$  when  $H_0$  is true. Shade the area that represents the  $P$ -value for the observed outcome  $\bar{x}=6.9\%$ .
- Calculate the  $P$ -value.
- Is the result significant at the  $\alpha=0.05$  level? Do you think the study gives strong evidence that the mean compensation of all CEOs went up?

### 6.134 Meaning of “statistically significant.”

When asked to explain the meaning of “statistically significant at the  $\alpha=0.01$  level,” a student says, “This means there is only probability 0.01 that the null hypothesis is true.” Is this an essentially correct explanation of statistical significance? Explain your answer.

### 6.135 More on the meaning of “statistically significant.”

Another student, when asked why statistical significance appears so often in research reports, says, “Because saying that results are significant tells us that they cannot easily be explained by chance variation alone.” Do you think that this statement is essentially correct? Explain your answer.

### 6.136 Roulette.

A roulette wheel has 18 red slots among its 38 slots. You observe many spins and record the number of times that red occurs. Now you want to use these data to test whether the probability of a red has the value that is correct for a fair roulette wheel. State the hypotheses  $H_0$  and  $H_a$  that you will test.

### 6.137 Simulation study of the confidence interval.

Use a computer to generate  $n=12$  observations from a Normal distribution with mean 25 and standard deviation 4:  $N(25,4)$ . Find the 95% confidence interval for  $\mu$ . Repeat this process 100 times and then count the number of times that the confidence interval includes the value  $\mu=25$ . Explain your results.

### 6.138 Simulation study of a test of significance.

Use a computer to generate  $n = 12$  observations from a Normal distribution with mean 25 and standard deviation 4:  $N(25,4)$ . Test the null hypothesis that  $\mu = 25$  using a two-sided significance test. Repeat this process 100 times and then count the number of times that you reject  $H_0$ . Explain your results.

### 6.139 Another simulation study of a test of significance.

Use the same procedure for generating data as in the previous exercise. Now test the null hypothesis that  $\mu = 23$ . Explain your results.

### 6.140 The handshake during employment interviews.

Nonverbal cues, such as eye contact and smiling, have been shown to positively influence the assessment of an interview. Because a firm handshake is often viewed as a sign of confidence and strength, it is thought that it may also influence the assessment. To investigate this, some researchers recruited 98 undergraduate students enrolled in an elective, one-credit career preparation course and had them participate in a mock interview.<sup>36</sup> The following table of means, broken down by gender, summarizes the interviewer’s impression of a series of characteristics associated with the interview. These impressions were all rated on a 1-to-5 scale with 1 representing “weak” and 5 representing “strong.” There were 50 females and 48 males in the study.

Characteristic	Men	Women
Conscientiousness	3.80	3.88
Extraversion	3.88	3.79
Agreeableness	3.72	3.94
Emotional stability	3.58	3.43
Openness to experience	3.50	3.45
Overall handshake	3.70	3.47
Handshake strength	3.64	3.11
Handshake vigor	3.42	3.25
Handshake grip	3.89	3.51
Handshake duration	3.65	3.50
Eye contact	3.90	3.96
Professional dress	4.33	4.53
Interviewer assessment	3.72	3.83

For each characteristic, compute the  $z$  statistic and the associated  $P$ -value for the comparison between the two groups. For all characteristics but the overall interviewer assessment, assume that the standard deviation of the difference is 0.10, so  $z$  is simply the difference in the means divided by 0.10. For interviewer assessment, assume that the standard deviation of the difference is 0.19. Note that you are performing 13 significance tests in this exercise. Keep this in mind when you interpret your results. Write a report summarizing your work.



### 6.141 Find published studies with confidence intervals.

Search the Internet or some journals that report research in your field and find two reports that provide an estimate with a margin of error or a confidence interval. For each report,

- describe the method used to collect the data.
- describe the variable being studied.
- give the estimate and the confidence interval.
- describe any practical difficulties that may have led to errors in addition to the sampling errors quantified by the margin of error.

# 7 Inference for Distributions

CHAPTER



- 7.1 Inference for the Mean of a Population**
- 7.2 Comparing Two Means**
- 7.3 Other Topics in Comparing Distributions**

## Introduction

We began our study of data analysis in Chapter 1 by learning graphical and numerical tools for describing the distribution of a single variable and for comparing several distributions. Our study of the practice of statistical inference begins in the same way, with inference about a single distribution and comparison of two distributions. Comparing more than two distributions requires more elaborate methods, which are presented in Chapters 12 and 13.

Two important aspects of any distribution are its center and spread. If the distribution is Normal, we describe its center by the mean  $\mu$  and its spread by the standard deviation  $\sigma$ . In this chapter, we will meet confidence intervals and significance tests for inference about a population mean  $\mu$  and the difference between two population means  $\mu_1 - \mu_2$ . The previous chapter emphasized the reasoning of significance tests and confidence intervals. Now we emphasize statistical practice, so we no longer assume that population standard deviations are known. This means that we move away from the standard Normal sampling distribution to a new family of sampling distributions. The  $t$  procedures for inference about means are among the most commonly used statistical methods.

## 7.1 Inference for the Mean of a Population

When you complete this section, you will be able to

- Distinguish the standard deviation of the sample mean from the standard error of the sample mean.
- Describe a level  $C$  confidence interval for the population mean in terms of an estimate and its margin of error.
- Construct a level  $C$  confidence interval for  $\mu$  from an SRS of size  $n$  from a large population.
- Perform a one-sample  $t$  significance test and summarize the results.
- Identify when the matched pairs  $t$  procedures should be used instead of two-sample  $t$  procedures.
- Explain when  $t$  procedures can be useful for non-Normal data.

Both confidence intervals and tests of significance for the mean  $\mu$  of a Normal population are based on the sample mean  $\bar{x}$ , which estimates the unknown  $\mu$ . The sampling distribution of  $\bar{x}$  depends on  $\sigma$ . This fact causes no difficulty when  $\sigma$  is known. When  $\sigma$  is unknown, however, we must estimate  $\sigma$  even though we are primarily interested in  $\mu$ . The sample standard deviation  $s$  is used to estimate the population standard deviation  $\sigma$ .



sampling distribution of  $\bar{x}$ , p. 307

### The $t$ distributions

Suppose that we have a simple random sample (SRS) of size  $n$  from a Normally distributed population with mean  $\mu$  and standard deviation  $\sigma$ . The sample mean  $\bar{x}$  is then Normally distributed with mean  $\mu$  and standard deviation  $\sigma/n$ . When  $\sigma$  is not known, we estimate it with the sample standard deviation  $s$  and then we estimate the standard deviation of  $\bar{x}$  by  $s/n$ . This quantity is called the *standard error* of the sample mean  $\bar{x}$  and we denote it by  $SE_{\bar{x}}$ .

### STANDARD ERROR

When the standard deviation of a statistic is estimated from the data, the result is called the **standard error** of the statistic. The standard error of the sample mean is

$$SE_{\bar{x}} = s/n$$

The term “standard error” is sometimes used for the actual standard deviation of a statistic. The estimated value is then called the “estimated standard error.” In this book we will use the term “standard error” only when the standard deviation of a statistic is estimated from the data. The term has this meaning in the output of many statistical computer packages and in research reports that apply statistical methods.

The standardized sample mean, or one-sample  $z$  statistic,

$$z = \frac{\bar{x} - \mu}{\sigma/n}$$

is the basis of the  $z$  procedures for inference about  $\mu$  when  $\sigma$  is known. This statistic has the standard Normal distribution  $N(0, 1)$ . When we substitute the standard error  $s/n$  for the standard deviation  $\sigma/n$  of  $x$  the statistic does *not* have a Normal distribution. It has a distribution that is new to us, called a *t distribution*.

## THE $t$ DISTRIBUTIONS

Suppose that an SRS of size  $n$  is drawn from an  $N(\mu, \sigma)$  population. Then the **one-sample  $t$  statistic**

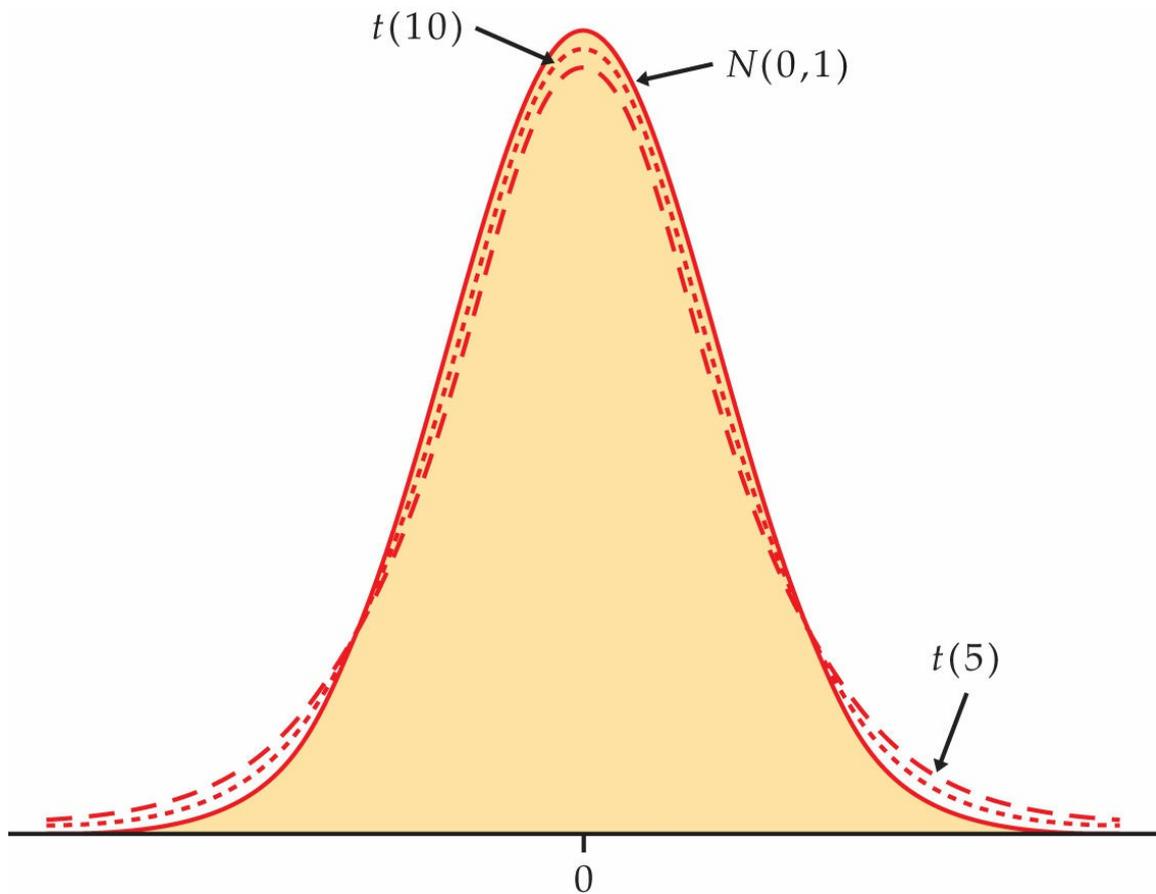
$$t = \frac{\bar{x} - \mu}{s/n}$$

has the  **$t$  distribution** with  $n - 1$  **degrees of freedom**.



degrees of freedom, p. 44

A particular  $t$  distribution is specified by giving the *degrees of freedom*. We use  $t(k)$  to stand for the  $t$  distribution with  $k$  degrees of freedom. The degrees of freedom for this  $t$  statistic come from the sample standard deviation  $s$  in the denominator of  $t$ . We showed earlier that  $s$  has  $n - 1$  degrees of freedom. Thus, there is a different  $t$  distribution for each sample size. There are also other  $t$  statistics with different degrees of freedom, some of which we will meet later in this chapter.



**FIGURE 7.1**

Density curves for the standard Normal,  $t(10)$ , and  $t(5)$  distributions. All are symmetric with center 0. The  $t$  distributions have more probability in the tails than the standard Normal distribution.

The density curves of the  $t(k)$  distributions are similar in shape to the standard Normal curve. That is, they are symmetric about 0 and are bell-shaped. Figure 7.1 compares the density curves of the standard Normal distribution and the  $t$  distributions with 5 and 10 degrees of freedom. The similarity in shape is apparent, as is the fact that the  $t$  distributions have more probability in the tails and less in the center.

In reference to the standardized sample mean, this greater spread is due to the extra variability caused by substituting the random variable  $s$  for the fixed parameter  $\sigma$ . Figure 7.1 also shows that as the degrees of freedom  $k$  increase, the  $t(k)$  density curve gets closer to the  $N(0, 1)$  curve. This reflects the fact that  $s$  will be closer to  $\sigma$  as the sample size increases.

The  $t$  distributions were discovered in 1908 by William S. Gosset. Gosset was a statistician employed by the Guinness brewing company, which prohibited its employees from publishing their discoveries that were brewing related. In this case, the company let him publish under the pen name “Student” using an example that did not involve brewing. The  $t$  distribution is often called “Student’s  $t$ ” in his honor.

Table D in the back of the book gives critical values  $t^*$  for the  $t$  distributions.

For convenience, we have labeled the table entries both by the value of  $p$  needed for significance tests and by the confidence level  $C$  (in percent) required for confidence intervals. The standard Normal critical values are in the bottom row of entries and labeled  $z^*$ . As in the case of the Normal table (Table A), computer software often makes Table D unnecessary.

## USE YOUR KNOWLEDGE

### 7.1 Apartment rates

You randomly choose 16 unfurnished one-bedroom apartments from a large number of advertisements in your local newspaper. You calculate that their mean monthly rent is \$600 and their standard deviation is \$55.

- (a) What is the standard error of the mean?
- (b) What are the degrees of freedom for a one-sample  $t$  statistic?

### 7.2 90% versus 95% confidence interval?

Refer to the previous exercise. You plan to construct a confidence interval for the average monthly rent of unfurnished one-bedroom apartments in your area. If you were to use 90% confidence, rather than 95% confidence, would the margin of error be larger or smaller? Does your answer depend on sample size? Explain your answer.

## The one-sample $t$ confidence interval



confidence interval, p. 356

With the  $t$  distributions to help us, we can now analyze a sample from a Normal population with unknown  $\sigma$  or a large sample from a non-Normal population with unknown  $\sigma$ . The one-sample  $t$  confidence interval is similar in both reasoning and computational detail to the  $z$  confidence interval of Chapter 6. There, the margin of error for the population mean was  $z^*\sigma/n$ . Here, we replace  $\sigma$  by its estimate  $s$  and  $z^*$  by  $t^*$ . This means that the margin of error for the population mean when we use the data to estimate  $\sigma$  is  $t^*s/n$ .

## THE ONE-SAMPLE $t$ CONFIDENCE INTERVAL

Suppose that an SRS of size  $n$  is drawn from a population having unknown mean  $\mu$ . A level  $C$  **confidence interval** for  $\mu$  is

$$\bar{x} \pm t^*sn$$

where  $t^*$  is the value for the  $t(n - 1)$  density curve with area  $C$  between  $-t^*$  and  $t^*$ . The quantity

$$t^*sn$$

is the **margin of error**. The confidence level is exactly  $C$  when the population distribution is Normal and is approximately correct for large  $n$  in other cases.

### Example

#### 7.1 Watching videos on a cell phone

The Nielsen Company is a global information and media company and one of the leading suppliers of media information. In their state-of-the-media report, they announced that U.S. cell phone subscribers average 5.4 hours per month watching videos on their phones.<sup>1</sup> We decide to construct a 95% confidence interval for the average time (hours per month) spent watching videos on cell phones among U.S. college students. We draw the following SRS of size 8 from this population:

11.9 2.8 3.0 6.2 4.7 9.8 11.1 7.8

The sample mean is

$$\bar{x} = 11.9 + 2.8 + \dots + 7.8 / 8 = 7.16$$

and the standard deviation is

$$s = \sqrt{(11.9 - 7.16)^2 + (2.8 - 7.16)^2 + \dots + (7.8 - 7.16)^2 / (8 - 1)} = 3.56$$

with degrees of freedom  $n - 1 = 7$ . The standard error is

$$SE_{\bar{x}} = s / \sqrt{n} = 3.56 / \sqrt{8} = 1.26$$

From Table D we find  $t^* = 2.365$ . The 95% confidence interval is

$$\bar{x} \pm t^*sn = 7.16 \pm 2.365(3.56)$$

$$=7.16 \pm (2.365)(1.26)$$

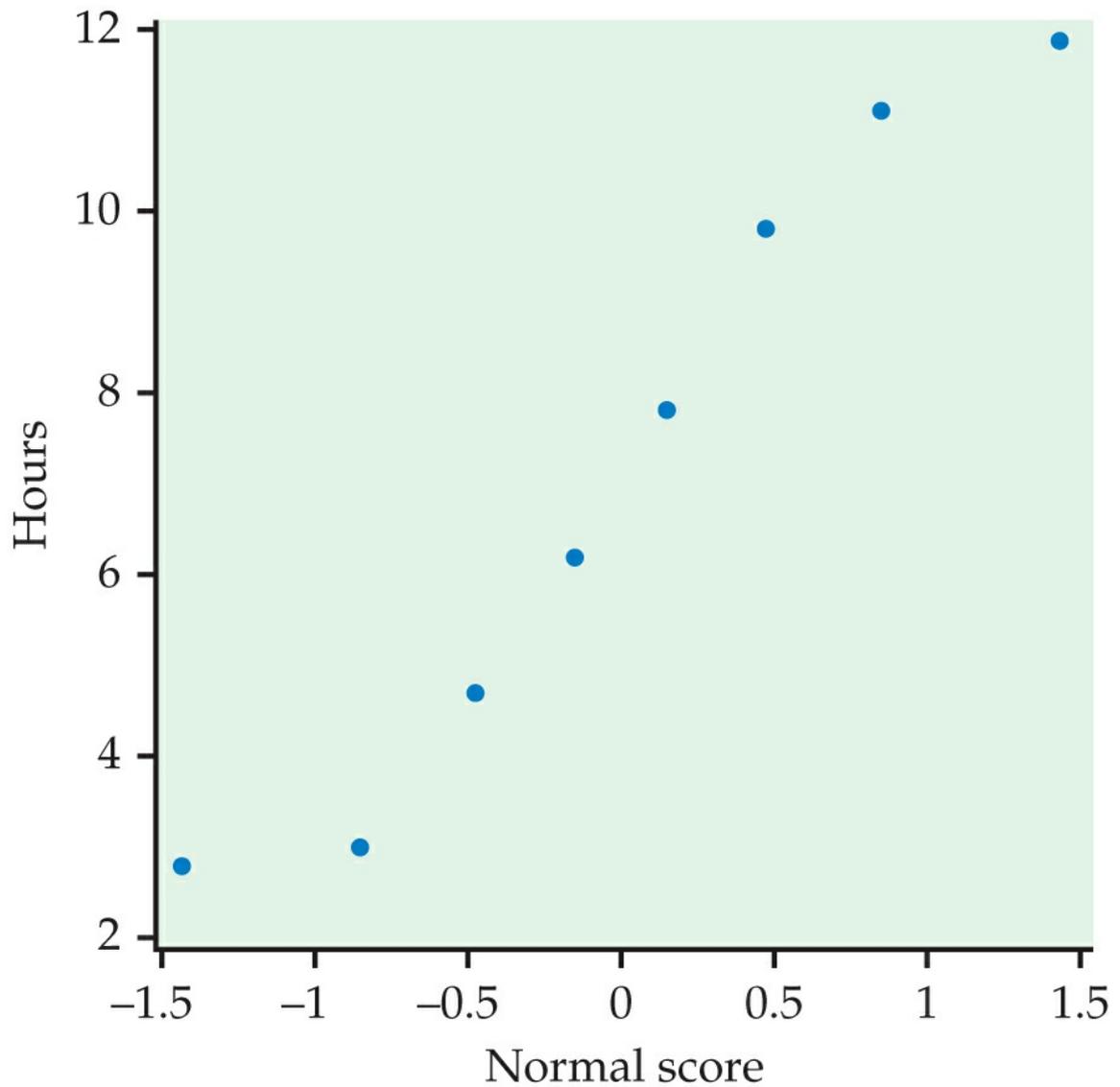
$$=7.16 \pm 2.98$$

$$=(4.2, 10.1)$$

We are 95% confident that among U.S. college students the average time spent watching videos on a cell phone is between 4.2 and 10.1 hours per month.

In this example we have given the actual interval (4.2, 10.1) hours per month as our answer. Sometimes we prefer to report the mean and margin of error: the mean time is 7.2 hours per month with a margin of error of 3.0 hours per month.

Valid interpretation of the  $t$  confidence interval in Example 7.1 rests on assumptions that appear reasonable here. First, we assume that our random sample is an SRS from the U.S. population of college student cell phone users. Second, we assume that the distribution of watching times is Normal. Figure 7.2 shows the Normal quantile plot. With only 8 observations, this assumption cannot be effectively checked. In fact, because a watching time cannot be negative, we might expect this distribution to be skewed to the right. With these data, however, there are no extreme outliers to suggest a severe departure from Normality.



**FIGURE 7.2**

Normal quantile plot of data in Example 7.1.

## USE YOUR KNOWLEDGE

### 7.3 More on apartment rents

Refer to Exercise 7.1 (page 420). Construct a 95% confidence interval for the mean monthly rent of all advertised one-bedroom apartments.

### 7.4 Finding critical $t^*$ -values

What critical value  $t^*$  from Table D should be used to construct

- (a) a 95% confidence interval when  $n = 12$ ?
- (b) a 99% confidence interval when  $n = 38$ ?
- (c) a 90% confidence interval when  $n = 81$ ?

## The one-sample $t$ test



significance test, p. 372

Significance tests using the standard error are also very similar to the  $z$  test that we studied in the last chapter.

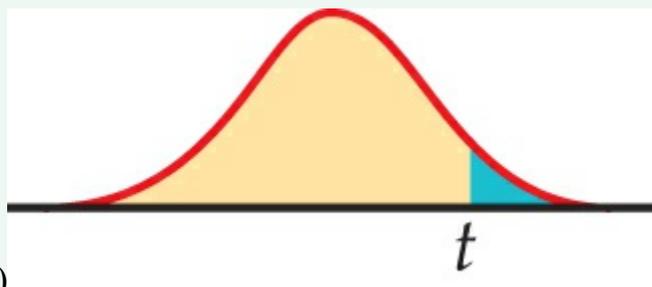
### THE ONE-SAMPLE $t$ TEST

Suppose that an SRS of size  $n$  is drawn from a population having unknown mean  $\mu$ . To test the hypothesis  $H_0 = \mu = \mu_0$  based on an SRS of size  $n$ , compute the **one-sample  $t$  statistic**

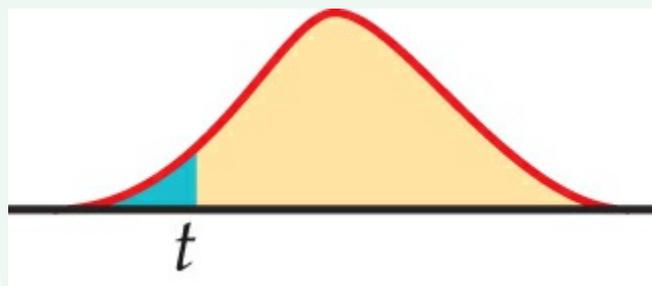
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

In terms of a random variable  $T$  having the  $t(n - 1)$  distribution, the  $p$ -value for a test of  $H_0$  against

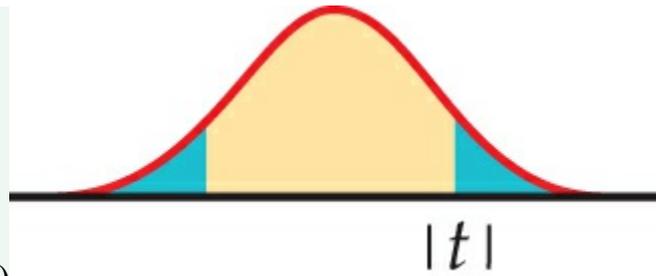
$H_a: \mu > \mu_0$  is  $P(T \geq t)$



$H_a: \mu < \mu_0$  is  $P(T \leq t)$



$$H_a: \mu \neq \mu_0 \text{ is } 2P(T \geq |t|)$$



These  $P$ -values are exact if the population distribution is Normal and are approximately correct for large  $N$  in other cases.

## Example

### 7.2 Significance test for watching videos on a cell phone

We want to test whether the average time that U.S. college students spend watching videos on their phones differs from the reported overall U.S. average at the 0.05 significance level. Specifically, we want to test

$$H_0: \mu = 5.4$$

$$H_a: \mu \neq 5.4$$

Recall that  $n = 8$ ,  $\bar{x} = 7.16$  and  $s = 3.56$ . The  $t$  test statistic is

$$\begin{aligned} t &= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{7.16 - 5.4}{3.56/\sqrt{8}} \\ &= 1.40 \end{aligned}$$

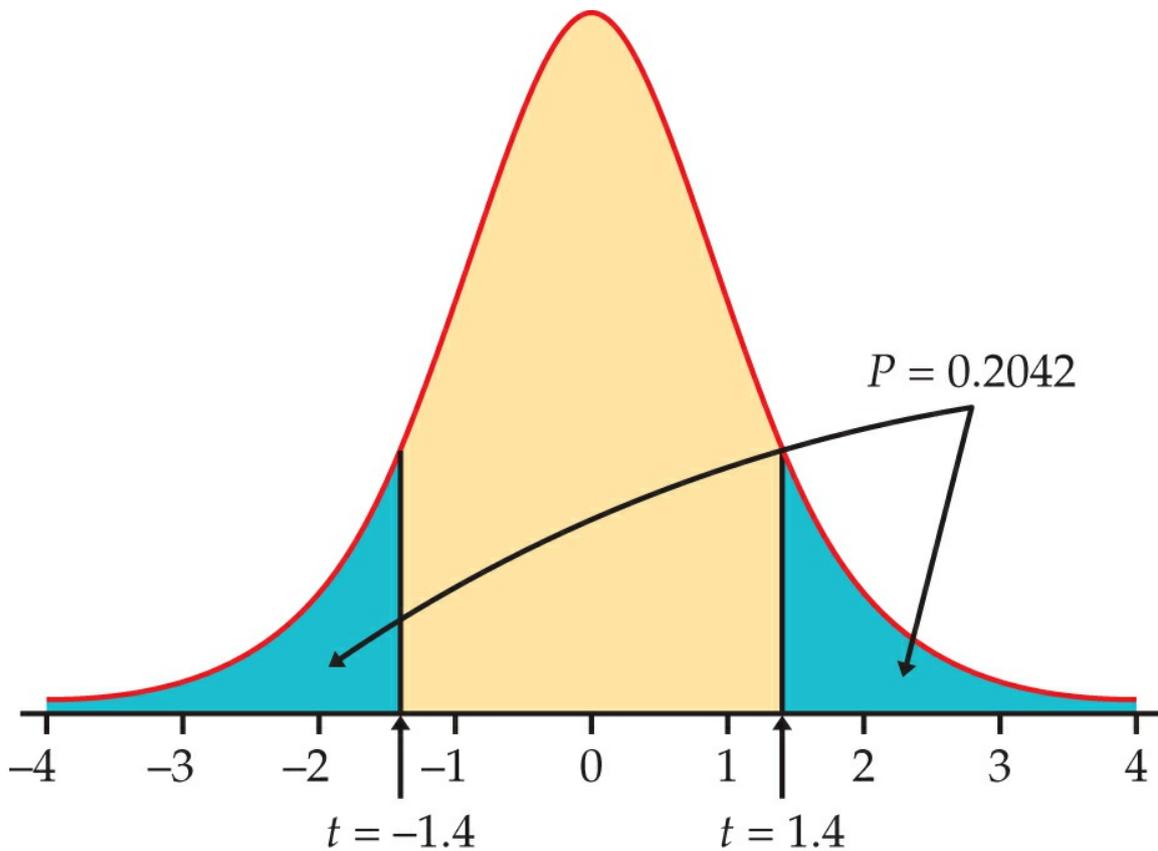
This means that the sample mean  $\bar{x} = 7.16$  is slightly less than 1.5 standard deviations away from the null hypothesized value  $\mu = 5.4$ . Because the degrees of freedom are  $n - 1 = 7$ , this  $t$  statistic has the  $t(7)$  distribution. Figure 7.3 shows that the  $P$ -value is  $2P(T \geq 1.40)$  where  $T$  has the  $t(7)$  distribution. From Table D we see that  $P(T \geq 1.119) = 0.15$  and  $P(T \geq 1.415) = 0.10$ .

Therefore, we conclude that the  $P$ -value is between  $2 \times 0.10 = 0.20$  and  $2 \times 0.15 = 0.30$ . Software gives the exact value as  $P = 0.2042$ . These data are compatible with a mean of 5.4 hours per month. Under  $H_0$  a difference this large or larger would occur about one time in five simply due to chance. There is not enough evidence to reject the null hypothesis at the 0.05 level.

$df = 7$



$p$	0.15	0.10
$t^*$	1.119	1.415



**FIGURE 7.3**

Sketch of the  $P$ -value calculation for Example 7.2.

In this example we tested the null hypothesis  $\mu = 5.4$  hours per month against the two-sided alternative  $\mu \neq 5.4$  hours per month because we had no prior suspicion that the average among college students would be larger or smaller. If we had suspected that the average would be larger, we would have used a one-sided test.

## Example

### 7.3 One-sided test for watching videos on a cell phone

For the problem described in the previous example, we want to test whether the U.S. college student average is larger than the overall U.S. population average. Here we test

$$H_0: \mu = 5.4$$

versus

$$H_a: \mu > 5.4$$

The  $t$  test statistic does not change:  $t = 1.40$ . As Figure 7.4 illustrates, however, the  $P$ -value is now  $P(T \geq 1.40)$ , half of the value in the previous example. From Table D we can determine that  $0.10 < P < 0.15$ ; software gives the exact value as  $P = 0.1021$ . Again, there is not enough evidence to reject the null hypothesis in favor of the alternative at the 0.05 significance level.



For the watching-videos problem, our conclusion did not depend on the choice between a one-sided and a two-sided test. Sometimes, however, this choice *will* affect the conclusion, and so this choice needs to be made prior to analysis. If in doubt, always use a two-sided test. *It is wrong to examine the data first and then decide to do a one-sided test in the direction indicated by the data.* Often a significant result for a two-sided test can be used to justify a one-sided test for *another* sample from the same population.

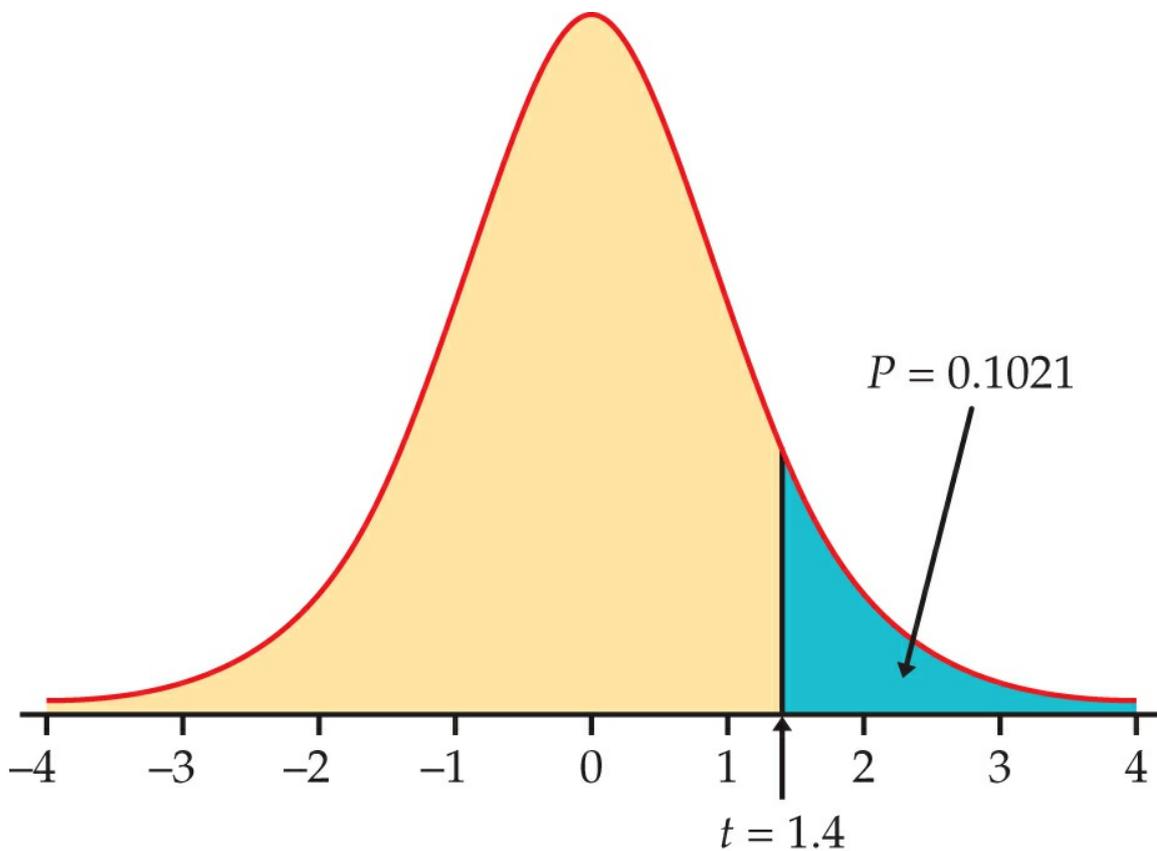


FIGURE 7.4

Sketch of the  $P$ -value calculation for Example 7.3.

For small data sets, such as the one in Example 7.1, it is easy to perform the computations for confidence intervals and significance tests with an ordinary calculator. For larger data sets, however, we prefer to use software or a statistical calculator.

## Example

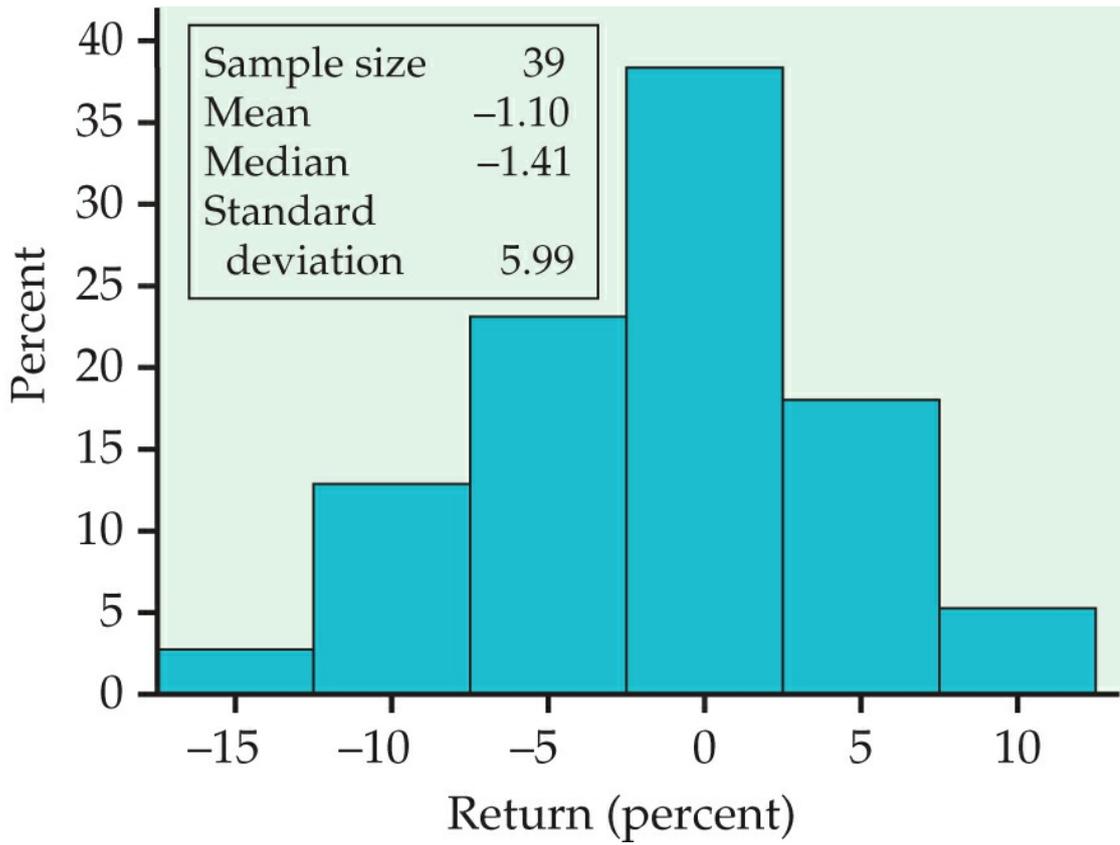
### 7.4 Stock portfolio diversification?

An investor with a stock portfolio worth several hundred thousand dollars sued his broker and brokerage firm because lack of diversification in his portfolio led to poor performance. Table 7.1 gives the rates of return for the 39 months that the account was managed by the broker.<sup>2</sup>

Figure 7.5 gives a histogram for these data and Figure 7.6 gives the Normal quantile plot. There are no outliers and the distribution shows no strong skewness. We are reasonably confident that the distribution of  $\bar{x}$  is approximately Normal, and we proceed with our inference based on Normal theory.

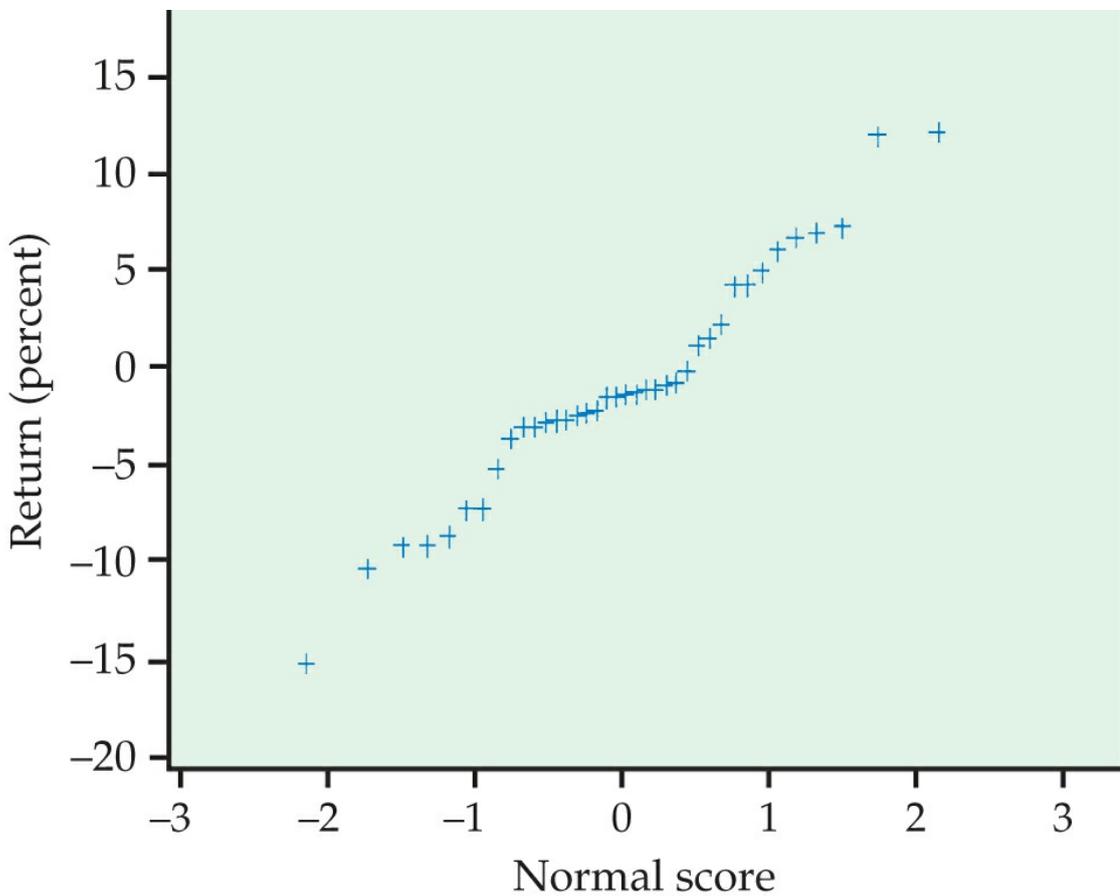
-8.36	1.63	-2.27	-2.93	-2.70	-2.93	-9.14	-2.64
6.82	-2.35	-3.58	6.13	7.00	-15.25	-8.66	-1.03
-9.16	-1.25	-1.22	-10.27	-5.11	-0.80	-1.44	1.28
-0.65	4.34	12.22	-7.21	-0.09	7.34	5.04	-7.24
-2.14	-1.01	-1.41	12.03	-2.56	4.33	2.35	

The arbitration panel compared these returns with the average of the Standard and Poor's 500 stock index for the same period. Consider the 39 monthly returns as a random sample from the population of monthly returns the brokerage firm would generate if it managed the account forever.



**FIGURE 7.5**

Histogram of monthly rates of return for a stock portfolio, for Example 7.4.



### FIGURE 7.6

Normal quantile plot for Example 7.4.

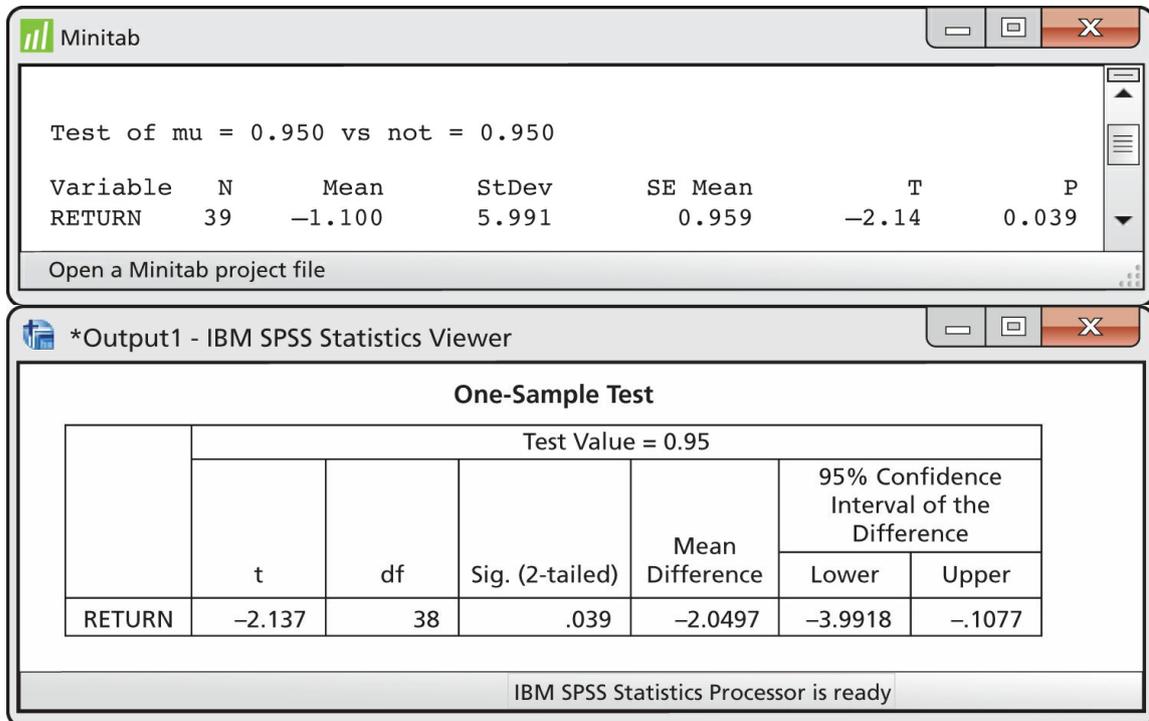
Are these returns compatible with a population mean of  $\mu = 0.95\%$ , the S&P 500 average? Our hypotheses are

$$H_0: \mu = 0.95$$

$$H_a: \mu \neq 0.95$$

Minitab and SPSS outputs appear in Figure 7.7. Output from other software will be similar.

Here is one way to report the conclusion: the mean monthly return on investment for this client's account was  $\bar{x} = -1.1\%$ . This is significantly worse than the performance of the S&P 500 stock index for the same period ( $t = -2.14$ ,  $df = 38$ ,  $P = 0.039$ ).



### FIGURE 7.7

Minitab and SPSS outputs for Example 7.4.

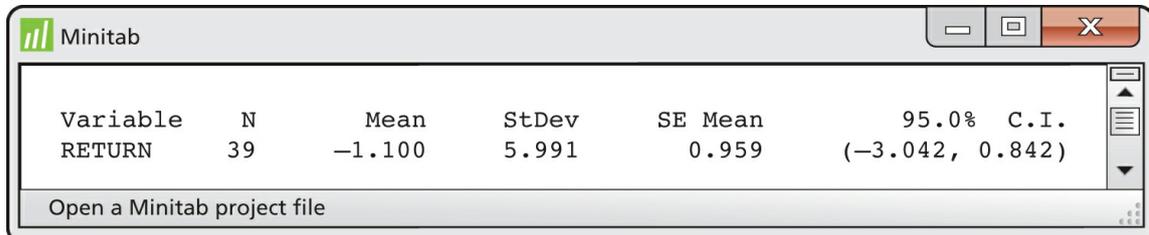
The hypothesis test in Example 7.4 leads us to conclude that the mean return on the client's account differs from that of the S&P 500 stock index. Now let's assess the return on the client's account with a confidence interval.

## Example

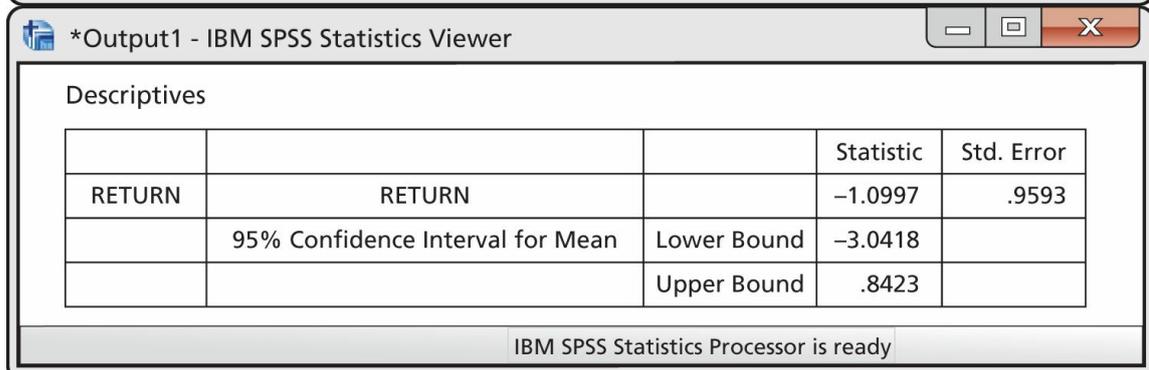
### 7.5 Estimating the mean monthly return

The mean monthly return on the client's portfolio was  $\bar{x} = -1.1\%$  and the standard deviation was  $s = 5.99\%$ . Figure 7.8 gives the Minitab, SPSS, and Excel outputs for a 95% confidence interval for the population mean  $\mu$ . Note that Excel gives the margin of error next to the label "Confidence Level(95.0%)" rather than the actual confidence interval. We see that the 95% confidence interval is  $(-3.04, 0.84)$  or (from Excel),  $-1.0997 \pm 1.9420$ .

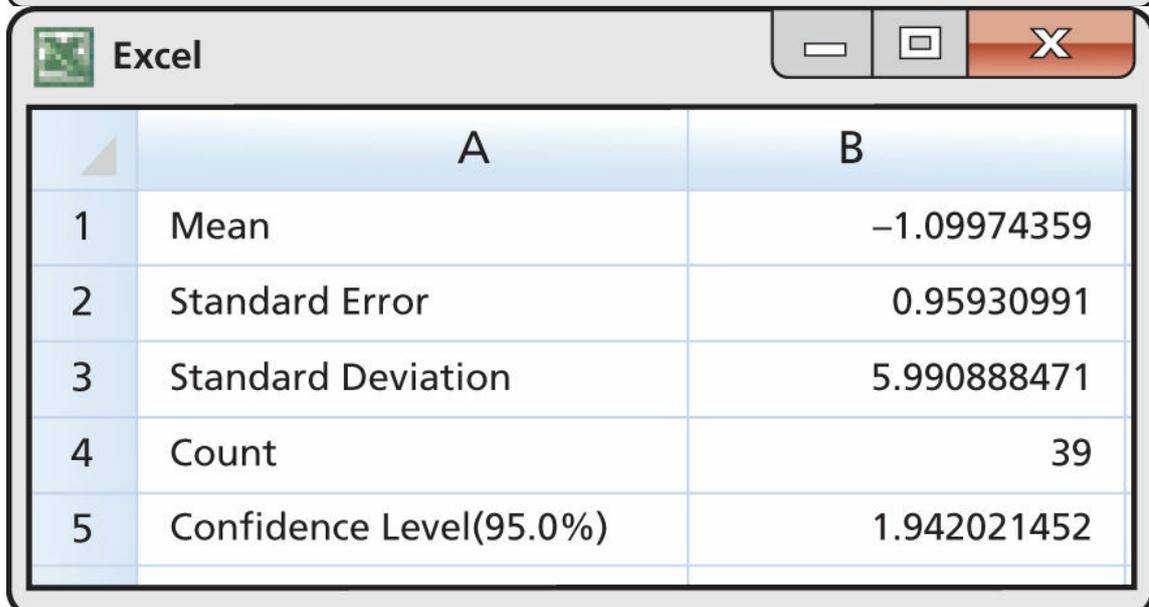
Because the S&P 500 return, 0.95%, falls outside this interval, we know that  $\mu$  differs significantly from 0.95% at the  $\alpha = 0.05$  level. Example 7.4 gave the actual  $P$ -value as  $P = 0.039$



Variable	N	Mean	StDev	SE Mean	95.0% C.I.
RETURN	39	-1.100	5.991	0.959	(-3.042, 0.842)



Descriptives				
			Statistic	Std. Error
RETURN	RETURN		-1.0997	.9593
	95% Confidence Interval for Mean	Lower Bound	-3.0418	
		Upper Bound	.8423	



	A	B
1	Mean	-1.09974359
2	Standard Error	0.95930991
3	Standard Deviation	5.990888471
4	Count	39
5	Confidence Level(95.0%)	1.942021452

## FIGURE 7.8

Minitab, SPSS, and Excel outputs for Example 7.5.

The confidence interval suggests that the broker's management of this account had a long-term mean somewhere between a loss of 3.04% and a gain of 0.84% per month. We are interested, not in the actual mean, but in the difference between the performance of the client's portfolio and that of the diversified S&P 500 stock index.

### Example

#### 7.6 Estimating the difference from a standard

Following the analysis accepted by the arbitration panel, we are considering the S&P 500 monthly average return as a constant standard. (It is easy to envision scenarios where we would want to treat this type of quantity as random.) The difference between the mean of the investor's account and the S&P 500 is  $\bar{x} - \mu = -1.10 - 0.95 = -2.05\%$ . In Example 7.5 we found that the 95% confidence interval for the investor's account was  $(-3.04, 0.84)$ .

To obtain the corresponding interval for the difference, subtract 0.95 from each of the endpoints. The resulting interval is  $(-3.04 - 0.95, 0.84 - 0.95)$  or  $(-3.99, -0.11)$ . We conclude with 95% confidence that the underperformance was between  $-3.99\%$  and  $-0.11\%$ . This interval is presented in the SPSS output of Figure 7.7. This estimate helps to set the compensation owed the investor.

The assumption that these 39 monthly returns represent an SRS from the population of monthly returns is certainly questionable. If the monthly S&P 500 returns were available, an alternative analysis would be to compare the average difference between each monthly return for this account and for the S&P 500. This method of analysis is discussed next.

### USE YOUR KNOWLEDGE

#### 7.5 Significance test using the $t$ distribution

A test of a null hypothesis versus a two-sided alternative gives  $t = 2.22$ .

(a) The sample size is 18. Is the test result significant at the 5% level? Explain how you obtained your answer.

(b) The sample size is 9. Is the test result significant at the 5% level? Explain how you obtained your answer.

(c) Sketch the two  $t$  distributions to illustrate your answers.

## 7.6 Significance test for apartment rents

Refer to Exercise 7.1 (page 420). Does this SRS give good reason to believe that the mean rent of all advertised one-bedroom apartments is greater than \$550? State the hypotheses, find the  $t$  statistic and its  $P$ -value, and state your conclusion.

## 7.7 Using software

In Example 7.1 (page 421) we calculated the 95% confidence interval for the U.S. college student average of hours per month spent watching videos on a cell phone. Use software to compute this interval and verify that you obtain the same interval.

## Matched pairs $t$ procedures



confounding, p. 173

The watching-videos problem of Example 7.1 concerns only a single population. We know that comparative studies are usually preferred to single-sample investigations because of the protection they offer against confounding. For that reason, inference about a parameter of a single distribution is less common than comparative inference.



matched pairs design, p. 186

One common comparative design, however, makes use of single-sample procedures. In a matched pairs study, subjects are matched in pairs, and their outcomes are compared within each matched pair. For example, an experiment to compare two cell phone packages might use pairs of subjects that are the same age, sex, and income level. The experimenter could toss a coin to assign the two

packages to the two subjects in each pair. The idea is that matched subjects are more similar than unmatched subjects, so comparing outcomes within each pair is more efficient (smaller  $\sigma$ ). Matched pairs are also common when randomization is not possible. One situation calling for matched pairs is when observations are taken on the same subjects under two different conditions or before and after some intervention. Here is an example.

## **Example**

### **7.7 Does a full moon affect behavior?**

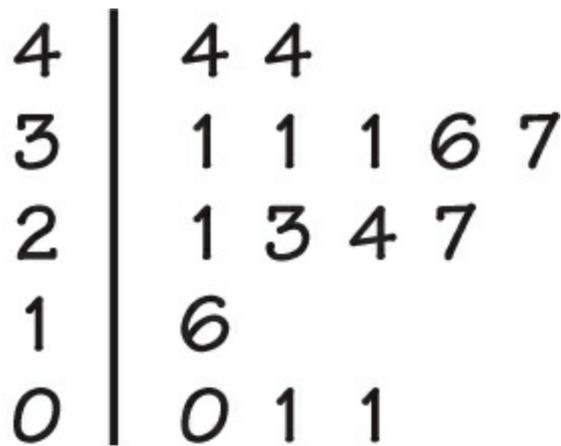
Many people believe that the moon influences the actions of some individuals. A study of dementia patients in nursing homes recorded various types of disruptive behaviors every day for 12 weeks. Days were classified as moon days if they were in a three-day period centered at the day of the full moon. For each patient the average number of disruptive behaviors was computed for moon days and for all other days. The data for the 15 subjects whose behaviors were classified as aggressive are presented in Table 7.2.<sup>3</sup> The patients in this study are not a random sample of dementia patients. However, we examine their data in the hope that what we find is not unique to this particular group of individuals and applies to other patients who have similar characteristics.



**TABLE 7.2** Aggressive Behaviors of Dementia Patients

Patient	Moon days	Other days	Difference	Patient	Moon days	Other days	Difference
1	3.33	0.27	3.06	9	6.00	1.59	4.41
2	3.67	0.59	3.08	10	4.33	0.60	3.73
3	2.67	0.32	2.35	11	3.33	0.65	2.68
4	3.33	0.19	3.14	12	0.67	0.69	-0.02
5	3.33	1.26	2.07	13	1.33	1.26	0.07
6	3.67	0.11	3.56	14	0.33	0.23	0.10
7	4.67	0.30	4.37	15	2.00	0.38	1.62
8	2.67	0.40	2.27				

To analyze these paired data, we first subtract the disruptive behaviors for other days from the disruptive behaviors for moon days. These 15 differences form a single sample. They appear in the “Difference” columns in Table 7.2. The first patient, for example, averaged 3.33 aggressive behaviors on moon days but only 0.27 aggressive behaviors on other days. The difference  $3.33 - 0.27 = 3.06$  is what we will use in our analysis.



**FIGURE 7.9**

Stemplot of differences in aggressive behaviors, for Examples 7.7 and 7.8.

Next, we examine the distribution of these differences. Figure 7.9 gives a stemplot of the differences. This plot indicates that there are three patients with very small differences, but there are no indications of extreme outliers or strong skewness. We will proceed with our analysis using the Normality-based methods of this section.

To assess whether there is a difference in aggressive behaviors on moon days versus other days, we test

$$H_0: \mu = 0$$

$$H_a: \mu \neq 0$$

Here  $\mu$  is the mean difference in aggressive behaviors, moon versus other days, for patients of this type. The null hypothesis says that aggressive behaviors occur at the same frequency for both types of days, and  $H_a$  says that the frequency of aggressive behaviors on moon days is not the same as on other days.

The 15 differences have

$$\bar{x} = 2.433 \text{ and } s = 1.460$$

The one-sample  $t$  statistic is therefore

$$\begin{aligned} t &= \frac{\bar{x} - 0}{s/\sqrt{n}} = \frac{2.433 - 0}{1.460/\sqrt{15}} \\ &= 6.45 \end{aligned}$$

The  $P$ -value is found from the  $t(14)$  distribution (remember that the degrees of freedom are 1 less than the sample size).

$df = 14$

$p$	0.001	0.0005
$t^*$	3.787	4.140

Table D shows that 6.45 lies beyond the upper 0.0005 critical value of the  $t(14)$  distribution. Since we are using a two-sided alternative, we know that the  $P$ -value is less than two times this value, or 0.0010. Software gives a value that is much smaller,  $P = 0.000015$ . In practice, there is little difference between these two  $P$ -values; the data provide clear evidence in favor of the alternative hypothesis. A mean difference this large is very unlikely to occur by chance if there is, in fact, no effect of the moon on aggressive behaviors. The positive mean difference indicates that more aggressive behaviors occur on moon days. In scholarly publications, the details of routine statistical procedures are omitted, and our test would be reported in the following form: “There were more aggressive behaviors on moon days than on other days ( $t = 6.45$ ,  $df = 14$ ,  $P < 0.001$ ).”



Note that we could have justified a one-sided alternative in this example. Based on previous research, we expect more aggressive behaviors on moon days, and the alternative  $H_a: \mu > 0$  is reasonable in this setting. The choice of the alternative here, however, has no effect on the conclusion: from Table D we determine that  $P$  is less than 0.0005 for the one-sided alternative; from software it is 0.000008. These are very small values and we would still report  $P < 0.001$ . *In most circumstances we cannot be absolutely certain about the direction of the difference, and the safest strategy is to use the two-sided alternative.*

The results of the significance test allow us to conclude that dementia patients exhibit more aggressive behaviors in the days around a full moon. What are the implications of the study for the administrators who run the facilities where these patients live? For example, should they increase staff on these days? To make these kinds of decisions, an estimate of the magnitude of the problem, with a margin of error, would be helpful.

## Example

### 7.8 A 95% confidence interval for the full-moon study

A 95% confidence interval for the mean difference in aggressive behaviors per

day requires the critical value  $t^* = 2.145$  from Table D. The margin of error is

$$\begin{aligned} t^*sn &= 2.145(1.46015) \\ &= 0.81 \end{aligned}$$

and the confidence interval is

$$\begin{aligned} \bar{x} \pm t^*sn &= 2.43 \pm 0.81 \\ &= (1.62, 3.24) \end{aligned}$$

The estimated average difference is 2.43 aggressive behaviors per day, with margin of error 0.81 for 95% confidence. The increase needs to be interpreted in terms of the baseline values. The average number of aggressive behaviors per day on other days for our 15 patients is 0.59; on moon days it is 3.02. This is approximately a 400% increase. If aggressive behaviors require a substantial amount of attention by staff, then administrators should be aware of the increased level of these activities during the full-moon period. Additional staff may be needed.

The following are key points to remember concerning matched pairs:

1. A matched pairs analysis is called for when subjects are matched in pairs or there are two measurements or observations on each individual and we want to examine the difference.
2. For each pair or individual, use the difference between the two measurements as the data for your analysis.
3. Use the one-sample confidence interval and significance-testing procedures that we learned in this section.

Use of the  $t$  procedures in Examples 7.7 and 7.8 faces several issues. First, no randomization is possible in a study like this. Our inference procedures assume that there is a process that generates these aggressive behaviors and that the process produces them at possibly different rates during the days near the full moon. Second, many of the patients in these nursing homes did not exhibit any disruptive behaviors. These were not included in our analysis, so our inference is restricted to patients who do exhibit disruptive behaviors.



A final difficulty is that the data show departures from Normality. In a matched pairs analysis, the  $t$  procedures are applied to the differences, so we are assuming that the differences are Normally distributed. Figure 7.9 gives a stemplot of the differences. There are 3 patients with very small differences in aggressive behaviors while the other 12 have a large increase. We have a dilemma here

similar to that in Example 7.1. *The data may not be Normal, and our sample size is very small.* We can try an alternative procedure that does not require the Normality assumption—but there is a price to pay. The alternative procedures have less power to detect differences. Despite these caveats, for Example 7.7 the  $P$ -value is so small that we are very confident that we have found an effect of the moon phase on behavior.

## USE YOUR KNOWLEDGE

### 7.8 Comparison of two energy drinks

Consider the following study to compare two popular energy drinks. For each subject, a coin was flipped to determine which drink to rate first. Each drink was rated on a 0 to 100 scale, with 100 being the highest rating.

	Subject				
Drink	1	2	3	4	5
A	48	90	83	96	93
B	49	78	66	88	71

Is there a difference in preference? State appropriate hypotheses and carry out a matched pairs  $t$  test for these data.

### 7.9 A 95% confidence interval for the difference in preference

Refer to the previous exercise. For the company producing Drink A, the real question is how much difference there is between the two preferences. Use the data in Exercise 7.8 to give a 95% confidence interval for the difference in preference between Drink A and Drink B.

## Robustness of the $t$ procedures

The results of one-sample  $t$  procedures are exactly correct only when the population is Normal. Real populations are never exactly Normal. The usefulness of the  $t$  procedures in practice therefore depends on how strongly they are affected by non-Normality. Procedures that are not strongly affected are called *robust*.

## ROBUST PROCEDURES

A statistical inference procedure is called **robust** if the required probability calculations are insensitive to violations of the assumptions made.



resistant measure, p. 32

The assumption that the population is Normal rules out outliers, so the presence of outliers shows that this assumption is not valid. The  $t$  procedures are not robust against outliers, because  $\bar{x}$  and  $s$  are not resistant to outliers.



In Example 7.7, there are three patients with fairly low values of the difference. Whether or not these are outliers is a matter of judgment. If we rerun the analysis without these three patients, the  $t$  statistic would increase to 11.89 and the  $P$ -value would be much lower. Careful inspection of the records may reveal some characteristic of these patients which distinguishes them from the others in the study. Without such information, it is difficult to justify excluding them from the analysis. *In general, we should be very cautious about discarding suspected outliers, particularly when they make up a substantial proportion of the data, as they do in this example.*

Fortunately, the  $t$  procedures are quite robust against non-Normality of the population except in the case of outliers or strong skewness. Larger samples improve the accuracy of  $P$ -values and critical values from the  $t$  distributions when the population is not Normal. This is true for two reasons:

1. The sampling distribution of the sample mean  $\bar{x}$  from a large sample is close to Normal (that's the central limit theorem). Normality of the individual observations is of little concern when the sample is large.



central limit theorem, p. 307

2. As the sample size  $n$  grows, the sample standard deviation  $s$  will be an accurate estimate of  $\sigma$  whether or not the population has a Normal distribution. This fact is closely related to the law of large numbers.



law of large numbers, p. 268



Constructing a Normal quantile plot, stemplot, or boxplot to check for skewness and outliers is an important preliminary to the use of  $t$  procedures for small samples. For most purposes, the one-sample  $t$  procedures can be safely used when  $n \geq 15$  unless an outlier or clearly marked skewness is present. *Except in the case of small samples, the assumption that the data are an SRS from the population of interest is more crucial than the assumption that the population distribution is Normal.* Here are practical guidelines for inference on a single mean:<sup>4</sup>

- *Sample size less than 15:* Use  $t$  procedures if the data are close to Normal. If the data are clearly non-Normal or if outliers are present, do not use  $t$ .
- *Sample size at least 15 and less than 40:* The  $t$  procedures can be used except in the presence of outliers or strong skewness.
- *Large samples:* The  $t$  procedures can be used even for clearly skewed distributions when the sample is large, roughly  $n \geq 40$ .

Consider, for example, some of the data we studied in Chapter 1. The service center call lengths in Table 1.2 (page 19) are strongly skewed to the right. Since there are 80 observations, we could use the  $t$  procedures here. On the other hand, many would prefer to use a transformation to make these data more nearly Normal. (See the material on inference for non-Normal populations on page 436 and in Chapter 16.) The time to start a business data in Exercise 1.47 (page 32) contain one outlier in a sample of size 25, which makes the use of  $t$  procedures more risky. Figure 1.29 (page 70) gives the Normal quantile plot for 60 IQ scores. These data appear to be Normal and we would apply the  $t$  procedures in this case.

## USE YOUR KNOWLEDGE

### 7.10 $t$ procedures for time to start a business?

Consider the data from Exercise 1.47 (page 32) but with Suriname removed. Would you be comfortable applying the  $t$  procedures in this case? Explain your answer.

## 7.11 $t$ procedures for ticket prices?

Consider the data on StubHub! ticket prices presented in Figure 1.31 (page 71). Would you be comfortable applying the  $t$  procedures in this case? In explaining your answer, recall that these  $t$  procedures focus on the mean  $\mu$ .

### The power of the $t$ test

The power of a statistical test measures its ability to detect deviations from the null hypothesis. In practice, we carry out the test in the hope of showing that the null hypothesis is false, so high power is important. The power of the one-sample  $t$  test for a specific alternative value of the population mean  $\mu$  is the probability that the test will reject the null hypothesis when the alternative value of the mean is true. To calculate the power, we assume a fixed level of significance, often  $\alpha = 0.05$ .

Calculation of the exact power of the  $t$  test takes into account the estimation of  $\sigma$  by  $s$  and is a bit complex. But an approximate calculation that acts as if  $\sigma$  were known is almost always adequate for planning a study. This calculation is very much like that for the  $z$  test:

1. Decide on a standard deviation, a significance level, whether the test is one-sided or two-sided, and an alternative value of  $\mu$  to detect.
2. Write the event that the test rejects  $H_0$  in terms of  $\bar{x}$
3. Find the probability of this event when the population mean has this alternative value.

Consider Example 7.7 (page 429), where we examined the effect of the moon on the aggressive behavior of dementia patients in nursing homes. Suppose that we wanted to perform a similar study in a different setting. How many patients should we include in our new study? To answer this question, we do a power calculation.



In Example 7.7, we found  $\bar{x} = 2.433$  and  $s = 1.460$ . Let's use  $s = 1.5$  for our calculations. *It is always better to use a value of the standard deviation that is a little larger than what we expect than to use one that is smaller.* This may give a sample size that is a little larger than we need. But we want to avoid a situation where we fail to find the effect that we are looking for because we did not have enough data.

Let's use  $\mu = 1.0$  as the alternative value to detect. We are very confident that the effect was larger than this in our previous study, and this amount of an increase

in aggressive behavior would still be important to those who work in these facilities. Finally, based on the previous study, we can justify using a one-sided alternative; we expect the moon days to be associated with an increase in aggressive behavior.

Now that we've decided on the necessary information for Step 1, we can proceed through the calculations of Steps 2 and 3.

## Example

### 7.9 Computing the power of a $t$ test

Based on our previous decisions, we'll compute the power of the  $t$  test for

$$H_0: \mu = 0$$

$$H_a: \mu > 0$$

when the alternative  $\mu = 1.0$ . We will use a 5% level of significance and  $s = 1.5$  for these calculations. The  $t$  test with  $n$  observations rejects  $H_0$  at the 5% significance level if the  $t$  statistic

$$t = \frac{\bar{x} - 0}{s/n}$$

exceeds the upper 5% point of  $t(n - 1)$ . Since this is a new study in a different setting, we'll assume that we'll recruit  $n = 20$  patients. The upper 5% point of  $t(19)$  is 1.729. The event that the test rejects  $H_0$  is therefore

$$t = \frac{\bar{x} - 1.5/20 \geq 1.729$$

$$\bar{x} \geq 1.729 + 1.5/20$$

$$\bar{x} \geq 0.580$$

The power is the probability that  $\bar{x} \geq 0.580$  when  $\mu = 1.0$ . Taking  $\sigma = 1.5$ , this probability is found by standardizing  $\bar{x}$ :

$$P(\bar{x} \geq 0.580 \text{ when } \mu = 1.0) = P(\bar{x} - 1.0 - 1.5/20 \geq 0.580 - 1.0 - 1.5/20)$$

$$= P(Z \geq -1.25)$$

$$= 1 - 0.1056 = 0.89$$

The power is 89% that we will detect an increase of 1.0 aggressive behavior per day during moon days. This is sufficient power for most situations. For many

studies, 80% is considered the standard value for desirable power. We could repeat the calculations for some smaller values of  $n$  to determine the smallest value that would meet the 80% criterion.



Power calculations are used in planning studies to ensure that we have a reasonable chance of detecting effects of interest. They give us some guidance in selecting a sample size. In making these calculations, we need assumptions about the standard deviation and the alternative of interest. In our example we assumed that the standard deviation would be 1.5, but in practice we are hoping that the value will be somewhere around this value. Similarly, we have used a somewhat arbitrary alternative of 1.0. This is a guess based on the results of the previous study. *Beware of putting too much trust in fine details of the results of these calculations.* They serve as a guide, not a mandate.

## USE YOUR KNOWLEDGE

### 7.12 Power and the alternative mean $\mu$

If you were to repeat the power calculation in Example 7.9 for a value of  $\mu$  that is greater than 1, would you expect the power to be higher or lower than 89%? Why?

### 7.13 More on power and the alternative mean $\mu$

Verify your answer to the previous question by doing the calculation for the alternative  $\mu$

### 7.14 Power and sample size $n$

If you were to repeat the power calculation in Example 7.9 using  $n = 25$  instead of  $n = 20$ , would you expect the power to be higher or lower than 89%? Why?

### 7.15 More on power and sample size $n$

Verify your answer to the previous question by doing the calculation for

the alternative  $\mu = 1$  and  $n = 25$ .

## Inference for non-Normal populations

We have not discussed how to do inference about the mean of a clearly non-Normal distribution based on a small sample. If you face this problem, you should consult an expert. Three general strategies are available:

1. In some cases a distribution other than a Normal distribution will describe the data well. There are many non-Normal models for data, and inference procedures for these models are available.
2. Because skewness is the chief barrier to the use of  $t$  procedures on data without outliers, you can attempt to transform skewed data so that the distribution is symmetric and as close to Normal as possible. Confidence levels and  $P$ -values from the  $t$  procedures applied to the transformed data will be quite accurate for even moderate sample sizes.
3. Use a **distribution-free** inference procedure. Such procedures do not assume that the population distribution has any specific form, such as Normal. Distribution-free procedures are often called **nonparametric procedures**. Chapter 15 discusses several of these procedures.

### distribution-free procedures

### nonparametric procedures

Each of these strategies can be effective, but each quickly carries us beyond the basic practice of statistics. We emphasize procedures based on Normal distributions because they are the most common in practice, because their robustness makes them widely useful, and (most important) because we are first of all concerned with understanding the principles of inference. We will therefore not discuss procedures for non-Normal continuous distributions. We will be content with illustrating by example the use of a transformation and of a simple distribution-free procedure.

## Transforming data

When the distribution of a variable is skewed, it often happens that a simple transformation results in a variable whose distribution is symmetric and even close to Normal. The most common transformation is the logarithm, or log. The logarithm tends to pull in the right tail of a distribution. For example, the data 2, 3, 4, 20 show an outlier in the right tail. Their common logarithms 0.30, 0.48, 0.60, 1.30 are much less skewed. Taking logarithms is a possible remedy for right-skewness. Instead of analyzing values of the original variable  $X$ , we compute their logarithms and analyze the values of  $X$ . Here is an example of this approach.

**TABLE 7.3****Length (in Seconds) of Audio Files Sampled from an iPod**

240	316	259	46	871	411	1366
233	520	239	259	535	213	492
315	696	181	357	130	373	245
305	188	398	140	252	331	47
309	245	69	293	160	245	184
326	612	474	171	498	484	271
207	169	171	180	269	297	266
1847						

## Example

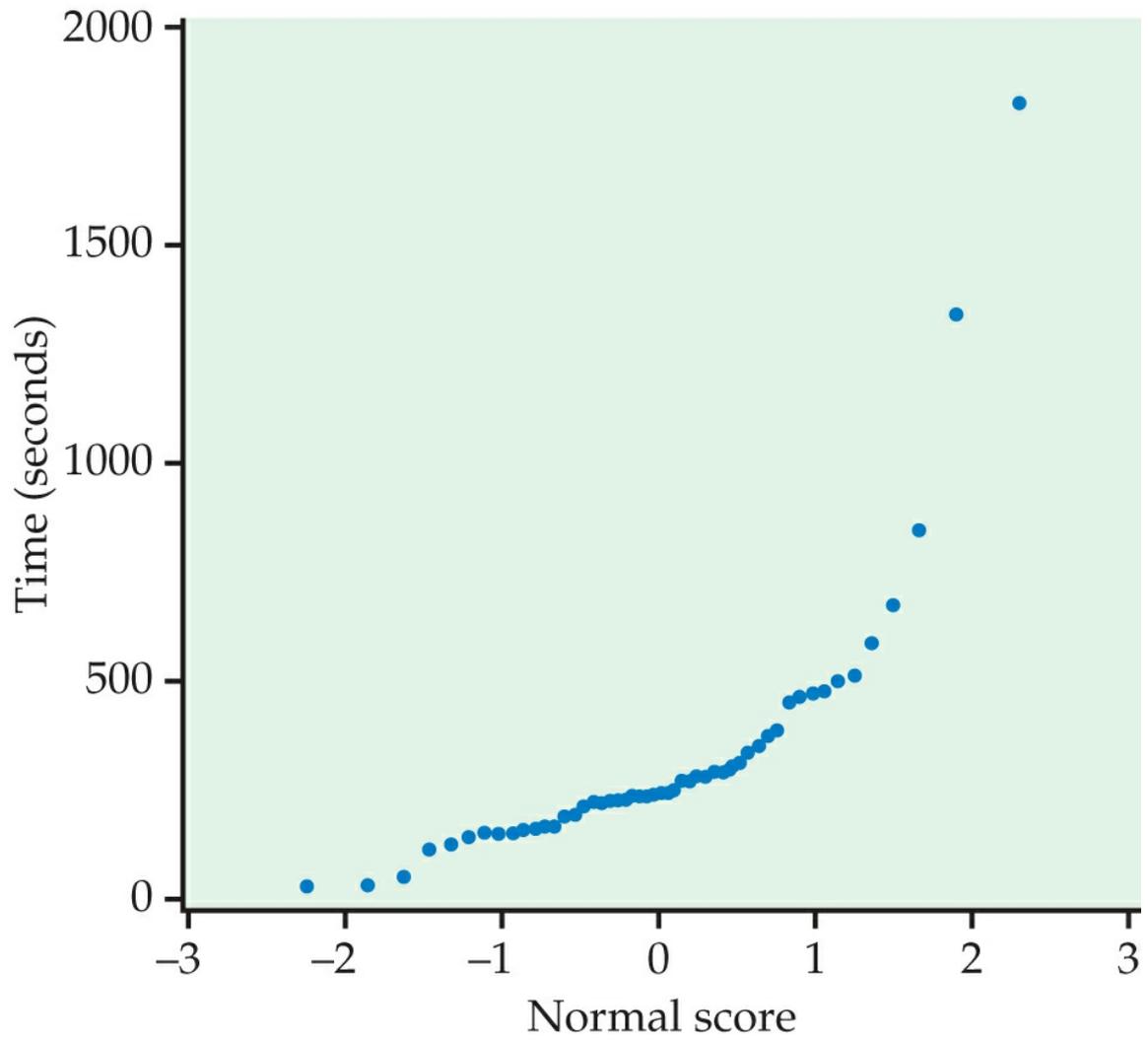
### 7.10 Length of audio files on an iPod

Table 7.3 presents data on the length (in seconds) of audio files found on an iPod. There was a total of 10,003 audio files, and 50 files were randomly selected using the “shuffle songs” command.<sup>5</sup> We would like to give a confidence interval for the average audio file length  $\mu$  for this iPod.



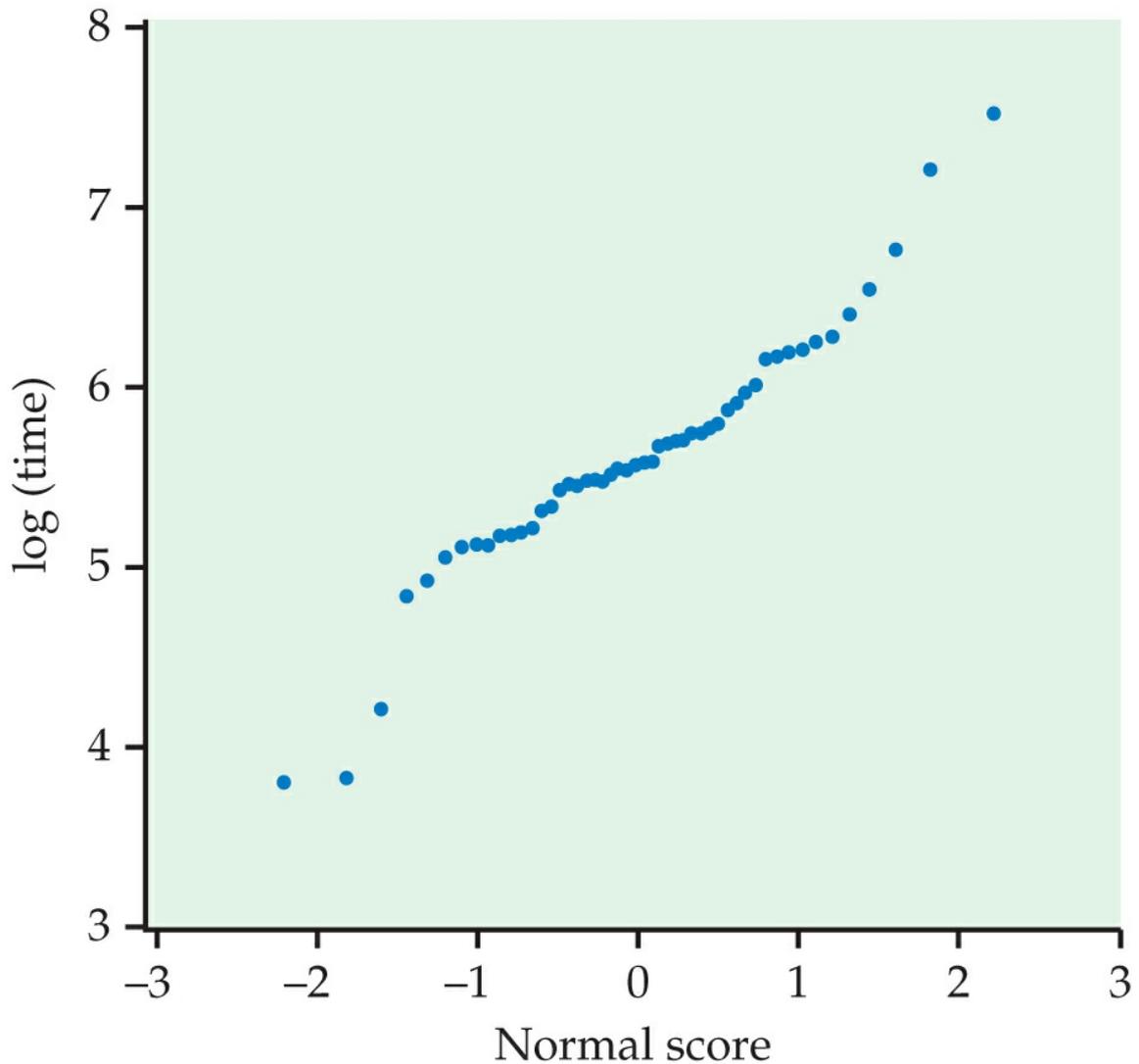
A Normal quantile plot of the audio data from Table 7.3 (Figure 7.10) shows that the distribution is skewed to the right. Because there are no extreme outliers, the sample mean of the 50 observations will nonetheless have an approximately Normal sampling distribution. The  $t$  procedures could be used for approximate inference. For more exact inference, we will transform the data so that the distribution is more nearly Normal. Figure 7.11 is a Normal quantile plot of the natural logarithms of the time measurements. The transformed data are very close to Normal, so  $t$  procedures will give quite

exact results.



**FIGURE 7.10**

Normal quantile plot of audio file lengths, for Example 7.10. This sort of pattern occurs when a distribution is skewed to the right.



**FIGURE 7.11**

Normal quantile plot of the logarithms of the audio file lengths, for Example 7.10. This distribution appears approximately Normal.

The application of the  $t$  procedures to the transformed data is straight-forward. Call the original length values from Table 7.3 the variable  $X$ . The transformed data are values of  $X_{\text{new}} = \log X$ . In most software packages, it is an easy task to transform data in this way and then analyze the new variable.

## Example

### 7.11 Software output of audio length data

Analysis of the natural log of the length values in Minitab produces the following output:

N	Mean	StDev	SE Mean	95.0% C.I.
50	5.6315	0.6840	0.0967	(5.4371, 5.8259)

For comparison, the 95%  $t$  confidence interval for the original mean  $\mu$  is found from the original data as follows:

N	Mean	StDev	SE Mean	95.0% C.I.
50	354.1	307.9	43.6	(266.6, 441.6)

The advantage of analyzing transformed data is that use of procedures based on the Normal distributions is better justified and the results are more exact. The disadvantage is that a confidence interval for the mean  $\mu$  in the original scale (in our example, seconds) cannot be easily recovered from the confidence interval for the mean of the logs. One approach based on the log Normal distribution<sup>6</sup> results in an interval of (285.5, 435.5), which is narrower and slightly asymmetric compared with the  $t$  interval.

### The sign test

Perhaps the most straightforward way to cope with non-Normal data is to use a *distribution-free*, or *nonparametric*, procedure. As the name indicates, these procedures do not require the population distribution to have any specific form, such as Normal. Distribution-free significance tests are quite simple and are available in most statistical software packages. Distribution-free tests have two drawbacks. First, they are generally less powerful than tests designed for use with a specific distribution, such as the  $t$  test. Second, we must often modify the statement of the hypotheses in order to use a distribution-free test. A distribution-free test concerning the center of a distribution, for example, is usually stated in terms of the median rather than the mean. This is sensible when the distribution may be skewed. But the distribution-free test does not ask the same question (Has the mean changed?) that the  $t$  test does. The simplest distribution-free test, and one of the most useful, is the **sign test**.

### sign test

Let's examine again the aggressive-behavior data of Example 7.7 (page 429). In that example we concluded that there was more aggressive behavior on moon days than on other days. The stemplot given in Figure 7.9 was not very reassuring concerning the assumption that the data are Normal. There were 3 patients with low values that seemed to be somewhat different from the observations on the other 12 patients. How does the sign test deal with these data?

## Example

### 7.12 Sign test for the full-moon effect

The sign test is based on the following simple observation: of the 15 patients in our sample, 14 had more aggressive behaviors on moon days than on other days. This sounds like convincing evidence in favor of a moon effect on behavior, but we need to do some calculations to confirm this.



Let  $P$  be the probability that a randomly chosen dementia patient will have more aggressive behaviors on moon days than on other days. The null hypothesis of “no moon effect” says that the moon days are no different from other days, so a patient is equally likely to have more aggressive behaviors on moon days as on other days. We therefore want to test

$$H_0:p = 1/2$$

$$H_a:p > 1/2$$

There are 15 patients in the study, so the number who have more aggressive behaviors on moon days has the binomial distribution  $B(15, 1/2)$  if  $H_0$  is true. The  $P$ -value for the observed count 14 is therefore  $P(X \geq 14)$ , where  $X$  has the  $B(15, 1/2)$  distribution. You can compute this probability with software or from the binomial probability formula:

$$\begin{aligned} P(X \geq 14) &= P(X = 14) + P(X = 15) \\ &= \binom{15}{14} (1/2)^{14} (1/2)^1 + \binom{15}{15} (1/2)^{15} (1/2)^0 \\ &= \binom{15}{14} (1/2)^{15} + \binom{15}{15} (1/2)^{15} \\ &= 0.000488 \end{aligned}$$

Using Table C we would approximate this value as 0.0005. As in Example 7.7, there is very strong evidence in favor of an increase in aggressive behavior on moon days.

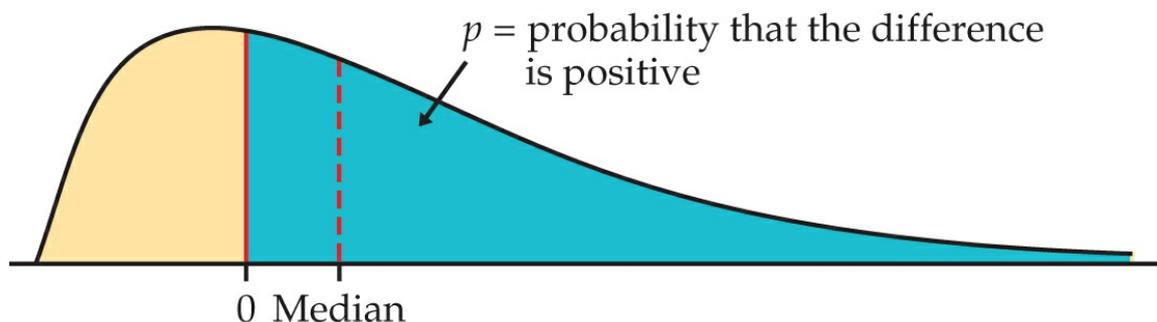
There are several varieties of sign test, all based on counts and the binomial distribution. The sign test for matched pairs (Example 7.12) is the most useful. The null hypothesis of “no effect” is then always  $H_0:p = 1/2$ . The alternative can be

one-sided in either direction or two-sided, depending on the type of change we are looking for. The test gets its name from the fact that we look only at the signs of the differences, not their actual values.

## THE SIGN TEST FOR MATCHED PAIRS

Ignore pairs with difference 0; the number of trials  $n$  is the count of the remaining pairs. The test statistic is the count  $X$  of pairs with a positive difference.  $P$ -values for  $X$  are based on the binomial  $B(n, 1/2)$  distribution.

The matched pairs  $t$  test in Example 7.7 tested the hypothesis that the mean of the distribution of differences (moon days minus other days) is 0. The sign test in Example 7.12 is in fact testing the hypothesis that the *median* of the differences is 0. If  $P$  is the probability that a difference is positive, then  $p = 1/2$  when the median is 0. This is true because the median of the distribution is the point with probability  $1/2$  lying to its right. As Figure 7.12 illustrates,  $p > 1/2$  when the median is greater than 0, again because the probability to the right of the median is always  $1/2$ . The sign test of  $H_0: p = 1/2$  against  $H_a: p \neq 1/2$  is a test of



**FIGURE 7.12**

Why the sign test tests the median difference: when the median is greater than 0, the probability  $p$  of a positive difference is greater than  $1/2$ , and vice versa.

$$H_0: \text{population median} = 0$$

$$H_a: \text{population median} > 0$$



The sign test in Example 7.12 makes no use of the actual differences—it just counts how many patients had more aggressive behaviors on moon days than on other days. Because the sign test uses so little of the available information, it is much less powerful than the  $t$  test when the population is close to Normal. *It is*

better to use a test that is powerful when we believe our assumptions are approximately satisfied than a less powerful test with fewer assumptions. There are other distribution-free tests that are more powerful than the sign test.<sup>7</sup>

## USE YOUR KNOWLEDGE

### 7.16 Sign test for energy drink comparison

Exercise 7.8 (page 432) gives data on the appeal of two popular energy drinks. Is there evidence that the medians are different? State the hypotheses, carry out the sign test, and report your conclusion.

## Section 7.1 Summary

Significance tests and confidence intervals for the mean  $\mu$  of a Normal population are based on the sample mean  $\bar{x}$  of an SRS. Because of the central limit theorem, the resulting procedures are approximately correct for other population distributions when the sample is large.

The standardized sample mean, or **one-sample  $z$  statistic**,

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

has the  $N(0,1)$  distribution. If the standard deviation  $\sigma/\sqrt{n}$  of  $\bar{x}$  is replaced by the **standard error  $s/\sqrt{n}$**  the **one-sample  $t$  statistic**

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

has the  **$t$  distribution** with  $n - 1$  degrees of freedom.

There is a  $t$  distribution for every positive **degrees of freedom  $k$** . All are symmetric distributions similar in shape to Normal distributions. The  $t(k)$  distribution approaches the  $N(0, 1)$  distribution as  $k$  increases.

A level  $C$  **confidence interval for the mean  $\mu$**  of a Normal population is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

where  $t^*$  is the value for the  $t(n - 1)$  density curve with area  $C$  between  $-t^*$  and  $t^*$ . The quantity

$$t^* \frac{s}{\sqrt{n}}$$

is the **margin of error**.

Significance tests for  $H_0: \mu = \mu_0$  are based on the  $t$  statistic.  $P$ -values or fixed significance levels are computed from the  $t(n - 1)$  distribution.

These one-sample procedures are used to analyze **matched pairs** data by first taking the differences within the matched pairs to produce a single sample.

The  $t$  procedures are relatively **robust** against non-Normal populations. The  $t$  procedures are useful for non-Normal data when  $15 \leq n < 40$  unless the data show outliers or strong skewness. When  $n \geq 40$ , the  $t$  procedures can be used even for clearly skewed distributions.

The **power** of the  $t$  test is calculated like that of the  $t$  test, using an approximate value for both  $\sigma$  and  $s$ .

Small samples from skewed populations can sometimes be analyzed by first applying a transformation (such as the logarithm) to obtain an approximately Normally distributed variable. The  $t$  procedures then apply to the transformed data.

The **sign test** is a **distribution-free test** because it uses probability calculations that are correct for a wide range of population distributions.

The sign test for “no treatment effect” in matched pairs counts the number of positive differences. The  $P$ -value is computed from the  $B(n, 1/2)$  distribution, where  $n$  is the number of nonzero differences. The sign test is less powerful than the  $t$  test in cases where use of the  $t$  test is justified.

## SECTION 7.1 Exercises

*For Exercises 7.1 and 7.2, see page 420; for Exercises 7.3 and 7.4, see page 422; for Exercises 7.5 to 7.7, see page 427; for Exercises 7.8 and 7.9, see page 432; for Exercises 7.10 and 7.11, see page 434; for Exercises 7.12 to 7.15, see page 436; and for Exercise 7.16, see page 440.*

### 7.17 Finding the critical value $t^*$

What critical value  $t^*$  from Table D should be used to calculate the margin of error for a confidence interval for the mean of the population in each of the following situations?

- (a) A 95% confidence interval based on  $n = 12$  observations.
- (b) A 95% confidence interval from an SRS of 21 observations.
- (c) A 90% confidence interval from a sample of size 21.
- (d) These cases illustrate how the size of the margin of error depends upon the confidence level and the sample size. Summarize these relationships.

### 7.18 Distribution of the $t$ statistic

Assume a sample size of  $n = 16$ . Draw a picture of the distribution of the  $t$  statistic under the null hypothesis. Use Table D and your picture to illustrate the values of the test statistic that would lead to rejection of the null hypothesis at the 5% level for a two-sided alternative.

### 7.19 More on the distribution of the $t$ statistic

Repeat the previous exercise for the two situations where the alternative is one-sided.

## 7.20 One-sided versus two-sided $P$ -values

Computer software reports  $\bar{x} = 11.2$  and  $P = 0.068$  for a  $t$  test of  $H_0: \mu = 0$  versus  $H_a: \mu \neq 0$ . Based on prior knowledge, you justified testing the alternative  $H_a: \mu > 0$ . What is the  $P$ -value for your significance test?

## 7.21 More on one-sided versus two-sided $P$ -values

Suppose that computer software reports  $\bar{x} = -11.2$  and  $P = 0.068$  for a  $t$  test of  $H_0: \mu = 0$  versus  $H_a: \mu \neq 0$ . Would this change your  $P$ -value for the alternative hypothesis in the previous exercise? Use a sketch of the distribution of the test statistic under the null hypothesis to illustrate and explain your answer.

## 7.22 A one-sample $t$ test

The one-sample  $t$  statistic for testing

$$H_0: \mu = 8$$

$$H_a: \mu > 8$$

from a sample of  $n = 16$  observations has the value  $t = 2.15$

- What are the degrees of freedom for this statistic?
- Give the two critical values  $t^*$  from Table D that bracket  $t$ .
- Between what two values does the  $P$ -value of the test fall?
- Is the value  $t = 2.15$  significant at the 5% level? Is it significant at the 1% level?
- If you have software available, find the exact  $P$ -value.

## 7.23 Another one-sample $t$ test

The one-sample  $t$  statistic for testing

$$H_0: \mu = 40$$

$$H_a: \mu \neq 40$$

from a sample of  $n = 27$  observations has the value  $t = 2.01$

- What are the degrees of freedom for  $t$ ?
- Locate the two critical values  $t^*$  from Table D that bracket  $t$ .
- Between what two values does the  $P$ -value of the test fall?
- Is the value  $t = 2.01$  statistically significant at the 5% level? At the 1% level?
- If you have software available, find the exact  $P$ -value.

## 7.24 A final one-sample $t$ test

The one-sample  $t$  statistic for testing

$$H_0: \mu = 20$$

$$H_a: \mu < 20$$

based on  $n = 11$  observations has the value  $t = -1.85$ .

- (a) What are the degrees of freedom for this statistic?
- (b) Between what two values does the  $P$ -value of the test fall?
- (c) If you have software available, find the exact  $P$ -value.

## 7.25 Two-sided to one-sided $P$ -value

Most software gives  $P$ -values for two-sided alternatives. Explain why you cannot always divide these  $P$ -values by 2 to obtain  $P$ -values for one-sided alternatives.

## 7.26 Number of friends on Facebook

Facebook recently examined all active Facebook users (more than 10% of the global population) and determined that the average user has 190 friends. This distribution takes only integer values, so it is certainly not Normal. It is also highly skewed to the right, with a median of 100 friends.<sup>8</sup> Consider the following SRS of  $n = 30$  Facebook users from your large university.

594	60	417	120	132	176	516	319	734	8
31	325	52	63	537	27	368	11	12	190
85	165	288	65	57	81	257	24	297	148

- (a) Are these data also heavily skewed? Use graphical methods to examine the distribution. Write a short summary of your findings.
- (b) Do you think it is appropriate to use the  $t$  methods of this section to compute a 95% confidence interval for the mean number of friends that Facebook users at your large university have? Explain why or why not.
- (c) Compute the sample mean and standard deviation, the standard error of the mean, and the margin of error for 95% confidence.
- (d) Report the 95% confidence interval for  $\mu$ , the average number of friends for Facebook users at your large university.

## 7.27 Alcohol content in beer

In February 2013, two California residents filed a class-action lawsuit against Anheuser-Busch, alleging the company was watering down beers to boost profits.<sup>9</sup> They argued that because water was being added, the true alcohol content of the beer by volume is less than the advertised amount. For example, they alleged that Budweiser beer has an alcohol content by volume of 4.7% instead of the stated 5%. Several media outlets picked up on this suit and hired independent labs to test samples of Budweiser beer and find the alcohol content. Below is a summary of these tests, each done on a single can.  BUD

4.94 5.00 4.99

- (a) Even though we have a very small sample, test the null hypothesis that the alcohol content is 4.7% by volume. Do the data provide evidence against the claim of the two residents?
- (b) Construct a 95% confidence interval for the true alcohol content in Budweiser.
- (c) U.S. government standards require that the true alcohol content in all cans and bottles be within  $\pm 0.3\%$  of the advertised level. Do these tests provide strong evidence that this is the case for Budweiser beer? Explain your answer.

## 7.28 Using the Internet on a computer

The Nielsen Company reported that U.S. residents aged 18 to 24 years spend an average of 35.5 hours per month using the Internet on a computer.<sup>10</sup> You think this is quite low compared with the amount of time that students at your university spend using the Internet on a computer, and you decide to do a survey to verify this. You collect an SRS of  $n = 50$  students and obtain  $\bar{x} = 40.1$  hours with  $s = 28.4$  hours.

- (a) Report the 95% confidence interval for  $\mu$  the average number of hours per month that students at your university use the Internet on a computer.
- (b) Use this interval to test whether the average time for students at your university is different from the average reported by Nielsen. Use the 5% significance level. Summarize your results.

## 7.29 Rudeness and its effect on onlookers

Many believe that an uncivil environment has a negative effect on people. A pair of researchers performed a series of experiments to test whether witnessing rudeness and disrespect affects task performance.<sup>11</sup> In one study, 34 participants met in small groups and witnessed the group organizer being rude to a “participant” who showed up late for the group meeting. After the exchange, each participant performed an individual brainstorming task in which he or she was asked to produce as many uses for a brick as possible in 5 minutes. The mean number of uses was 7.88 with a standard deviation of 2.35.

- (a) Suppose that prior research has shown that the average number of uses a person can produce in 5 minutes under normal conditions is 10. Given that the researchers hypothesize that witnessing this rudeness will decrease performance, state the appropriate null and alternative hypotheses.
- (b) Carry out the significance test using a significance level of 0.05. Give the  $P$ -value and state your conclusion.

## 7.30 Fuel efficiency $t$ test

Computers in some vehicles calculate various quantities related to performance. One of these is the fuel efficiency, or gas mileage, usually expressed as miles per gallon (mpg). For one vehicle equipped in this way, the miles per gallon were recorded each time the gas tank was filled, and the computer was then reset.<sup>12</sup> Here are the mpg values for a random sample of 20 of these records:  MPG

41.5	50.7	36.6	37.3	34.2	45.0	48.0	43.2	47.7	42.2
43.2	44.6	48.4	46.4	46.8	39.2	37.3	43.5	44.3	43.3

- (a) Describe the distribution using graphical methods. Is it appropriate to analyze these data using methods based on Normal distributions? Explain why or why not.
- (b) Find the mean, standard deviation, standard error, and margin of error for 95% confidence.
- (c) Report the 95% confidence interval for  $\mu$ , the mean miles per gallon for this vehicle based on these

data.

### 7.31 Tree diameter confidence interval

A study of 584 longleaf pine trees in the Wade Tract in Thomas County, Georgia, is described in Example 6.1 (page 352). For each tree in the tract, the researchers measured the diameter at breast height (DBH). This is the diameter of the tree at a height of 4.5 feet, and the units are centimeters (cm). Only trees with DBH greater than 1.5 cm were sampled. Here are the diameters of a random sample of 40 of these trees:



10.5	13.3	26.0	18.3	52.2	9.2	26.1	17.6	40.5	31.8
47.2	11.4	2.7	69.3	44.4	16.9	35.7	5.4	44.2	2.2
4.3	7.8	38.1	2.2	11.4	51.5	4.9	39.7	32.6	51.8
43.6	2.3	44.6	31.5	40.3	22.3	43.3	37.5	29.1	27.9

- Use a histogram or stemplot and a boxplot to examine the distribution of DBHs. Include a Normal quantile plot if you have the necessary software. Write a careful description of the distribution.
- Is it appropriate to use the methods of this section to find a 95% confidence interval for the mean DBH of all trees in the Wade Tract? Explain why or why not.
- Report the mean with the margin of error and the confidence interval. Write a short summary describing the meaning of the confidence interval.
- Do you think these results would apply to other similar trees in the same area? Give reasons for your answer.

### 7.32 Nutritional intake among Canadian high-performance male athletes

Recall Exercise 6.72 (page 393). For one part of the study,  $n = 114$  male athletes from eight Canadian sports centers were surveyed. Their average caloric intake was 3077.0 kilocalories per day (kcal/d) with a standard deviation of 987.0. The recommended amount is 3421.7. Is there evidence that Canadian high-performance male athletes are deficient in their caloric intake?

- State the appropriate  $H_0$  and  $H_a$  to test this.
- Carry out the test, give the  $P$ -value, and state your conclusion.
- Construct a 95% confidence interval for the average deficiency in caloric intake.

### 7.33 The return-trip effect

We often feel that the return trip from a destination takes less time than the trip to the destination even though the distance traveled is usually identical. To better understand this effect, a group of researchers ran a series of experiments.<sup>13</sup> In one experiment, they surveyed 69 participants who had just returned from a day trip by bus. Each was asked to rate how long the return trip had taken, compared with the initial trip, on an 11-point scale from  $-5 =$  a lot shorter to  $5 =$  a lot longer. The sample mean was  $-0.55$  and the sample standard deviation was 2.16.

- These data are integer values. Do you think we can still use the  $t$ -based methods of this section? Explain your answer.
- Is there evidence that the mean rating is different from zero? Carry out the significance test using  $\alpha =$

0.05 and summarize the results.

### 7.34 Stress levels in parents of children with ADHD

In a study of parents who have children with attention-deficit/hyperactivity disorder (ADHD), parents were asked to rate their overall stress level using the Parental Stress Scale (PSS).<sup>14</sup> This scale has 18 items that contain statements regarding both positive and negative aspects of parenthood. Respondents are asked to rate their agreement with each statement using a 5-point scale (1 = strongly disagree to 5 = strongly agree). The scores are summed such that a higher score indicates greater stress. The mean rating for the 50 parents in the study was reported as 52.98 with a standard deviation of 10.34.

- Do you think that these data are approximately Normally distributed? Explain why or why not.
- Is it appropriate to use the methods of this section to compute a 90% confidence interval? Explain why or why not.
- Find the 90% margin of error and the corresponding confidence interval. Write a sentence explaining the interval and the meaning of the 90% confidence level.
- To recruit parents for the study, the researchers visited a psychiatric outpatient service in Rohtak, India, and selected 50 consecutive families who met the inclusion and exclusion criteria. To what extent do you think the results can be generalized to all parents with children who have ADHD in India or in other locations around the world?

### 7.35 Are the parents feeling extreme stress?

Refer to the previous exercise. The researchers considered a score greater than 45 to represent extreme stress. Is there evidence that the average stress level for the parents in this study is above this level? Perform a test of significance using  $\alpha = 0.10$  and summarize your results.

### 7.36 Food intake and weight gain

If we increase our food intake, we generally gain weight. Nutrition scientists can calculate the amount of weight gain that would be associated with a given increase in calories. In one study, 16 nonobese adults, aged 25 to 36 years, were fed 1000 calories per day in excess of the calories needed to maintain a stable body weight. The subjects maintained this diet for 8 weeks, so they consumed a total of 56,000 extra calories.<sup>15</sup> According to theory, 3500 extra calories will translate into a weight gain of 1 pound. Therefore, we expect each of these subjects to gain  $56,000/3500 = 16$  pounds (lb). Here are the weights before and after the 8-week period, expressed in kilograms (kg):  **WTGAIN**

Subject	1	2	3	4	5	6	7	8
Weight before	55.7	54.9	59.6	62.3	74.2	75.6	70.7	53.3
Weight after	61.7	58.8	66.0	66.2	79.0	82.3	74.3	59.3
Subject	9	10	11	12	13	14	15	16
Weight before	73.3	63.4	68.1	73.7	91.7	55.9	61.7	57.8
Weight after	79.1	66.0	73.4	76.9	93.1	63.0	68.2	60.3

- For each subject, subtract the weight before from the weight after to determine the weight change.
- Find the mean and the standard deviation for the weight change.
- Calculate the standard error and the margin of error for 95% confidence. Report the 95% confidence

interval for weight change in a sentence that explains the meaning of the 95%.

- (d) Convert the mean weight gain in kilograms to mean weight gain in pounds. Because there are 2.2 kg per pound, multiply the value in kilograms by 2.2 to obtain pounds. Do the same for the standard deviation and the confidence interval.
- (e) Test the null hypothesis that the mean weight gain is 16 lb. Be sure to specify the null and alternative hypotheses, the test statistic with degrees of freedom, and the  $P$ -value. What do you conclude?
- (f) Write a short paragraph explaining your results.

### 7.37 Food intake and NEAT

Nonexercise activity thermogenesis (NEAT) provides a partial explanation for the results you found in the previous analysis. NEAT is energy burned by fidgeting, maintenance of posture, spontaneous muscle contraction, and other activities of daily living. In the study of the previous exercise, the 16 subjects increased their NEAT by 328 calories per day, on average, in response to the additional food intake. The standard deviation was 256.

- (a) Test the null hypothesis that there was no change in NEAT versus the two-sided alternative. Summarize the results of the test and give your conclusion.
- (b) Find a 95% confidence interval for the change in NEAT. Discuss the additional information provided by the confidence interval that is not evident from the results of the significance test.

### 7.38 Potential insurance fraud?

Insurance adjusters are concerned about the high estimates they are receiving from Jocko's Garage. To see if the estimates are unreasonably high, each of 10 damaged cars was taken to Jocko's and to another garage and the estimates (in dollars) were recorded. Here are the results:  **JOCKO**

Car	1	2	3	4	5
Jocko's	1410	1550	1250	1300	900
Other	1250	1300	1250	1200	950
Car	6	7	8	9	10
Jocko's	1520	1750	3600	2250	2840
Other	1575	1600	3380	2125	2600

- (a) For each car, subtract the estimate of the other garage from Jocko's estimate. Find the mean and the standard deviation for this difference.
- (b) Test the null hypothesis that there is no difference between the estimates of the two garages. Be sure to specify the null and alternative hypotheses, the test statistic with degrees of freedom, and the  $P$ -value. What do you conclude using the 0.05 significance level?
- (c) Construct a 95% confidence interval for the difference in estimates.
- (d) The insurance company is considering seeking repayment from 1000 claims filed with Jocko's last year. Using your answer to part (c), what repayment would you recommend the insurance company seek? Explain your answer.

### 7.39 Fuel efficiency comparison $t$ test

Refer to Exercise 7.30. In addition to the computer calculating miles per gallon, the driver also recorded the miles per gallon by dividing the miles driven by the number of gallons at fill-up. The driver wants to determine if these calculations are different.  **MPGDIFF**

Fill-up	1	2	3	4	5	6	7	8	9	10
Computer	41.5	50.7	36.6	37.3	34.2	45.0	48.0	43.2	47.7	42.2
Driver	36.5	44.2	37.2	35.6	30.5	40.5	40.0	41.0	42.8	39.2
Fill-up	11	12	13	14	15	16	17	18	19	20
Computer	43.2	44.6	48.4	46.4	46.8	39.2	37.3	43.5	44.3	43.3
Driver	38.8	44.5	45.4	45.3	45.7	34.2	35.2	39.8	44.9	47.5

- (a) State the appropriate  $H_0$  and  $H_a$
- (b) Carry out the test using a significance level of 0.05. Give the  $P$ -value, and then interpret the result.

## 7.40 Counts of picks in a one-pound bag

A guitar supply company must maintain strict oversight on the number of picks they package for sale to customers. Their current advertisement specifies between 900 and 1000 picks in every bag. An SRS of 36 one-pound bags of picks was collected as part of a quality improvement effort within the company. The number of picks in each bag is shown in the following table.  **PICKS**

924925967909959937970936952  
 919965921913886956962916945  
 957912961950923935969916952  
 917977940924957920986895923

- (a) Create (i) a histogram or a stemplot, (ii) a boxplot, and (iii) a Normal quantile plot of these counts. Write a careful description of the distribution. Make sure to note any outliers, and comment on the skewness and Normality of the data.
- (b) Based on your observations in part (a), is it appropriate to analyze these data using the  $t$  procedures? Briefly explain your response.
- (c) Find the mean, the standard deviation, and the standard error of the mean for this sample.
- (d) Calculate the 90% confidence interval for the mean number of picks in a one-pound bag.

## 7.41 Significance test for the average number of picks

Refer to the previous exercise.

- (a) Do these data provide evidence that the average number of picks in a one-pound bag is greater than 925? Using a significance level of 5%, state your hypotheses, the  $P$ -value, and your conclusions.
- (b) Do these data provide evidence that the average number of picks in a one-pound bag is greater than 935? Using a significance level of 5%, state your hypotheses, the  $P$ -value, and your conclusion.
- (c) Explain the relationship between your conclusions in parts (a) and (b) and the 90% confidence interval calculated in the previous problem.

## 7.42 A customer satisfaction survey

Many organizations are doing surveys to determine the satisfaction of their customers. Attitudes toward various aspects of campus life were the subject of one such study conducted at Purdue University. Each item was rated on a 1 to 5 scale, with 5 being the highest rating. The average response of 2368 first-year students to “Feeling welcomed at Purdue” was 3.92 with a standard deviation of 1.02. Assuming that the respondents are an SRS, give a 90% confidence interval for the mean of all first-year students.

### 7.43 Comparing operators of a DXA machine

Dual-energy X-ray absorptiometry (DXA) is a technique for measuring bone health. One of the most common measures is total body bone mineral content (TBBMC). A highly skilled operator is required to take the measurements. Recently, a new DXA machine was purchased by a research lab, and two operators were trained to take the measurements. TBBMC for eight subjects was measured by both operators.<sup>16</sup> The units are grams (g). A comparison of the means for the two operators provides a check on the training they received and allows us to determine if one of the operators is producing measurements that are consistently higher than the other. Here are the data:  TBBMC

Operator	Subject							
	1	2	3	4	5	6	7	8
1	1.328	1.342	1.075	1.228	0.939	1.004	1.178	1.286
2	1.323	1.322	1.073	1.233	0.934	1.019	1.184	1.304

- (a) Take the difference between the TBBMC recorded for Operator 1 and the TBBMC for Operator 2. Describe the distribution of these differences. Is it appropriate to analyze these data using the  $t$  methods? Explain why or why not.
- (b) Use a significance test to examine the null hypothesis that the two operators have the same mean. Be sure to give the test statistic with its degrees of freedom, the  $P$ -value, and your conclusion.
- (c) The sample here is rather small, so we may not have much power to detect differences of interest. Use a 95% confidence interval to provide a range of differences that are compatible with these data.
- (d) The eight subjects used for this comparison were not a random sample. In fact, they were friends of the researchers whose ages and weights were similar to these of the types of people who would be measured with this DXA machine. Comment on the appropriateness of this procedure for selecting a sample, and discuss any consequences regarding the interpretation of the significance-testing and confidence interval results.

### 7.44 Another comparison of DXA machine operators

Refer to the previous exercise. TBBMC measures the total amount of mineral in the bones. Another important variable is total body bone mineral density (TBBMD). This variable is calculated by dividing TBBMC by the area corresponding to bone in the DXA scan. The units are grams per squared centimeter ( $\text{g}/\text{cm}^2$ ). Here are the TBBMD values for the same subjects:  TBBMD

Operator	Subject							
	1	2	3	4	5	6	7	8
1	4042	3703	2626	2673	1724	2136	2808	3322
2	4041	3697	2613	2628	1755	2140	2836	3287

Analyze these data using the questions in the previous exercise as a guide.

## 7.45 Assessment of a foreign-language institute

The National Endowment for the Humanities sponsors summer institutes to improve the skills of high school teachers of foreign languages. One such institute hosted 20 French teachers for 4 weeks. At the beginning of the period, the teachers were given the Modern Language Association's listening test of understanding of spoken French. After 4 weeks of immersion in French in and out of class, the listening test was given again. (The actual French spoken in the two tests was different, so that simply taking the first test should not improve the score on the second test.) The maximum possible score on the test is 36.<sup>17</sup> Here are the data:  **SUMLANG**

Teacher	Pretest	Posttest	Gain	Teacher	Pretest	Posttest	Gain
1	32	34	2	11	30	36	6
2	31	31	0	12	20	26	6
3	29	35	6	13	24	27	3
4	10	16	6	14	24	24	0
5	30	33	3	15	31	32	1
6	33	36	3	16	30	31	1
7	22	24	2	17	15	15	0
8	25	28	3	18	32	34	2
9	32	26	-6	19	23	26	3
10	20	26	6	20	23	26	3

To analyze these data, we first subtract the pretest score from the posttest score to obtain the improvement for each teacher. These 20 differences form a single sample. They appear in the "Gain" columns. The first teacher, for example, improved from 32 to 34, so the gain is  $34 - 32 = 2$

- State appropriate null and alternative hypotheses for examining the question of whether or not the course improves French spoken-language skills.
- Describe the gain data. Use numerical and graphical summaries.
- Perform the significance test. Give the test statistic, the degrees of freedom, and the  $P$ -value. Summarize your conclusion.
- Give a 95% confidence interval for the mean improvement.

## 7.46 Length of calls to a customer service center

Refer to the lengths of calls to a customer service center in Table 1.2 (page 19). Give graphical and numerical summaries for these data. Compute a 95% confidence interval for the mean call length. Comment on the validity of your interval.  **CALLS80**

## 7.47 Sign test for potential insurance fraud

The differences in the repair estimates in Exercise 7.38 can also be analyzed using a sign test. Set up the appropriate null and alternative hypotheses, carry out the test, and summarize the results. How do these results compare with those that you obtained in Exercise 7.38?  **JOCKO**

## 7.48 Sign test for the comparison of operators

The differences in the TBBMC measures in Exercise 7.43 can also be analyzed using a sign test. Set up the appropriate null and alternative hypotheses, carry out the test, and summarize the results. How do these results compare with those that you obtained in Exercise 7.43?  **TBBMC**

### 7.49 Another sign test for the comparison of operators

TBBMD values for the same subjects that you studied in the previous exercise are given in Exercise 7.44. Answer the questions given in the previous exercise for TBBMD.  **TBBMD**

### 7.50 Sign test for assessment of a foreign-language institute

Use the sign test to assess whether the summer institute of Exercise 7.45 improves French listening skills. State the hypotheses, give the  $P$ -value using the binomial table (Table C), and report your conclusion.  **SUMLANG**

### 7.51 Sign test for fuel efficiency comparison

Use the sign test to assess whether the computer calculates a higher mpg than the driver in Exercise 7.39. State the hypotheses, give the  $P$ -value using the binomial table (Table C), and report your conclusion.  **MPGDIFF**

### 7.52 Insulation study

A manufacturer of electric motors tests insulation at a high temperature (250°C) and records the number of hours until the insulation fails.<sup>18</sup> The data for 5 specimens are

446 326 372 377 310

The small sample size makes judgment from the data difficult, but engineering experience suggests that the logarithm of the failure time will have a Normal distribution. Take the logarithms of the 5 observations, and use  $t$  procedures to give a 90% confidence interval for the mean of the log failure time for insulation of this type.  **INSULAT**

### 7.53 Power of the comparison of DXA machine operators

Suppose that the bone researchers in Exercise 7.43 wanted to be able to detect an alternative mean difference of 0.002. Find the power for this alternative for a sample size of 15. Use the standard deviation that you found in Exercise 7.43 for these calculations.

### 7.54 Sample size calculations

You are designing a study to test the null hypothesis that  $\mu = 0$  versus the alternative that  $\mu$  is positive. Assume that  $\sigma$  is 15. Suppose that it would be important to be able to detect the alternative  $\mu = 2$ . Perform power calculations for a variety of sample sizes and determine how large a sample you would need to detect this alternative with power of at least 0.80.

### 7.55 Determining the sample size

Consider Example 7.9 (page 435). What is the minimum sample size needed for the power to be greater than 80% when  $\mu = 0.75$ ?

## 7.2 Comparing Two Means

When you complete this section, you will be able to

- Describe a level  $C$  confidence interval for the difference between two population means in terms of an estimate and its margin of error.
- Construct a level  $C$  confidence interval for the difference between two population means  $\mu_1 - \mu_2$  from two SRSs of size  $n_1$  and  $n_2$ , respectively.
- Perform a two-sample  $t$  significance test and summarize the results.
- Explain when the  $t$  procedures can be useful for non-Normal data.

A psychologist wants to compare male and female college students' impressions of personality based on selected Facebook pages. A nutritionist is interested in the effect of increased calcium on blood pressure. A bank wants to know which of two incentive plans will most increase the use of its debit cards. Two-sample problems such as these are among the most common situations encountered in statistical practice.

### TWO-SAMPLE PROBLEMS

- The goal of inference is to compare the means of the response variable in two groups.
- Each group is considered to be a sample from a distinct population.
- The responses in each group are independent of those in the other group.



randomized comparative experiment, p. 180

A two-sample problem can arise from a randomized comparative experiment that randomly divides the subjects into two groups and exposes each group to a different treatment. A two-sample problem can also arise when comparing random samples separately selected from two populations. Unlike the matched pairs designs studied earlier, there is no matching of the units in the two samples, and the two samples may be of different sizes. As a result, inference procedures for two-sample data differ from those for matched pairs.

We can present two-sample data graphically by a back-to-back stemplot (for

small samples) or by side-by-side boxplots (for larger samples). Now we will apply the ideas of formal inference in this setting. When both population distributions are symmetric, and especially when they are at least approximately Normal, a comparison of the mean responses in the two populations is most often the goal of inference.

We have two independent samples, from two distinct populations (such as subjects given a treatment and those given a placebo). The same response variable is measured for both samples. We will call the variable  $x_1$  in the first population and  $x_2$  in the second because the variable may have different distributions in the two populations. Here is the notation that we will use to describe the two populations:

Population	Variable	Mean	Standard deviation
1	$x_1$	$\mu_1$	$\sigma_1$
2	$x_2$	$\mu_2$	$\sigma_2$

We want to compare the two population means, either by giving a confidence interval for  $\mu_1 - \mu_2$  or by testing the hypothesis of no difference,  $H_0: \mu_1 = \mu_2$

Inference is based on two independent SRSs, one from each population. Here is the notation that describes the samples:

Population	Sample size	Sample mean	Sample standard deviation
1	$n_1$	$\bar{x}_1$	$s_1$
2	$n_2$	$\bar{x}_2$	$s_2$

Throughout this section, the subscripts 1 and 2 show the population to which a parameter or a sample statistic refers.

## The two-sample $z$ statistic

The natural estimator of the difference  $\mu_1 - \mu_2$  is the difference between the sample means,  $\bar{x}_1 - \bar{x}_2$ . If we are to base inference on this statistic, we must know its sampling distribution. First, the mean of the difference  $\bar{x}_1 - \bar{x}_2$  is the difference between the means  $\mu_1 - \mu_2$ . This follows from the addition rule for means and the fact that the mean of any  $\bar{x}$  is the mean of the population. Second, to compute its variance, we use the addition rule for variances. Because the samples are independent, their sample means  $\bar{x}_1$  and  $\bar{x}_2$  are independent random variables. Thus, the variance of the difference  $\bar{x}_1 - \bar{x}_2$  is the sum of their variances, which is



rules for means, p. 272



rules for variances, p. 275

$$\sigma_1^2/n_1 + \sigma_2^2/n_2$$

We now know the mean and variance of the distribution of  $\bar{x}_1 - \bar{x}_2$  in terms of the parameters of the two populations. If the two population distributions are both Normal, then the distribution of  $\bar{x}_1 - \bar{x}_2$  is also Normal. This is true because each sample mean alone is Normally distributed and because a difference between independent Normal random variables is also Normal.

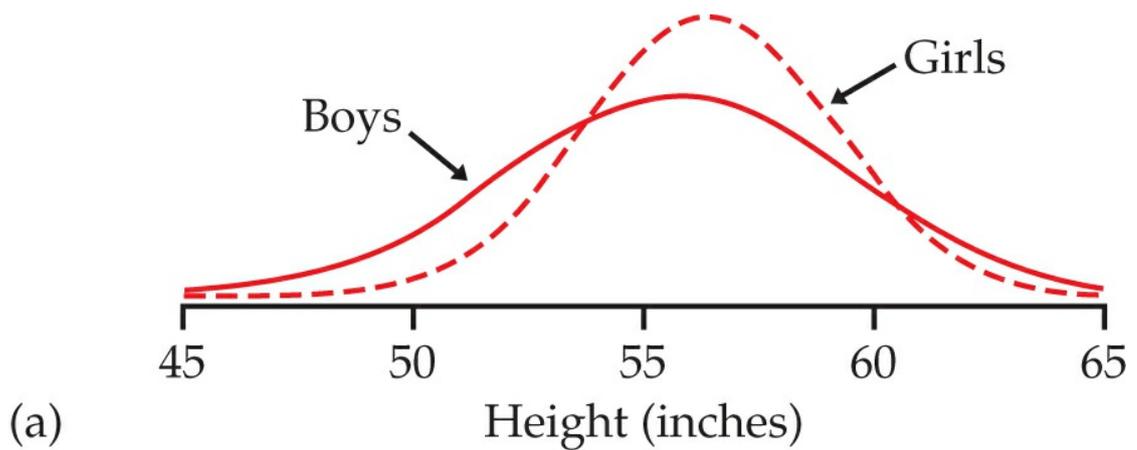
## Example

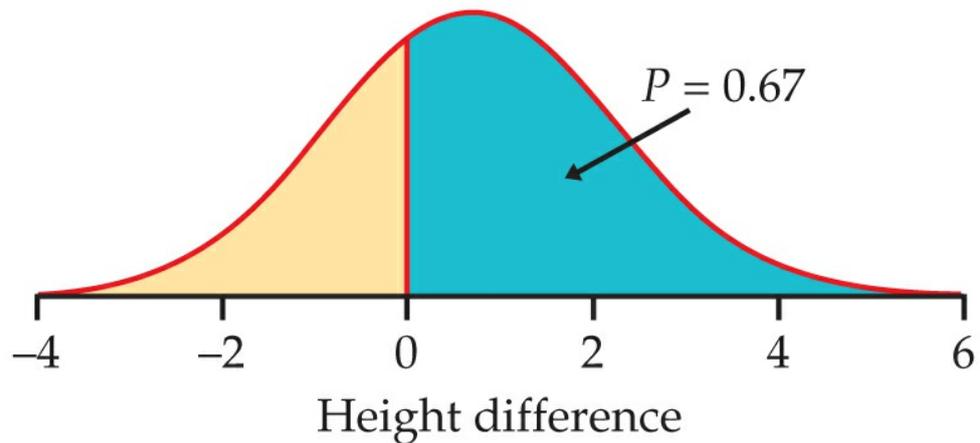
### 7.13 Heights of 10-year-old girls and boys

A fourth-grade class has 12 girls and 8 boys. The children's heights are recorded on their 10th birthdays. What is the chance that the girls are taller than the boys? Of course, it is very unlikely that all the girls are taller than all the boys. We translate the question into the following: what is the probability that the mean height of the girls is greater than the mean height of the boys?



Based on information from the National Health and Nutrition Examination Survey, we assume that the heights (in inches) of 10-year-old girls are  $N(56.4, 2.7)$  and the heights of 10-year-old boys are  $N(55.7, 3.8)$ <sup>19</sup> The heights of the students in our class are assumed to be random samples from these populations. The two distributions are shown in Figure 7.13(a).





**FIGURE 7.13**

Distributions for Example 7.13. (a) Distributions of heights of 10-year-old boys and girls.  
 (b) Distribution of the difference between the mean heights of 12 girls and 8 boys.

The difference  $\bar{x}_1 - \bar{x}_2$  between the female and male mean heights varies in different random samples. The sampling distribution has mean

$$\mu_1 - \mu_2 = 56.4 - 55.7 = 0.7 \text{ inch}$$

and variance

$$\begin{aligned} \sigma^2_{\bar{x}_1 - \bar{x}_2} &= \sigma_1^2/n_1 + \sigma_2^2/n_2 = 2.72/12 + 3.828/8 \\ &= 2.41 \end{aligned}$$

The standard deviation of the difference in sample means is therefore  $\sqrt{2.41} = 1.55$

If the heights vary Normally, the difference in sample means is also Normally distributed. The distribution of the difference in heights is shown in Figure 7.13(b). We standardize  $\bar{x}_1 - \bar{x}_2$  by subtracting its mean (0.7) and dividing by its standard deviation (1.55). Therefore, the probability that the girls, on average, are taller than the boys is

$$\begin{aligned} P(\bar{x}_1 - \bar{x}_2 > 0) &= P((\bar{x}_1 - \bar{x}_2) - 0.7 / 1.55 > 0 - 0.7 / 1.55) \\ &= P(Z > -0.45) = 0.6736 \end{aligned}$$

Even though the population mean height of 10-year-old girls is greater than the population mean height of 10-year-old boys, the probability that the sample mean of the girls is greater than the sample mean of the boys in our class is only 67%. *Large samples are needed to see the effects of small differences.*



As Example 7.13 reminds us, any Normal random variable has the  $N(0, 1)$

distribution when standardized. We have arrived at a new  $z$  statistic.

## TWO-SAMPLE $z$ STATISTIC

Suppose that  $\bar{x}_1$  is the mean of an SRS of size  $n_1$  drawn from an  $N(\mu_1, \sigma_1)$  population and that  $\bar{x}_2$  is the mean of an independent SRS of size  $n_2$  drawn from an  $N(\mu_2, \sigma_2)$  population. Then the **two-sample  $z$  statistic**

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has the standard Normal  $N(0, 1)$  sampling distribution.

In the unlikely event that both population standard deviations are known, the two-sample  $z$  statistic is the basis for inference about  $\mu_1 - \mu_2$ . Exact  $z$  procedures are seldom used, however, because  $\sigma_1$  and  $\sigma_2$  are rarely known. In Chapter 6, we discussed the one-sample  $z$  procedures in order to introduce the ideas of inference. Here we move directly to the more useful  $t$  procedures.

## The two-sample $t$ procedures

Suppose now that the population standard deviations  $\sigma_1$  and  $\sigma_2$  are not known. We estimate them by the sample standard deviations  $s_1$  and  $s_2$  from our two samples. Following the pattern of the one-sample case, we substitute the standard errors for the standard deviations used in the two-sample  $z$  statistic. The result is the *two-sample  $t$  statistic*:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Unfortunately, this statistic does *not* have a  $t$  distribution. A  $t$  distribution replaces the  $N(0, 1)$  distribution only when a single standard deviation ( $\sigma$ ) in a  $z$  statistic is replaced by its sample standard deviation ( $s$ ). In this case, we replace two standard deviations ( $\sigma_1$  and  $\sigma_2$ ) by their estimates ( $s_1$  and  $s_2$ ), which does not produce a statistic having a  $t$  distribution.

Nonetheless, we can approximate the distribution of the two-sample  $t$  statistic by using the  $t(k)$  distribution with an **approximation for the degrees of freedom  $k$** . We use these approximations to find approximate values of  $t^*$  for confidence intervals and to find approximate  $P$ -values for significance tests. Here are two approximations:

### approximations for the degrees of freedom

1. Use a value of  $k$  that is calculated from the data. In general, it will not be a whole number.

2. Use  $k$  equal to the smaller of  $n_1 - 1$  and  $n_2 - 1$

In practice, the choice of approximation rarely makes a difference in our conclusion. Most statistical software uses the first option to approximate the  $t(k)$  distribution for two-sample problems unless the user requests another method. Use of this approximation without software is a bit complicated; we will give the details later in this section (see page 460).

If you are not using software, the second approximation is preferred. This approximation is appealing because it is conservative.<sup>20</sup> Margins of error for the level  $C$  confidence intervals are a bit larger than they need to be, so the true confidence level is larger than  $C$ . For significance testing, the true  $P$ -values are a bit smaller than those we obtain from this approximation; thus, for tests at a fixed significance level, we are a little less likely to reject  $H_0$  when it is true.

## The two-sample $t$ confidence interval

### THE TWO-SAMPLE $t$ CONFIDENCE INTERVAL

Suppose that an SRS of size  $n_1$  is drawn from a Normal population with unknown mean  $\mu_1$  and that an independent SRS of size  $n_2$  is drawn from another Normal population with unknown mean  $\mu_2$ . The **confidence interval for  $\mu_1 - \mu_2$**  given by

$$(\bar{x}_1 - \bar{x}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

has confidence level at least  $C$  no matter what the population standard deviations may be. Here,  $t^*$  is the value for the  $t(k)$  density curve with area  $C$  between  $-t^*$  and  $t^*$ . The value of the degrees of freedom  $k$  is approximated by software or we use the smaller of  $n_1 - 1$  and  $n_2 - 1$ . Similarly, we can use either software or the conservative approach with Table D to approximate the value of  $t^*$ .

### EXAMPLE

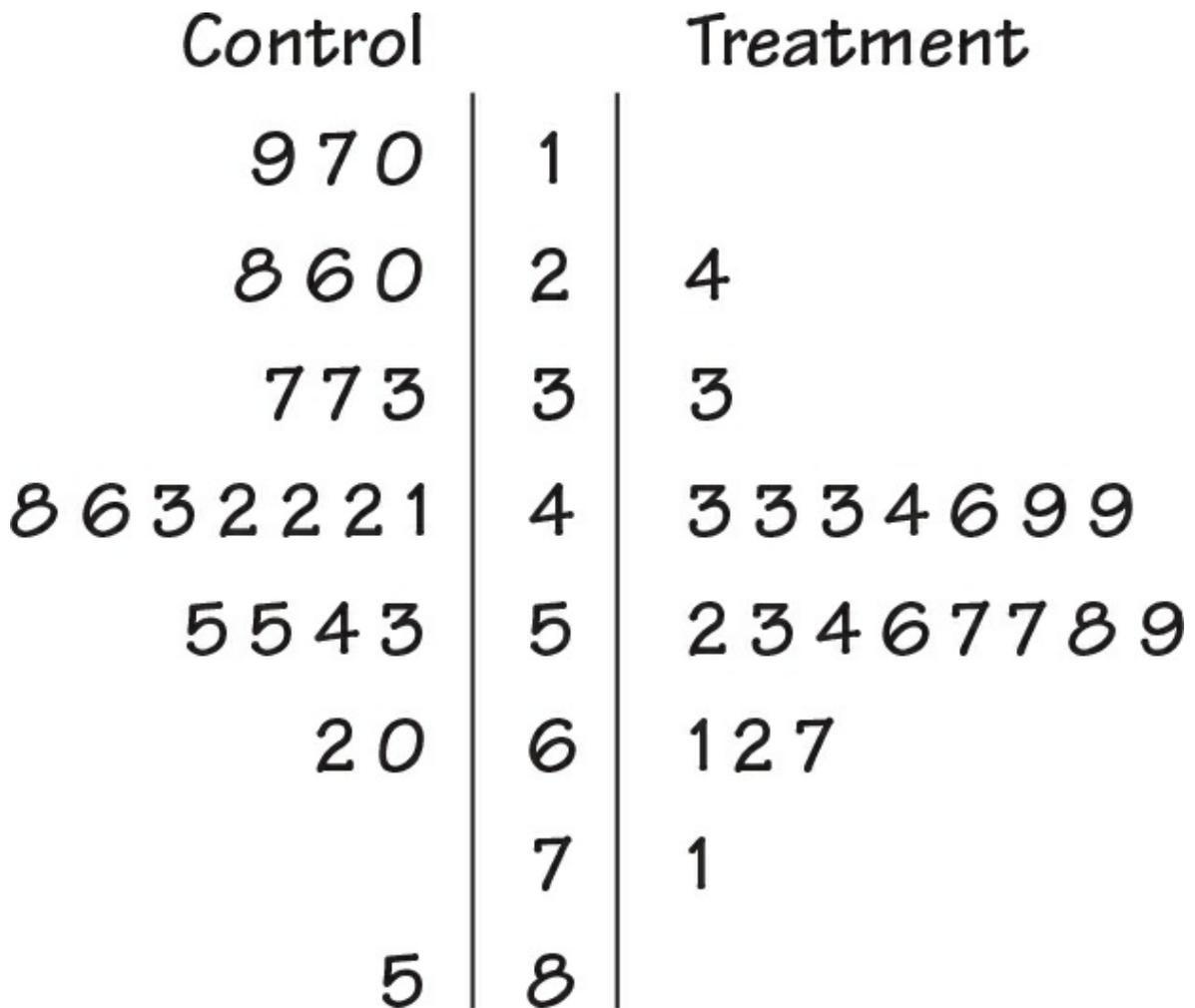
#### 7.14 Directed reading activities assessment

An educator believes that new directed reading activities in the classroom will

help elementary school pupils improve some aspects of their reading ability. She arranges for a third-grade class of 21 students to take part in these activities for an eight-week period. A control classroom of 23 third-graders follows the same curriculum without the activities. At the end of the eight weeks, all students are given a Degree of Reading Power (DRP) test, which measures the aspects of reading ability that the treatment is designed to improve. The data appear in Table 7.4.<sup>21</sup>



First examine the data:

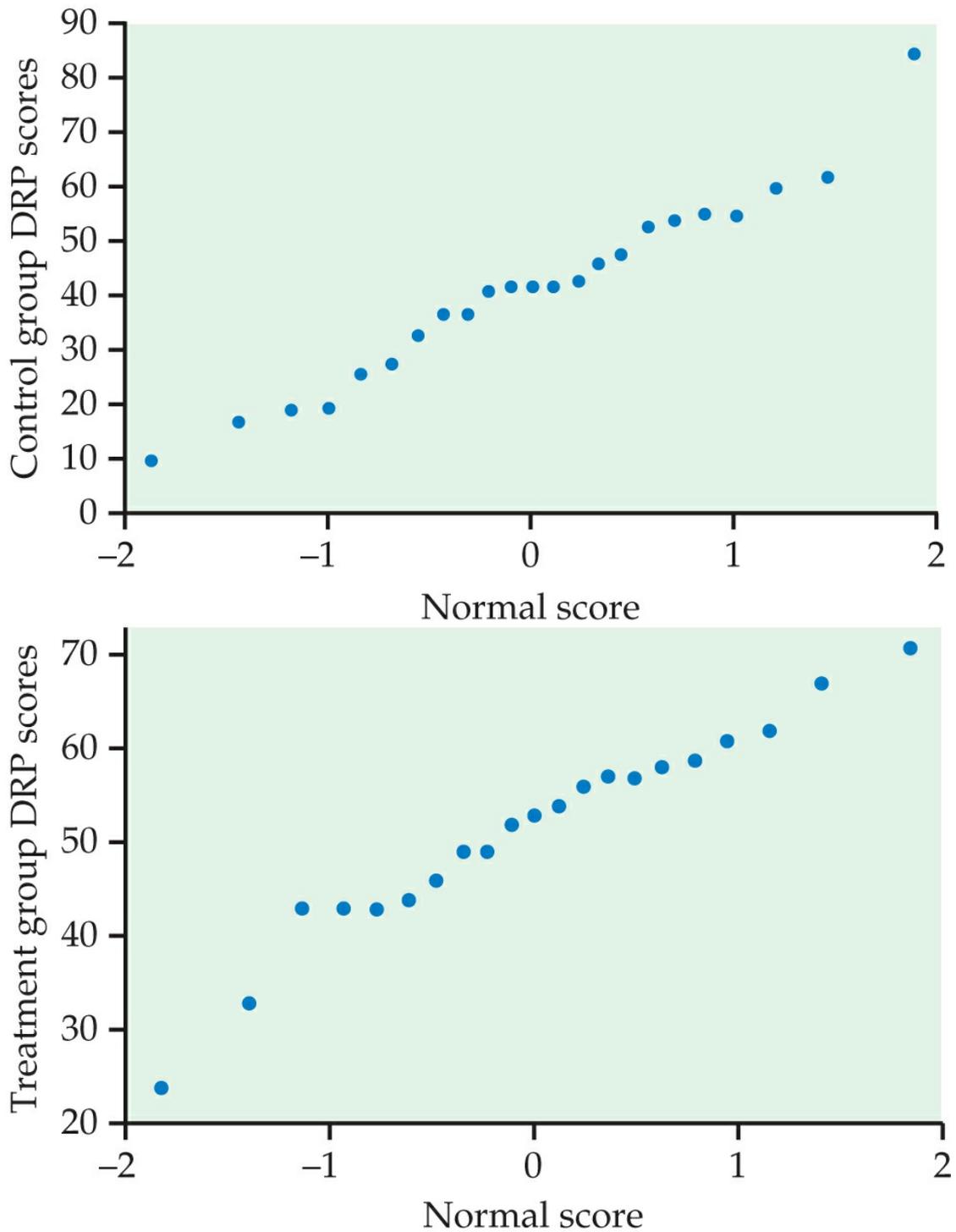


The back-to-back stemplot suggests that there is a mild outlier in the control group but no deviation from Normality serious enough to forbid use of  $t$  procedures. Separate Normal quantile plots for both groups (Figure 7.14) confirm that both distributions are approximately Normal. The scores of the treatment group appear to be somewhat higher than those of the control group. The summary statistics are

Group	$n$	$\bar{x}$	$s$
Treatment	21	51.48	11.01
Control	23	41.52	17.15

**TABLE 7.4**      **DRP Scores for Third-Graders**

Treatment group				Control group			
24	61	59	46	42	33	46	37
43	44	52	43	43	41	10	42
58	67	62	57	55	19	17	55
71	49	54		26	54	60	28
43	53	57		62	20	53	48



**FIGURE 7.14**

Normal quantile plots of the DRP scores in Table 7.4.

To describe the size of the treatment effect, let's construct a confidence interval for the difference between the treatment group and the control group means. The interval is

$$(\bar{x}_1 - \bar{x}_2) \pm t^* s_1 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (51.48 - 41.52) \pm t^* \sqrt{11.01221 + 17.15223}$$

$$= 9.96 \pm 4.31t^*$$

Using software, the degrees of freedom are 37.9 and  $t^* = 2.025$ . This approximation gives

$$9.96 \pm (4.31 \times 2.025) = 9.96 \pm 8.72 = (1.2, 18.7)$$

The conservative approach uses the  $t(20)$  distribution. Table D gives  $t^* = 2.086$ . With this approximation we have

$$9.96 \pm (4.31 \times 2.086) = 9.96 \pm 8.99 = (1.0, 18.9)$$

We see that the conservative approach does, in fact, give a wider interval than the more accurate approximation used by software. However, the difference is pretty small.

We estimate the mean improvement to be about 10 points, but with a margin of error of almost 9 points with either method. Unfortunately, the data do not allow a very precise estimate of the size of the average improvement.

## USE YOUR KNOWLEDGE

### 7.56 Two-sample $t$ confidence interval

Assume that  $\bar{x}_1=110$ ,  $\bar{x}_2=120$ ,  $s_1 = 8$ ,  $s_2 = 12$ ,  $n_1 = 50$ , and  $n_2 = 50$ . Find a 95% confidence interval for the difference in the corresponding values of  $\mu$  using the second approximation for degrees of freedom. Does this interval include more or fewer values than a 99% confidence interval would? Explain your answer.

### 7.57 Another two-sample $t$ confidence interval

Assume that  $\bar{x}_1=110$ ,  $\bar{x}_2=120$ ,  $s_1 = 8$ ,  $s_2 = 12$ ,  $n_1 = 10$ , and  $n_2 = 10$ . Find a 95% confidence interval for the difference in the corresponding values of  $\mu$  using the second approximation for degrees of freedom. Would you reject the null hypothesis that the population means are equal in favor of the two-sided alternative at significance level 0.05? Explain.

## The two-sample $t$ significance test

The same ideas that we used for the two-sample  $t$  confidence interval also apply to

*two-sample t significance tests.* We can use either software or the conservative approach with Table D to approximate the  $P$ -value.

## THE TWO-SAMPLE $t$ SIGNIFICANCE TEST

Suppose that an SRS of size  $n_1$  is drawn from a Normal population with unknown mean  $\mu_1$  and that an independent SRS of size  $n_2$  is drawn from another Normal population with unknown mean  $\mu_2$ . To test the hypothesis  $H_0: \mu_1 = \mu_2$  compute the **two-sample  $t$  statistic**

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

and use  $P$ -values or critical values for the  $t(k)$  distribution, where the degrees of freedom  $k$  either are approximated by software or are the smaller of  $n_1 - 1$  and  $n_2 - 1$ .

### Example

#### 7.15 Is there an improvement?

For the DRP study described in Example 7.14, we hope to show that the treatment (Group 1) performs better than the control (Group 2). For a formal significance test the hypotheses are

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 > \mu_2$$

The two-sample  $t$  test statistic is

$$\begin{aligned} t &= \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \\ &= \frac{51.48 - 41.52}{\sqrt{11.0122/21 + 17.1522/23}} \\ &= 2.31 \end{aligned}$$

The  $P$ -value for the one-sided test is  $P(T \geq 2.31)$ . Software gives the approximate  $P$ -value as 0.0132 and uses 37.9 as the degrees of freedom. For the second approximation, the degrees of freedom  $k$  are equal to the smaller of

$$n_1 - 1 = 21 - 1 = 20 \text{ and } n_2 - 1 = 23 - 1 = 22$$

Comparing 2.31 with the entries in Table D for 20 degrees of freedom, we see that  $P$  lies between 0.01 and 0.02. The data strongly suggest that directed reading activity improves the DRP score ( $t = 2.31$ ,  $df = 20$ ,  $0.01 < P < 0.02$ ).

If your software gives  $P$ -values for only the two-sided alternative,  $2P(T \geq |t|)$ , you need to divide the reported value by 2 after checking that the means differ in the direction specified by the alternative hypothesis.

The design of the study in Example 7.14 is not ideal. Random assignment of students was not possible in a school environment, so existing third-grade classes were used. The effect of the reading programs is therefore confounded with any other differences between the two classes. The classes were chosen to be as similar as possible—for example, in terms of the social and economic status of the students. Extensive pretesting showed that the two classes were, on the average, quite similar in reading ability at the beginning of the experiment. To avoid the effect of two different teachers, the researcher herself taught reading in both classes during the eight-week period of the experiment. We can therefore be somewhat confident that the two-sample test is detecting the effect of the treatment and not some other difference between the classes. This example is typical of many situations in which an experiment is carried out but randomization is not possible.

## USE YOUR KNOWLEDGE

### 7.58 Comparison of two web page designs

You want to compare the daily number of hits for two different MySpace page designs that advertise your indie rock band. You assign the next 30 days to either Design A or Design B, 15 days to each.

- Would you use a one-sided or a two-sided significance test for this problem? Explain your choice.
- If you use Table D to find the critical value, what are the degrees of freedom using the second approximation?
- If you perform the significance test using  $\alpha = 0.05$ , how large (positive or negative) must the  $t$  statistic be to reject the null hypothesis that the two designs result in the same average number of hits?

### 7.59 More on the comparison of two web page designs

Refer to the previous exercise. If the  $t$  statistic for comparing the mean

hits was 2.18, what  $P$ -value would you report? What would you conclude using  $\alpha = 0.05$ ?

## Robustness of the two-sample procedures

The two-sample  $t$  procedures are more robust than the one-sample  $t$  methods. When the sizes of the two samples are equal and the distributions of the two populations being compared have similar shapes, probability values from the  $t$  table are quite accurate for a broad range of distributions when the sample sizes are as small as  $n_1 = n_2 = 5$ .<sup>22</sup> When the two population distributions have different shapes, larger samples are needed.

The guidelines for the use of one-sample  $t$  procedures can be adapted to two-sample procedures by replacing “sample size” with the “sum of the sample sizes”  $n_1 + n_2$ . Specifically,

- *If  $n_1 + n_2$  is less than 15:* Use  $t$  procedures if the data are close to Normal. If the data in either sample are clearly non-Normal or if outliers are present, do not use  $t$ .
- *If  $n_1 + n_2$  is at least 15 and less than 40:* The  $t$  procedures can be used except in the presence of outliers or strong skewness.
- *Large samples:* The  $t$  procedures can be used even for clearly skewed distributions when the sample is large, roughly  $n_1 + n_2 \geq 40$ .



These guidelines are rather conservative, especially when the two samples are of equal size. *In planning a two-sample study, choose equal sample sizes if you can.* The two-sample  $t$  procedures are most robust against non-Normality in this case, and the conservative probability values are most accurate.

Here is an example with large sample sizes that are almost equal. Even if the distributions are not Normal, we are confident that the sample means will be approximately Normal. The two-sample  $t$  test is very robust in this case.

### Example

#### 7.16 Timing of food intake and weight loss

There is emerging evidence of a relationship between timing of feeding and weight regulation. In one study, researchers followed 402 obese or overweight individuals through a 20-week weight-loss treatment.<sup>23</sup> To investigate the timing of food intake, participants were grouped into early eaters and late eaters, based on the timing of their main meal. Here are the summary statistics of their weight loss over the 20 weeks, in kilograms (kg):



Group	$n$	$\bar{x}$	$s$
Early eater	202	9.9	5.8
Late eater	200	7.7	6.1

The early eaters lost more weight on average. Can we conclude that these two groups are not the same? Or is this observed difference merely what we could expect to see given the variation among participants?

While other evidence suggests that early eaters should lose more weight,

the researchers did not specify a direction for the difference. Thus, the hypotheses are

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

Because the samples are large, we can confidently use the  $t$  procedures even though we lack the detailed data and so cannot verify the Normality condition. The two-sample  $t$  statistic is

$$\begin{aligned} t &= \frac{\bar{x}_1 - \bar{x}_2}{s_1/\sqrt{n_1} + s_2/\sqrt{n_2}} \\ &= \frac{9.9 - 7.75}{\sqrt{8.22/202 + 6.12/200}} \\ &= 3.71 \end{aligned}$$

The conservative approach finds the  $P$ -value by comparing 3.71 to critical values for the  $t(199)$  distribution because the smaller sample has 200 observations. Because Table D does not contain a row for 199 degrees of freedom, we will be even more conservative and use the first row in the table with degrees of freedom less than 199. This means we'll use the  $t(100)$  distribution to compute the  $P$ -value.

Our calculated value of  $t$  is larger than the  $p = 0.0005$  entry in the table. We must double the table tail area  $p$  because the alternative is two-sided, so we conclude that the  $P$ -value is less than 0.001. The data give conclusive evidence that early eaters lost more weight, on average, than late eaters ( $t = 3.71$ ,  $df = 100$ ,  $P < 0.001$ ).

In this example the exact  $P$ -value is very small because  $t = 3.71$  says that the observed difference in means is over 3.5 standard errors above the hypothesized difference of zero ( $\mu_1 = \mu_2$ ). In this study, the researchers also compared energy intake and energy expenditure between late and early eaters. Despite the observed weight loss difference of 2.2 kg, no significant differences in these variables were found.



In this and other examples, we can choose which population to label 1 and which to label 2. After inspecting the data, we chose early eaters as Population 1 because this choice makes the  $t$  statistic a positive number. This avoids any possible confusion from reporting a negative value for  $t$ . *Choosing the population labels is **not** the same as choosing a one-sided alternative after looking at the data.* Choosing hypotheses after seeing a result in the data is a violation of sound statistical practice.

## Inference for small samples

Small samples require special care. We do not have enough observations to examine the distribution shapes, and only extreme outliers stand out. The power of significance tests tends to be low, and the margins of error of confidence intervals tend to be large. Despite these difficulties, we can often draw important conclusions from studies with small sample sizes. If the size of an effect is very large, it should still be evident even if the  $n$ 's are small.

### Example

#### 7.17 Timing of food intake

In the setting of Example 7.16, let's consider a much smaller study that collects weight loss data from only 5 participants in each eating group. Also, given the results of this past example, we choose the one-sided alternative. The data are

Group	Weight loss (kg)				
Early eater	6.3	15.1	9.4	16.8	10.2
Late eater	7.8	0.2	1.5	11.5	4.6

First, examine the distributions with a back-to-back stemplot (the data are rounded to the nearest integer).

Early		Late
	0	0 2
9 6	0	5 8
0	1	2
7 5	1	

While there is variation among weight losses within each group, there is also a noticeable separation. The early-eaters group contains 4 of the 5 greatest losses, and the late-eaters group contains 4 of the 5 lowest losses. A significance test can confirm whether this pattern can arise just by chance or if the early-eaters group has a higher mean. We test

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 > \mu_2$$

The average weight loss is higher in the early-eater group ( $t = 2.28$ ,  $df = 7.96$ ,  $P = 0.0262$ ). The difference in sample means is 6.44 kg.

SAS

**The TTEST Procedure**

Variable: loss

Grp	N	Mean	StdDev	StdErr	Minimum	Maximum
Early	5	11.5600	4.3062	1.9258	6.3000	16.8000
Late	5	5.1200	4.6224	2.0672	0.2000	11.5000
Diff (1-2)		6.4400	4.4671	2.8252		

Grp	Method	Mean	95%	CL Mean	Std Dev	95%	CL Std Dev
Early		11.5600	6.2132	16.9068	4.3062	2.5800	12.3740
Late		5.1200	-0.6195	10.8595	4.6224	2.7695	13.2829
Diff (1-2)	Pooled	6.4400	-0.0750	12.9550	4.4671	3.0173	8.5579
Diff (1-2)	Satterthwaite	6.4400	-0.0807	12.9607			

Method	Variances	DF	t Value	Pr >  t
Pooled	Equal	8	2.28	0.0521
Satterthwaite	Unequal	7.9601	2.28	0.0523

	A	B	C
1	<b>t-Test: Two-Sample Assuming Unequal Variances</b>		
2			
3		<i>Early</i>	<i>Late</i>
4	Mean	11.56	5.12
5	Variance	18.543	21.367
6	Observations	5	5
7	Hypothesized Mean Difference	0	
8	Df	8	
9	t Stat	2.27944966	
10	P(T<=t) one-tail	0.026058036	
11	t Critical one-tail	1.859548033	
12	P(T<=t) two-tail	0.052116073	
13	t Critical two-tail	2.306004133	

**JMP**

**Oneway Analysis of WTLOSS By Group**

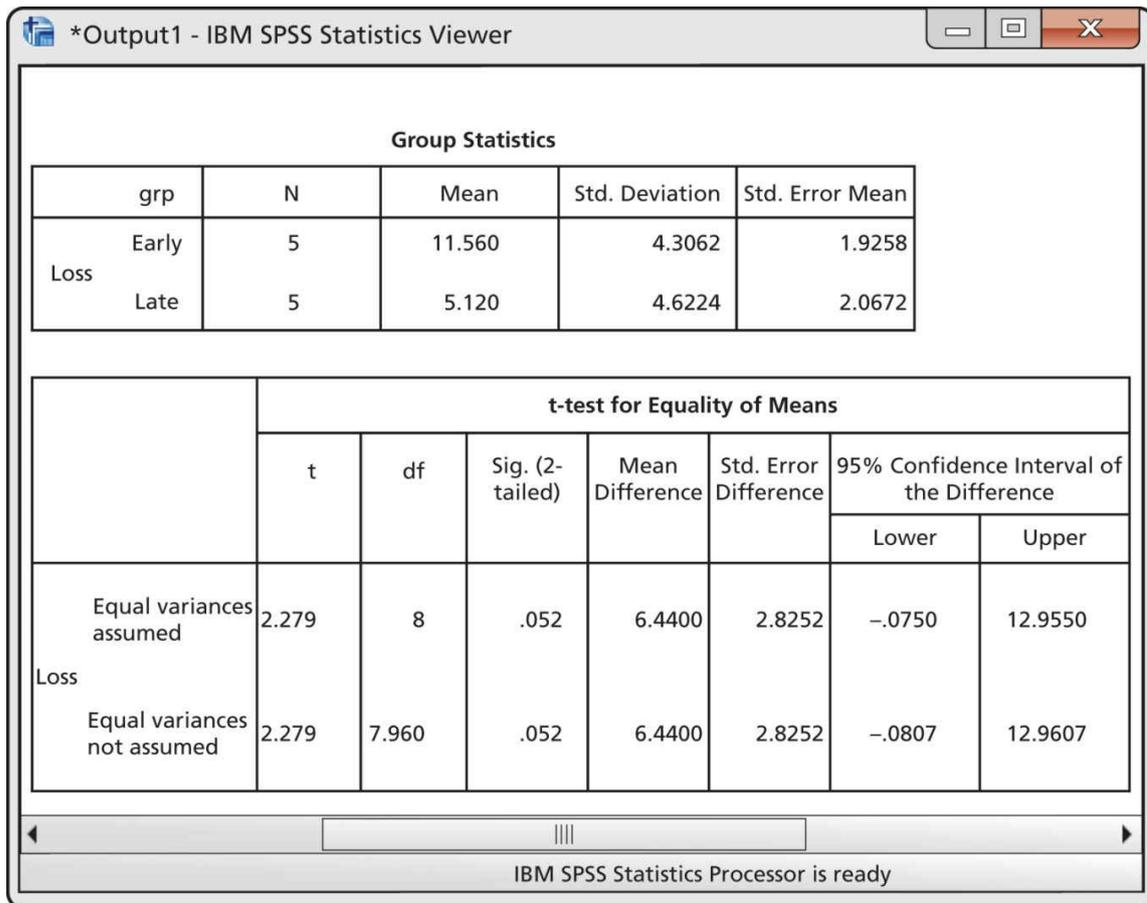
**Means and Std Deviations**

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
Early	5	11.5600	4.30616	1.9258	6.213	16.907
Late	5	5.1200	4.62245	2.0672	-0.620	10.860

**t Test**

Late-early  
Assuming unequal variances

Difference	-6.440	t Ratio	-2.27945
Std Err Dif	2.825	DF	7.960145
Upper CL Dif	0.081	Prob >  t	0.0523
Lower CL Dif	-12.961	Prob > t	0.9739
Confidence	0.95	Prob < t	0.0261*



**FIGURE 7.15**

SAS, Excel, JMP, and SPSS output for Example 7.17.

Figure 7.15 gives outputs for this analysis from several software packages. Although the formats differ, the basic information is the same. All report the sample sizes, the sample means and standard deviations (or variances), the  $t$  statistic, and its  $P$ -value. All agree that the  $P$ -value is small, though some give more detail than others. Software often labels the groups in alphabetical order. Always check the means first and report the statistic (you may need to change the sign) in an appropriate way. We do not need to do that here. Be sure to also mention the size of the effect you observed, such as “The mean weight loss for the early eaters was 6.44 kg higher than for the late eaters.”

There are two other things to notice in the outputs. First, SAS and SPSS only give results for the two-sided alternative. To get the  $P$ -value for the one-sided alternative, we must first check the mean difference to make sure it is in the proper direction. If it is, we divide the given  $P$ -value by 2. Also, SAS and SPSS report the results of *two*  $t$  procedures: a special procedure that assumes that the two population variances are equal and the general two-sample procedure that we have just studied. We don’t recommend the “equal-variances” procedures, but we describe them later, in the section on pooled two-sample  $t$  procedures.

## Software approximation for the degrees of freedom

We noted earlier that the two-sample  $t$  statistic does not have a  $t$  distribution. Moreover, the distribution changes as the unknown population standard deviations  $\sigma_1$  and  $\sigma_2$  change. However, the distribution can be approximated by a  $t$  distribution with degrees of freedom given by

$$df = \frac{(s_1^2 n_1 + s_2^2 n_2) \left( \frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} \right)}{(s_1^2 n_1 + s_2^2 n_2) \left( \frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} \right) + 2}$$

This is the approximation used by most statistical software. It is quite accurate when both sample sizes  $n_1$  and  $n_2$  are 5 or larger.

## Example

### 7.18 Degrees of freedom for directed reading assessment

For the DRP study of Example 7.14, the following table summarizes the data:

Group	$n$	$\bar{x}$	$s$
1	21	51.48	11.01
2	23	41.52	17.15

For greatest accuracy, we will use critical points from the  $t$  distribution with degrees of freedom given by the equation above:

$$\begin{aligned} df &= \frac{(11.01^2 21 + 17.15^2 23) \left( \frac{1}{21 - 1} + \frac{1}{23 - 1} \right)}{(11.01^2 21 + 17.15^2 23) \left( \frac{1}{21 - 1} + \frac{1}{23 - 1} \right) + 2} \\ &= 34.4869099 = 37.86 \end{aligned}$$

This is the value that we reported in Examples 7.14 and 7.15, where we gave the results produced by software.

The number  $df$  given by the preceding approximation is always at least as large as the smaller of  $n_1 - 1$  and  $n_2 - 1$ . On the other hand, the number  $df$  is never larger than the sum  $n_1 + n_2 - 2$  of the two individual degrees of freedom. The number  $df$  is generally not a whole number. There is a  $t$  distribution with any positive degrees of freedom, even though Table D contains entries only for whole-number degrees of freedom. When the number  $df$  is small and is not a whole number, interpolation between entries in Table D may be needed to obtain an accurate critical value or  $P$ -value. Because of this and the need to calculate  $df$ , we do not recommend regular use of this approximation if a computer is not doing the arithmetic. With a computer, however, the more accurate procedures are painless.

## USE YOUR KNOWLEDGE

### 7.60 Calculating the degrees of freedom

Assume that  $s_1 = 13$ ,  $s_2 = 8$ ,  $n_1 = 28$ , and  $n_2 = 24$ . Find the approximate degrees of freedom.

## The pooled two-sample $t$ procedures

There is one situation in which a  $t$  statistic for comparing two means has exactly a  $t$  distribution. This is when the two Normal population distributions have the *same* standard deviation. As we've done with other  $t$  statistics, we will first develop the  $z$  statistic and then, from it, the  $t$  statistic. In this case, notice that we need to substitute only a single standard error when we go from the  $z$  to the  $t$  statistic. This is why the resulting  $t$  statistic has a  $t$  distribution.

Call the common—and still unknown—standard deviation of both populations  $\sigma$ . Both sample variances  $s_1^2$  and  $s_2^2$  estimate  $\sigma^2$ . The best way to combine these two estimates is to average them with weights equal to their degrees of freedom. This gives more weight to the sample variance from the larger sample, which is reasonable. The resulting estimator of  $\sigma^2$  is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

This is called the **pooled estimator of  $\sigma^2$**  because it combines the information in both samples.

### pooled estimator of $\sigma^2$

When both populations have variance  $\sigma^2$ , the addition rule for variances says that  $\bar{x}_1 - \bar{x}_2$  has variance equal to the *sum* of the individual variances, which is

$$\sigma^2/n_1 + \sigma^2/n_2 = \sigma^2(1/n_1 + 1/n_2)$$

The standardized difference between means in this equal-variance case is therefore

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{1/n_1 + 1/n_2}}$$

This is a special two-sample  $z$  statistic for the case in which the populations have the same  $\sigma$ . Replacing the unknown  $\sigma$  by the estimate  $s_p$  gives a  $t$  statistic. The degrees of freedom are  $n_1 + n_2 - 2$ , the sum of the degrees of freedom of the two sample variances. This  $t$  statistic is the basis of the pooled two-sample  $t$  inference procedures.

## THE POOLED TWO-SAMPLE $t$ PROCEDURES

Suppose that an SRS of size  $n_1$  is drawn from a Normal population with unknown mean  $\mu_1$  and that an independent SRS of size  $n_2$  is drawn from another Normal population with unknown mean  $\mu_2$ . Suppose also that the two populations have the same standard deviation. A level  $C$  **confidence interval for  $\mu_1 - \mu_2$**  is

$$(\bar{x}_1 - \bar{x}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Here  $t^*$  is the value for the  $t(n_1 + n_2 - 2)$  density curve with area  $C$  between  $-t^*$  and  $t^*$ .

To test the hypothesis  $H_0: \mu_1 = \mu_2$ , compute the **pooled two-sample  $t$  statistic**

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

In terms of a random variable  $t$  having the  $t(n_1 + n_2 - 2)$  distribution, the  $P$ -value for a test of  $H_0$  against

$$H_a: \mu_1 > \mu_2 \text{ is } P(T \geq t)$$

$$H_a: \mu_1 < \mu_2 \text{ is } P(T \leq t)$$

$$H_a: \mu_1 \neq \mu_2 \text{ is } 2P(T \geq |t|)$$

## EXAMPLE

### 7.19 Calcium and blood pressure

Does increasing the amount of calcium in our diet reduce blood pressure? Examination of a large sample of people revealed a relationship between calcium intake and blood pressure, but such observational studies do not establish causation. Animal experiments, however, showed that calcium supplements do reduce blood pressure in rats, justifying an experiment with human subjects. A randomized comparative experiment gave one group of 10 black men a calcium supplement for 12 weeks. The control group of 11 black men received a placebo that appeared identical. (In fact, a block design with

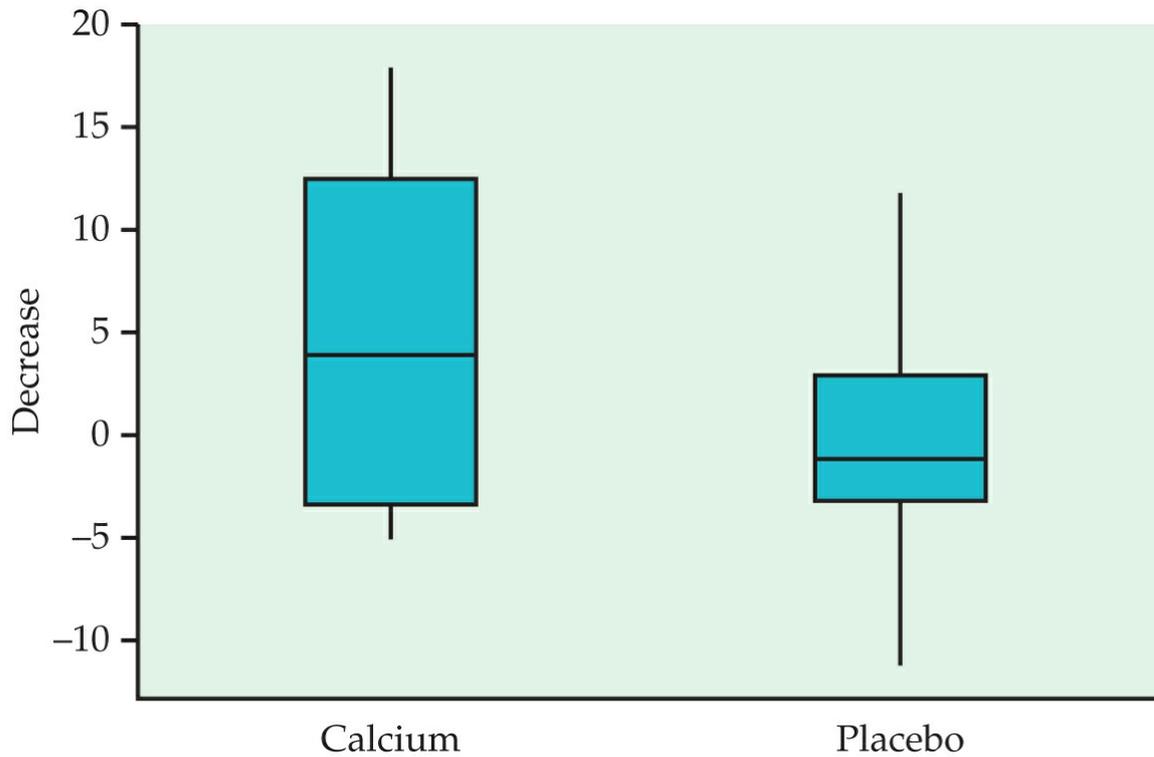
black and white men as the blocks was used. We will look only at the results for blacks, because the earlier survey suggested that calcium is more effective for blacks.) The experiment was double-blind. Table 7.5 gives the seated systolic (heart contracted) blood pressure for all subjects at the beginning and end of the 12-week period, in millimeters of mercury (mm Hg). Because the researchers were interested in decreasing blood pressure, Table 7.5 also shows the decrease for each subject. An increase appears as a negative entry.<sup>24</sup>



BP\_CA

**TABLE 7.5** Seated Systolic Blood Pressure (mm Hg)

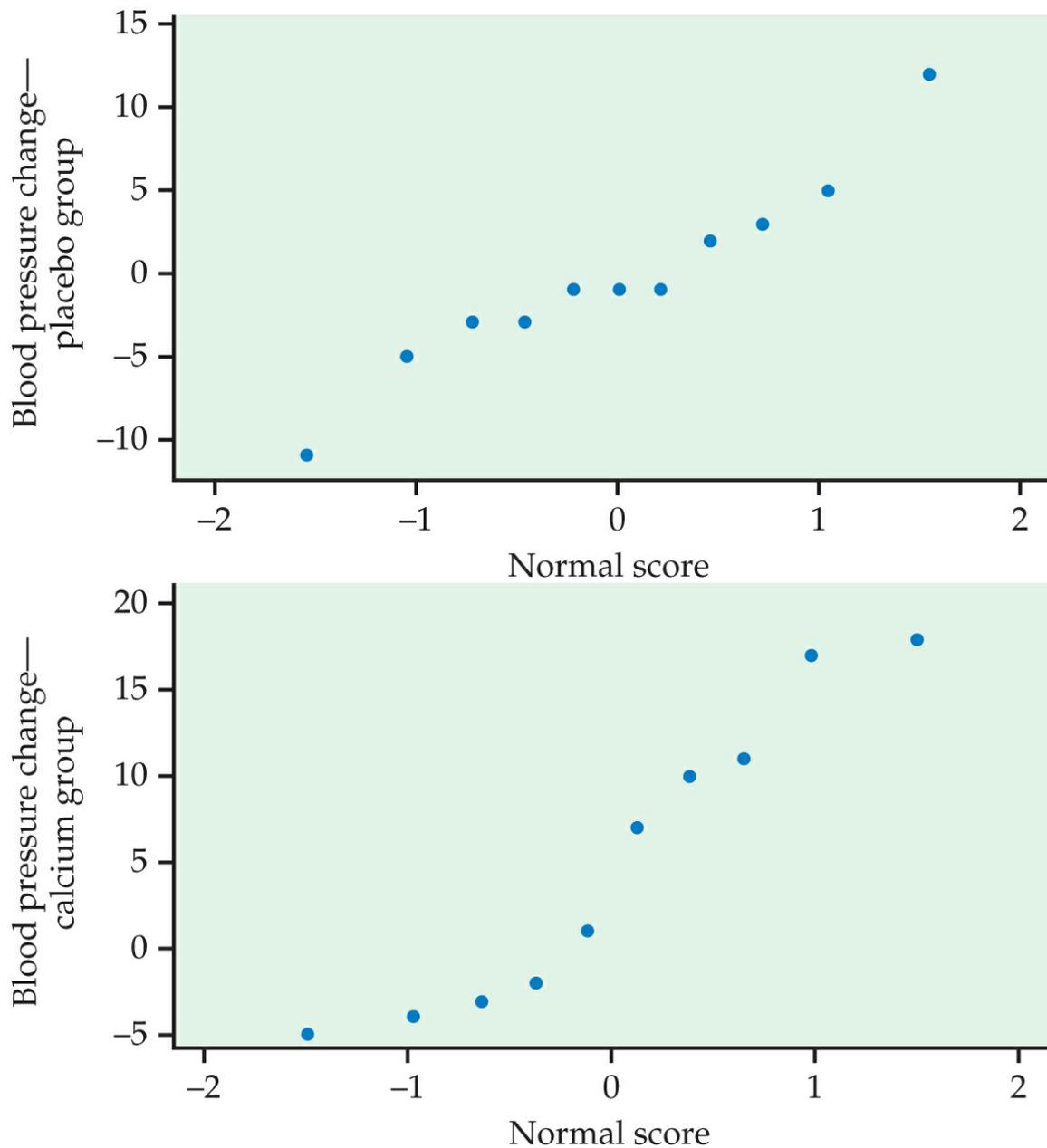
Calcium Group			Placebo Group		
Begin	End	Decrease	Begin	End	Decrease
107	100	7	123	124	-1
110	114	-4	109	97	12
123	105	18	112	113	-1
129	112	17	102	105	-3
112	115	-3	98	95	3
111	116	-5	114	119	-5
107	106	1	119	114	5
112	102	10	114	112	2
136	125	11	110	121	-11
102	104	-2	117	118	-1
			130	133	-3



**FIGURE 7.16**

Side-by-side boxplots of the decrease in blood pressure from Table 7.5.

As usual, we first examine the data. To compare the effects of the two treatments, take the response variable to be the amount of the decrease in blood pressure. Inspection of the data reveals that there are no outliers. Side-by-side boxplots and Normal quantile plots (Figures 7.16 and 7.17) give a more detailed picture. The calcium group has a somewhat short left tail, but there are no severe departures from Normality that will prevent use of  $t$  procedures. To examine the question of the researchers who collected these data, we perform a significance test.



**FIGURE 7.17**  
Normal quantile plots of the change in blood pressure from Table 7.5.

## EXAMPLE

### 7.20 Does increased calcium reduce blood pressure?

Take Group 1 to be the calcium group and Group 2 to be the placebo group. The evidence that calcium lowers blood pressure more than a placebo is

assessed by testing

$$H_0: \mu_1 = \mu_2$$

$$H_a > \mu_2$$

Here are the summary statistics for the decrease in blood pressure:

Group	Treatment	$n$	$\bar{x}$	$s$
1	Calcium	10	5.000	8.743
2	Placebo	11	-0.273	5.901

The calcium group shows a drop in blood pressure, and the placebo group has a small increase. The sample standard deviations do not rule out equal population standard deviations. A difference this large will often arise by chance in samples this small. We are willing to assume equal population standard deviations. The pooled sample variance is

$$\begin{aligned} sp^2 &= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 / (n_1 + n_2 - 2) \\ &= (10 - 1)8.743^2 + (11 - 1)5.901^2 / (10 + 11 - 2) = 54.536 \end{aligned}$$

so that

$$sp = \sqrt{54.536} = 7.385$$

The pooled two-sample  $t$  statistic is

$$\begin{aligned} t &= (\bar{x}_1 - \bar{x}_2) / sp \sqrt{1/n_1 + 1/n_2} \\ &= (5.000 - (-0.273)) / (7.385 \sqrt{1/10 + 1/11}) \\ &= 5.2733 / 3.227 = 1.634 \end{aligned}$$

The  $P$ -value is  $P(T \geq 1.634)$  where  $t$  has the  $t(19)$  distribution.

$df = 19$

$p$	0.10	0.05
$t^*$	1.328	1.729

From Table D we can see that  $P$  falls between the  $\alpha = 0.10$  and  $\alpha = 0.05$  levels. Statistical software gives the exact value  $P = 0.059$ . The experiment found evidence that calcium reduces blood pressure, but the evidence falls a bit short of the traditional 5% and 1% levels.

Sample size strongly influences the  $P$ -value of a test. An effect that fails to be significant at a specified level  $\alpha$  in a small sample can be significant in a larger sample. In the light of the rather small samples in Example 7.20, the evidence for some effect of calcium on blood pressure is rather good. The published account of the study combined these results for blacks with the results for whites and adjusted

for pretest differences among the subjects. Using this more detailed analysis, the researchers were able to report a  $P$ -value of 0.008.

Of course, a  $P$ -value is almost never the last part of a statistical analysis. To make a judgment regarding the size of the effect of calcium on blood pressure, we need a confidence interval.

## EXAMPLE

### 7.21 How different are the calcium and placebo groups?

We estimate that the effect of calcium supplementation is the difference between the sample means of the calcium and the placebo groups,  $\bar{x}_1 - \bar{x}_2 = 5.273$  mm Hg. A 90% confidence interval for  $\mu_1 - \mu_2$  uses the critical value  $t^* = 1.729$  from the  $t(19)$  distribution. The interval is

$$\begin{aligned} (\bar{x}_1 - \bar{x}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} &= [5.000 - (-0.273)] \pm (1.729)(7.385) \sqrt{\frac{1}{10} + \frac{1}{11}} \\ &= 5.273 \pm 5.579 \end{aligned}$$

We are 90% confident that the difference in means is in the interval  $(-0.306, 10.852)$ . The calcium treatment reduced blood pressure by about 5.3 mm Hg more than a placebo on the average, but the margin of error for this estimate is 5.6 mm Hg.



The pooled two-sample  $t$  procedures are anchored in statistical theory and so have long been the standard version of the two-sample  $t$  in textbooks. *But they require the assumption that the two unknown population standard deviations are equal.* As we shall see in Section 7.3, this assumption is hard to verify. The pooled  $t$  procedures are therefore a bit risky. They are reasonably robust against both non-Normality and unequal standard deviations when the sample sizes are nearly the same. When the samples are quite different in size, the pooled  $t$  procedures become sensitive to unequal standard deviations and should be used with caution unless the samples are large. Unequal standard deviations are quite common. In particular, it is not unusual for the spread of data to increase when its center gets larger. Statistical software often calculates both the pooled and the unpooled  $t$  statistics, as in Figure 7.15.

## USE YOUR KNOWLEDGE

### 7.61 Timing of food intake revisited

Figure 7.15 (pages 458–460) gives the outputs from four software packages for comparing the weight loss of two groups with different eating schedules. Some of the software reports both pooled and unpooled analyses. Which outputs give the pooled results? What are the pooled  $t$  and its  $P$ -value?

### 7.62 Equal sample sizes

The software outputs in Figure 7.15 give the *same value* for the pooled and unpooled  $t$  statistics. Do some simple algebra to show that this is always true when the two sample sizes  $n_1$  and  $n_2$  are the same. In other cases, the two  $t$  statistics usually differ.

## SECTION 7.2 Summary

Significance tests and confidence intervals for the difference between the means  $\mu_1$  and  $\mu_2$  of two Normal populations are based on the difference  $\bar{x}_1 - \bar{x}_2$  between the sample means from two independent SRSs. Because of the central limit theorem, the resulting procedures are approximately correct for other population distributions when the sample sizes are large.

When independent SRSs of sizes  $n_1$  and  $n_2$  are drawn from two Normal populations with parameters  $\mu_1, \sigma_1$  and  $\mu_2, \sigma_2$  the **two-sample  $z$  statistic**

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has the  $N(0, 1)$  distribution.

The **two-sample  $t$  statistic**

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

does *not* have a  $t$  distribution. However, good approximations are available.

**Conservative inference procedures** for comparing  $\mu_1$  and  $\mu_2$  are obtained from the two-sample  $t$  statistic by using the  $t(k)$  distribution with degrees of freedom  $k$  equal to the smaller of  $n_1 - 1$  and  $n_2 - 1$ .

**More accurate probability values** can be obtained by estimating the degrees of freedom from the data. This is the usual procedure for statistical software.

An approximate level  $C$  **confidence interval** for  $\mu_1 - \mu_2$  is given by

$$(\bar{x}_1 + \bar{x}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Here,  $t^*$  is the value for the  $t(k)$  density curve with area  $C$  between  $-t^*$  and  $t^*$  where  $k$  is computed from the data by software or is the smaller of  $n_1 - 1$  and  $n_2 - 1$ . The quantity

$$t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

is the **margin of error**.

Significance tests for  $H_0: \mu_1 = \mu_2$  use the **two-sample  $t$  statistic**

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

The  $P$ -value is approximated using the  $t(k)$  distribution where  $k$  is estimated from the data using software or is the smaller of  $n_1 - 1$  and  $n_2 - 1$ .

The guidelines for practical use of two-sample  $t$  procedures are similar to those for one-sample  $t$  procedures. Equal sample sizes are recommended.

If we can assume that the two populations have equal variances, **pooled two-sample  $t$  procedures** can be used. These are based on the **pooled estimator**

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

of the unknown common variance and the  $t(n_1 + n_2 - 2)$  distribution. We do not recommend this procedure for regular use.

## SECTION 7.2 Exercises

*For Exercises 7.56 and 7.57, see pages 453–454; for Exercises 7.58 and 7.59, see page 455; for Exercise 7.60, see page 461; and for Exercises 7.61 and 7.62, see page 466.*

*In exercises that call for two-sample  $t$  procedures, you may use either of the two approximations for the degrees of freedom that we have discussed: the value given by your software or the smaller of  $n_1 - 1$  and  $n_2 - 1$ . Be sure to state clearly which approximation you have used.*

### 7.63 What is wrong?

In each of the following situations explain what is wrong and why.

- A researcher wants to test  $H_0: \bar{x}_1 = \bar{x}_2$  versus the two-sided alternative  $H_a: \bar{x}_1 \neq \bar{x}_2$
- A study recorded the IQ scores of 100 college freshmen. The scores of the 56 males in the study were compared with the scores of all 100 freshmen using the two-sample methods of this section.
- A two-sample  $t$  statistic gave a  $P$ -value of 0.94. From this we can reject the null hypothesis with 90% confidence.
- A researcher is interested in testing the one-sided alternative  $H_a: \mu < \mu_2$ . The significance test gave  $t = 2.15$ . Since the  $P$ -value for the two-sided alternative is 0.036, he concluded that his  $P$ -value

was 0.018.

## 7.64 Basic concepts

For each of the following, answer the question and give a short explanation of your reasoning.

(a) A 95% confidence interval for the difference between two means is reported as (0.8, 2.3). What can you conclude about the results of a significance test of the null hypothesis that the population means are equal versus the two-sided alternative?

(b) Will larger samples generally give a larger or smaller margin of error for the difference between two sample means?

## 7.65 More basic concepts

For each of the following, answer the question and give a short explanation of your reasoning.

(a) A significance test for comparing two means gave  $t = -1.97$  with 10 degrees of freedom. Can you reject the null hypothesis that the  $\mu$ 's are equal versus the two-sided alternative at the 5% significance level?

(b) Answer part (a) for the one-sided alternative that the difference between means is negative.

## 7.66 Effect of the confidence level

Assume that  $\bar{x}_1=100$ ,  $\bar{x}_2=115$ ,  $s_1 = 19$ ,  $s_2 = 16$ ,  $n_1 = 50$ , and  $n_2 = 40$ . Find a 95% confidence interval for the difference between the corresponding values of  $\mu$ . Does this interval include more or fewer values than a 99% confidence interval would? Explain your answer.

## 7.67 Trustworthiness and eye color

Why do we naturally tend to trust some strangers more than others? One group of researchers decided to study the relationship between eye color and trustworthiness.<sup>25</sup> In their experiment the researchers took photographs of 80 students (20 males with brown eyes, 20 males with blue eyes, 20 females with brown eyes, and 20 females with blue eyes), each seated in front of a white background looking directly at the camera with a neutral expression. These photos were cropped so the eyes were horizontal and at the same height in the photo and so the neckline was visible. They then recruited 105 participants to judge the trustworthiness of each student photo. This was done using a 10-point scale, where 1 meant very untrustworthy and 10 very trustworthy. The 80 scores from each participant were then converted to  $z$ -scores, and the average  $z$ -score of each photo (across all 105 participants) was used for the analysis. Here is a summary of the results:

Eye color	$n$	$\bar{x}$	$s$
Brown	40	0.55	1.68
Blue	40	-0.38	1.53

Can we conclude from these data that brown-eyed students appear more trustworthy compared to their blue-eyed counterparts? Test the hypothesis that the average scores for the two groups are the same.

## 7.68 Facebook use in college

Because of Facebook's rapid rise in popularity among college students, there is a great deal of interest in the relationship between Facebook use and academic performance. One study collected information on  $n = 1839$  undergraduate students to look at the relationships among frequency of Facebook use, participation in Facebook activities, time spent preparing for class, and overall GPA.<sup>26</sup>

Students reported preparing for class an average of 706 minutes per week with a standard deviation of 526 minutes. Students also reported spending an average of 106 minutes per day on Facebook with a standard deviation of 93 minutes; 8% of the students reported spending no time on Facebook.

- (a) Construct a 95% confidence interval for the average number of minutes per week a student prepares for class.
- (b) Construct a 95% confidence interval for the average number of minutes per week a student spends on Facebook. (*Hint*: Be sure to convert from minutes per day to minutes per week.)
- (c) Explain why you might expect the population distributions of these two variables to be highly skewed to the right. Do you think this fact makes your confidence intervals invalid? Explain your answer.

### 7.69 Possible biases?

Refer to the previous exercise. The authors state:

*All students surveyed were U.S. residents admitted through the regular admissions process at a 4-year, public, primarily residential institution in the northeastern United States ( $N = 3866$ ). Students were sent a link to a survey hosted on SurveyMonkey.com, a survey-hosting website, through their university-sponsored email accounts. For the students who did not participate immediately, two additional reminders were sent, 1 week apart. Participants were offered a chance to enter a drawing to win one of 90 \$10 Amazon.com gift cards as incentive. A total of 1839 surveys were completed for an overall response rate of 48%.*

Discuss how these factors influence your interpretation of the results of this survey.

### 7.70 Comparing means

Refer to Exercise 7.68. Suppose that you wanted to compare the average minutes per week spent on Facebook with the average minutes per week spent preparing for class.

- (a) Provide an estimate of this difference.
- (b) Explain why it is incorrect to use the two-sample  $t$  test to see if the means differ.

### 7.71 Sadness and spending

The "misery is not miserly" phenomenon refers to a person's spending judgment going haywire when the person is sad. In a study, 31 young adults were given \$10 and randomly assigned to either a sad or a neutral group. The participants in the sad group watched a video about the death of a boy's mentor (from *The Champ*), and those in the neutral group watched a video on the Great Barrier Reef.

After the video, each participant was offered the chance to trade \$0.50 increments of the \$10 for an insulated water bottle.<sup>27</sup> Here are the data:  **SADNESS**

Group	Purchase price (\$)										
Neutral	0.00	2.00	0.00	1.00	0.50	0.00	0.50				
	2.00	1.00	0.00	0.00	0.00	0.00	1.00				
Sad	3.00	4.00	0.50	1.00	2.50	2.00	1.50	0.00	1.00		
	1.50	1.50	2.50	4.00	3.00	3.50	1.00	3.50			

- Examine each group's prices graphically. Is use of the  $t$  procedures appropriate for these data? Carefully explain your answer.
- Make a table with the sample size, mean, and standard deviation for each of the two groups.
- State appropriate null and alternative hypotheses for comparing these two groups.
- Perform the significance test at the  $\alpha = 0.05$  level, making sure to report the test statistic, degrees of freedom, and  $P$ -value. What is your conclusion?
- Construct a 95% confidence interval for the mean difference in purchase price between the two groups.

## 7.72 Wine labels with animals?

Traditional brand research argues that successful logos are ones that are highly relevant to the product they represent. However, a market research firm recently reported that nearly 20% of all table wine brands introduced in the last three years feature an animal on the label. Since animals have little to do with the product, why are marketers using this tactic?

Some researchers have proposed that consumers who are "primed" (in other words, they've thought about the image earlier in an unrelated context) process visual information more easily.<sup>28</sup> To demonstrate this, the researchers randomly assigned participants to either a primed or a nonprimed group. Each participant was asked to indicate his or her attitude toward a product on a seven-point scale (from 1 = dislike very much to 7 = like very much). A bottle of MagicCoat pet shampoo, with a picture of a collie on the label, was the product. Prior to giving this score, however, participants were asked to do a word find where four of the words were common to both groups (pet, grooming, bottle, label) and four were either related to the product image (dog, collie, puppy, woof) or conflicted with the image (cat, feline, kitten, meow). The following table contains the responses listed from smallest to largest.  **BPREF**

Group	Brand attitude																						
Primed	2	2	3	3	3	4	4	4	4	4	4	4	4	4	4	4	5	5	5	5	5	5	
Nonprimed	1	1	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	4	4	4	5

- Examine the scores of each group graphically. Is it appropriate to use the two-sample  $t$  procedures? Explain your answer.
- Test whether these two groups show the same preference for this product. Use a two-sided alternative hypothesis and a significance level of 5%.
- Construct a 95% confidence interval for the difference in average preference.
- Write a short summary of your conclusions.

## 7.73 Drive-thru customer service

*QSRMagazine.com* assessed 2053 drive-thru visits at quick-service restaurants.<sup>29</sup> One benchmark assessed was customer service. Responses ranged from “Rude (1)” to “Very Friendly (5).” The following table breaks down the responses according to two of the chains studied. 

Chain	Rating				
	1	2	3	4	5
Taco Bell	5	3	54	109	136
McDonald’s	2	22	73	165	100

- Comment on the appropriateness of  $t$  procedures for these data.
- Report the means and standard deviations of the ratings for each chain separately.
- Test whether the two chains, on average, have the same customer satisfaction. Use a two-sided alternative hypothesis and a significance level of 5%.
- Construct a 95% confidence interval for the difference in average satisfaction.

### 7.74 Diet and mood

Researchers were interested in comparing the long-term psychological effects of being on a high-carbohydrate, low-fat (LF) diet versus a high-fat, low-carbohydrate (LC) diet.<sup>30</sup> A total of 106 overweight and obese participants were randomly assigned to one of these two energy-restricted diets. At 52 weeks, 32 LC dieters and 33 LF dieters remained. Mood was assessed using a total mood disturbance score (TMDS), where a lower score is associated with a less negative mood. A summary of these results follows:

Group	$n$	$\bar{x}$	$s$
LC	32	47.3	28.3
LF	33	19.3	25.8

- Is there a difference in the TMDS at Week 52? Test the null hypothesis that the dieters’ average mood in the two groups is the same. Use a significance level of 0.05.
- Critics of this study focus on the specific LC diet (that is, the science) and the dropout rate. Explain why the dropout rate is important to consider when drawing conclusions from this study.

### 7.75 Comparison of dietary composition

Refer to Example 7.16 (page 456). That study also broke down the dietary composition of the main meal. The following table summarizes the total fats, protein, and carbohydrates in the main meal (g) for the two groups:

	Early eaters ( $n = 202$ )		Late eaters ( $n = 200$ )	
	$\bar{x}$	$s$	$\bar{x}$	$s$
Fats	23.1	12.5	21.4	8.2
Protein	27.6	8.6	25.7	6.8
Carbohydrates	64.1	21.0	63.5	20.8

- Is it appropriate to use the two-sample  $t$  procedures that we studied in this section to analyze these data for group differences? Give reasons for your answer.

- (b) Describe appropriate null and alternative hypotheses for comparing the two groups in terms of fats consumed.
- (c) Carry out the significance test using  $\alpha = 0.05$ . Report the test statistic with the degrees of freedom and the  $P$ -value. Write a short summary of your conclusion.
- (d) Find a 95% confidence interval for the difference between the two means. Compare the information given by the interval with the information given by the significance test.

### 7.76 More on dietary composition

Refer to the previous exercise. Repeat parts (b) through (d) for protein and carbohydrates. Write a short summary of your findings.

### 7.77 Dust exposure at work

Exposure to dust at work can lead to lung disease later in life. One study measured the workplace exposure of tunnel construction workers.<sup>31</sup> Part of the study compared 115 drill and blast workers with 220 outdoor concrete workers. Total dust exposure was measured in milligram years per cubic meter ( $\text{mg} \cdot \text{y}/\text{m}^3$ ). The mean exposure for the drill and blast workers was  $18.0 \text{ mg} \cdot \text{y}/\text{m}^3$  with a standard deviation of  $7.8 \text{ mg} \cdot \text{y}/\text{m}^3$ . For the outdoor concrete workers, the corresponding values were  $6.5 \text{ mg} \cdot \text{y}/\text{m}^3$  and  $3.4 \text{ mg} \cdot \text{y}/\text{m}^3$ .

- (a) The sample included all workers for a tunnel construction company who received medical examinations as part of routine health checkups. Discuss the extent to which you think these results apply to other similar types of workers.
- (b) Use a 95% confidence interval to describe the difference in the exposures. Write a sentence that gives the interval and provides the meaning of 95% confidence.
- (c) Test the null hypothesis that the exposures for these two types of workers are the same. Justify your choice of a one-sided or two-sided alternative. Report the test statistic, the degrees of freedom, and the  $P$ -value. Give a short summary of your conclusion.
- (d) The authors of the article describing these results note that the distributions are somewhat skewed. Do you think that this fact makes your analysis invalid? Give reasons for your answer.

### 7.78 Not all dust is the same

Not all dust particles that are in the air around us cause problems for our lungs. Some particles are too large and stick to other areas of our body before they can get to our lungs. Others are so small that we can breathe them in and out and they will not deposit in our lungs. The researchers in the study described in the previous exercise also measured respirable dust. This is dust that deposits in our lungs when we breathe it. For the drill and blast workers, the mean exposure to respirable dust was  $6.3 \text{ mg} \cdot \text{y}/\text{m}^3$  with a standard deviation of  $2.8 \text{ mg} \cdot \text{y}/\text{m}^3$ . The corresponding values for the outdoor concrete workers were  $1.4 \text{ mg} \cdot \text{y}/\text{m}^3$  and  $0.7 \text{ mg} \cdot \text{y}/\text{m}^3$ . Analyze these data using the questions in the previous exercise as a guide.

### 7.79 Change in portion size

A study of food portion sizes reported that over a 17-year period, the average size of a soft drink consumed by Americans aged 2 years and older increased from 13.1 ounces (oz) to 19.9 oz. The

authors state that the difference is statistically significant with  $P < 0.01$ .<sup>32</sup> Explain what additional information you would need to compute a confidence interval for the increase, and outline the procedure that you would use for the computations. Do you think that a confidence interval would provide useful additional information? Explain why or why not.

## 7.80 Beverage consumption

The results in the previous exercise were based on two national surveys with a very large number of individuals. Here is a study that also looked at beverage consumption, but the sample sizes were much smaller. One part of this study compared 20 children who were 7 to 10 years old with 5 children who were 11 to 13.<sup>33</sup> The younger children consumed an average of 8.2 oz of sweetened drinks per day while the older ones averaged 14.5 oz. The standard deviations were 10.7 oz and 8.2 oz, respectively.

- (a) Do you think that it is reasonable to assume that these data are Normally distributed? Explain why or why not. (*Hint*: Think about the 68–95–99.7 rule.)
- (b) Using the methods in this section, test the null hypothesis that the two groups of children consume equal amounts of sweetened drinks versus the two-sided alternative. Report all details of the significance-testing procedure with your conclusion.
- (c) Give a 95% confidence interval for the difference in means.
- (d) Do you think that the analyses performed in parts (b) and (c) are appropriate for these data? Explain why or why not.
- (e) The children in this study were all participants in an intervention study at the Cornell Summer Day Camp at Cornell University. To what extent do you think that these results apply to other groups of children?

## 7.81 Study design is important!

Recall Exercise 7.58 (page 455). You are concerned that day of the week may affect the number of hits. So to compare the two MySpace page designs, you choose two successive weeks in the middle of a month. You flip a coin to assign one Monday to the first design and the other Monday to the second. You repeat this for each of the seven days of the week. You now have 7 hit amounts for each design. It is *incorrect* to use the two-sample  $t$  test to see if the mean hits differ for the two designs. Carefully explain why.

## 7.82 New computer monitors?

The purchasing department has suggested that all new computer monitors for your company should be flat screens. You want data to assure you that employees will like the new screens. The next 20 employees needing a new computer are the subjects for an experiment.

- (a) Label the employees 01 to 20. Randomly choose 10 to receive flat screens. The remaining 10 get standard monitors.
- (b) After a month of use, employees express their satisfaction with their new monitors by responding to the statement “I like my new monitor” on a scale from 1 to 5, where 1 represents “strongly disagree,” 2 is “disagree,” 3 is “neutral,” 4 is “agree,” and 5 stands for “strongly agree.” The employees with the flat screens have average satisfaction 4.8 with standard deviation 0.7. The employees with the standard monitors have average 3.0 with standard deviation 1.5. Give a 95% confidence interval for the difference in the mean satisfaction scores for all employees.

(c) Would you reject the null hypothesis that the mean satisfaction for the two types of monitors is the same versus the two-sided alternative at significance level 0.05? Use your confidence interval to answer this question. Explain why you do not need to calculate the test statistic.

### 7.83 Why randomize?

Refer to the previous exercise. A coworker suggested that you give the flat screens to the next 10 employees who need new screens and the standard monitor to the following 10. Explain why your randomized design is better.

### 7.84 Does ad placement matter?

Corporate advertising tries to enhance the image of the corporation. A study compared two ads from two sources, the *Wall Street Journal* and the *National Enquirer*. Subjects were asked to pretend that their company was considering a major investment in Performax, the fictitious sportswear firm in the ads. Each subject was asked to respond to the question “How trustworthy was the source in the sportswear company ad for Performax?” on a 7-point scale. Higher values indicated more trustworthiness.<sup>34</sup> Here is a summary of the results:

Ad source	$n$	$\bar{x}$	$s$
<i>Wall Street Journal</i>	66	4.77	1.50
<i>National Enquirer</i>	61	2.43	1.64

- (a) Compare the two sources of ads using a  $t$  test. Be sure to state your null and alternative hypotheses, the test statistic with degrees of freedom, the  $P$ -value, and your conclusion.  
 (b) Give a 95% confidence interval for the difference.  
 (c) Write a short paragraph summarizing the results of your analyses.

### 7.85 Size of trees in the northern and southern halves

The study of 584 longleaf pine trees in the Wade Tract in Thomas County, Georgia, had several purposes. Are trees in one part of the tract more or less like trees in any other part of the tract or are there differences? In Example 6.1 (page 352) we examined how the trees were distributed in the tract and found that the pattern was not random. In this exercise we will examine the sizes of the trees. In Exercise 7.31 (page 443) we analyzed the sizes, measured as diameter at breast height (DBH), for a random sample of 40 trees. Here we divide the tract into northern and southern halves and take random samples of 30 trees from each half. Here are the diameters in centimeters (cm) of the sampled trees:

	27.8	14.5	39.1	3.2	58.8	55.5	25.0	5.4	19.0	30.6
North	15.1	3.6	28.4	15.0	2.2	14.2	44.2	25.7	11.2	46.8
	36.9	54.1	10.2	2.5	13.8	43.5	13.8	39.7	6.4	4.8
	44.4	26.1	50.4	23.3	39.5	51.0	48.1	47.2	40.3	37.4
South	36.8	21.7	35.7	32.0	40.4	12.8	5.6	44.3	52.9	38.0
	2.6	44.6	45.5	29.1	18.7	7.0	43.8	28.3	36.9	51.6

- (a) Use a back-to-back stemplot and side-by-side boxplots to examine the data graphically. Describe the patterns in the data.  
 (b) Is it appropriate to use the methods of this section to compare the mean DBH of the trees in the north half of the tract with the mean DBH of the trees in the south half? Give reasons for your answer.

- (c) What are appropriate null and alternative hypotheses for comparing the two samples of tree DBHs? Give reasons for your choices.
- (d) Perform the significance test. Report the test statistic, the degrees of freedom, and the  $P$ -value. Summarize your conclusion.
- (e) Find a 95% confidence interval for the difference in mean DBHs. Explain how this interval provides additional information about this problem.

### 7.86 Size of trees in the eastern and western halves

Refer to the previous exercise. The Wade Tract can also be divided into eastern and western halves. Here are the DBHs of 30 randomly selected longleaf pine trees from each half: 

	23.5	43.5	6.6	11.5	17.2	38.7	2.3	31.5	10.5	23.7
East	13.8	5.2	31.5	22.1	6.7	2.6	6.3	51.1	5.4	9.0
	43.0	8.7	22.8	2.9	22.3	43.8	48.1	46.5	39.8	10.9
	17.2	44.6	44.1	35.5	51.0	21.6	44.1	11.2	36.0	42.1
West	3.2	25.5	36.5	39.0	25.9	20.8	3.2	57.7	43.3	58.0
	21.7	35.6	30.9	40.6	30.7	35.6	18.2	2.9	20.4	11.4

Using the questions in the previous exercise, analyze these data.

### 7.87 Sales of a small appliance across months

A market research firm supplies manufacturers with estimates of the retail sales of their products from samples of retail stores. Marketing managers are prone to look at the estimate and ignore sampling error. Suppose that an SRS of 70 stores this month shows mean sales of 53 units of a small appliance, with standard deviation 12 units. During the same month last year, an SRS of 55 stores gave mean sales of 50 units, with standard deviation 10 units. An increase from 50 to 53 is a rise of 6%. The marketing manager is happy because sales are up 6%.

- (a) Use the two-sample  $t$  procedure to give a 95% confidence interval for the difference in mean number of units sold at all retail stores.
- (b) Explain in language that the manager can understand why he cannot be certain that sales rose by 6%, and that in fact sales may even have dropped.

### 7.88 An improper significance test

A friend has performed a significance test of the null hypothesis that two means are equal. His report states that the null hypothesis is rejected in favor of the alternative that the first mean is larger than the second. In a presentation on his work, he notes that the first sample mean was larger than the second mean and this is why he chose this particular one-sided alternative.

- (a) Explain what is wrong with your friend's procedure and why.
- (b) Suppose that he reported  $t = 1.70$  with a  $P$ -value of 0.06. What is the correct  $P$ -value that he should report?

### 7.89 Breast-feeding versus baby formula

A study of iron deficiency among infants compared samples of infants following different feeding regimens. One group contained breast-fed infants, while the infants in another group were fed a standard baby formula without any iron supplements. Here are summary results on blood hemoglobin levels at 12 months of age:<sup>35</sup>

Group	$n$	$\bar{x}$	$s$
Breast-fed	23	13.3	1.7
Formula	19	12.4	1.8

(a) Is there significant evidence that the mean hemoglobin level is higher among breast-fed babies? State  $H_0$  and  $H_a$  and carry out a  $t$  test. Give the  $P$ -value. What is your conclusion?

(b) Give a 95% confidence interval for the mean difference in hemoglobin level between the two populations of infants.

(c) State the assumptions that your procedures in parts (a) and (b) require in order to be valid.

## 7.90 Revisiting the sadness and spending study

In Exercise 7.71 (page 468), the purchase price of a water bottle was analyzed using the two-sample  $t$  procedures that do not assume equal standard deviations. Compare the means using a significance test and find the 95% confidence interval for the difference using the pooled methods. How do the results compare with those you obtained in Exercise 7.71?  **BPREF**

## 7.91 Revisiting wine labels with animals

In Exercise 7.72 (page 469), attitudes toward a product were compared using the two-sample  $t$  procedures that do not assume equal standard deviations. Compare the means using a significance test and find the 95% confidence interval for the difference using the pooled methods. How do the results compare with those you obtained in Exercise 7.72?

## 7.92 Revisiting dietary composition

In Exercise 7.75 (page 469), the total amount of fats was analyzed using the two-sample  $t$  procedures that do not assume equal standard deviations. Examine the standard deviations for the two groups and verify that it is appropriate to use the pooled procedures for these data. Compare the means using a significance test and find the 95% confidence interval for the difference using the pooled methods. How do the results compare with those you obtained in Exercise 7.75?

## 7.93 Revisiting the size of trees

Refer to the Wade Tract DBH data in Exercise 7.85 (page 471), where we compared a sample of trees from the northern half of the tract with a sample from the southern half. Because the standard deviations for the two samples are quite close, it is reasonable to analyze these data using the pooled procedures. Perform the significance test and find the 95% confidence interval for the difference in means using these methods. Summarize your results and compare them with what you found in

Exercise 7.85.  **NSPINES**

## 7.94 Revisiting the food-timing study

Example 7.16 (page 456) gives summary statistics for weight loss in early eaters and late eaters. The two sample standard deviations are quite similar, so we may be willing to assume equal population standard deviations. Calculate the pooled  $t$  test statistic and its degrees of freedom from the summary statistics. Use Table D to assess significance. How do your results compare with the unpooled analysis in the example?

## 7.95 Computing the degrees of freedom

Use the Wade Tract data in Exercise 7.85 to calculate the software approximation to the degrees of freedom using the formula on page 460. Verify your calculation with software.

## 7.96 Again computing the degrees of freedom

Use the Wade Tract data in Exercise 7.86 to calculate the software approximation to the degrees of freedom using the formula on page 460. Verify your calculation with software.

## 7.97 Revisiting the dust exposure study

The data on occupational exposure to dust that we analyzed in Exercise 7.77 (page 470) come from two groups of workers that are quite different in size. This complicates the issue regarding pooling because the sample that is larger will dominate the calculations.

- Calculate the software approximation to the degrees of freedom using the formula on page 460. Then verify your calculations with software.
- Find the pooled estimate of the standard deviation. Write a short summary comparing it with the estimates of the standard deviations that come from each group.
- Find the standard error of the difference in sample means that you would use for the method that does not assume equal variances. Do the same for the pooled approach. Compare these two estimates with each other.
- Perform the significance test and find the 95% confidence interval using the pooled methods. How do these results compare with those you found in Exercise 7.77?
- Exercise 7.78 has data for the same workers but for respirable dust. Here the standard deviations differ more than those in Exercise 7.77 do. Answer parts (a) through (d) for these data. Write a summary of what you have found in this exercise.

## 7.98 Revisiting the small-sample example

Refer to Example 7.17 (page 457). This is a case where the sample sizes are quite small. With only 5 observations per group, we have very little information to make a judgment about whether the population standard deviations are equal. The potential gain from pooling is large when the sample sizes are small. Assume that we will perform a two-sided test using the 5% significance level. 

### EATER

- Find the critical value for the unpooled  $t$  test statistic that does not assume equal variances. Use the minimum of  $n_1 - 1$  and  $n_2 - 1$  for the degrees of freedom.
- Find the critical value for the pooled  $t$  test statistic.
- How does comparing these critical values show an advantage of the pooled test?

## 7.3 Other Topics in Comparing Distributions

When you complete this section, you will be able to

- Perform an  $F$  test for the equality of two variances.
- Argue why this  $F$  test is of very little value in practice. In other words, identify when this test can be used and, more importantly, when it cannot.
- Determine the sample size necessary to have adequate power to detect a scaled difference in means of size  $\delta$ .

In this section we discuss three topics that are related to the material that we have already covered in this chapter. If we can do inference for means, it is natural to ask if we can do something similar for spread. The answer is Yes, but there are many cautions. We also discuss robustness and show how to find the power for the two-sample  $t$  test. If you plan to design studies, you should become familiar with this last topic.

### Inference for population spread

The two most basic descriptive features of a distribution are its center and spread. In a Normal population, these aspects are measured by the mean and the standard deviation. We have described procedures for inference about population means for Normal populations and found that these procedures are often useful for non-Normal populations as well. It is natural to turn next to inference about the standard deviations of Normal populations. Our recommendation here is short and clear: don't do it without expert advice.



We will describe the  $F$  test for comparing the spread of two Normal populations. *Unlike the  $t$  procedures for means, the  $F$  test and other procedures for standard deviations are extremely sensitive to non-Normal distributions.* This lack of robustness does not improve in large samples. It is difficult in practice to tell whether a significant  $P$ -value is evidence of unequal population spreads or simply evidence that the populations are not Normal. Consequently, we do not recommend use of inference about population standard deviations in basic statistical practice.<sup>36</sup>

It was once common to test equality of standard deviations as a preliminary to performing the pooled two-sample  $t$  test for equality of two population means. It is

better practice to check the distributions graphically, with special attention to skewness and outliers, and to use the software-based two-sample  $t$  that does not require equal standard deviations. In the words of one distinguished statistician, “To make a preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!”<sup>37</sup>

### The $F$ test for equality of spread

Because of the limited usefulness of procedures for inference about the standard deviations of Normal distributions, we will present only one such procedure. Suppose that we have independent SRSs from two Normal populations, a sample of size  $n_1$  from  $N(\mu_1, \sigma_1)$  and a sample of size  $n_2$  from  $N(\mu_2, \sigma_2)$ . The population means and standard deviations are all unknown. The hypothesis of equal spread

$$H_0: \sigma_1 = \sigma_2$$

is tested against

$$H_a: \sigma_1 \neq \sigma_2$$

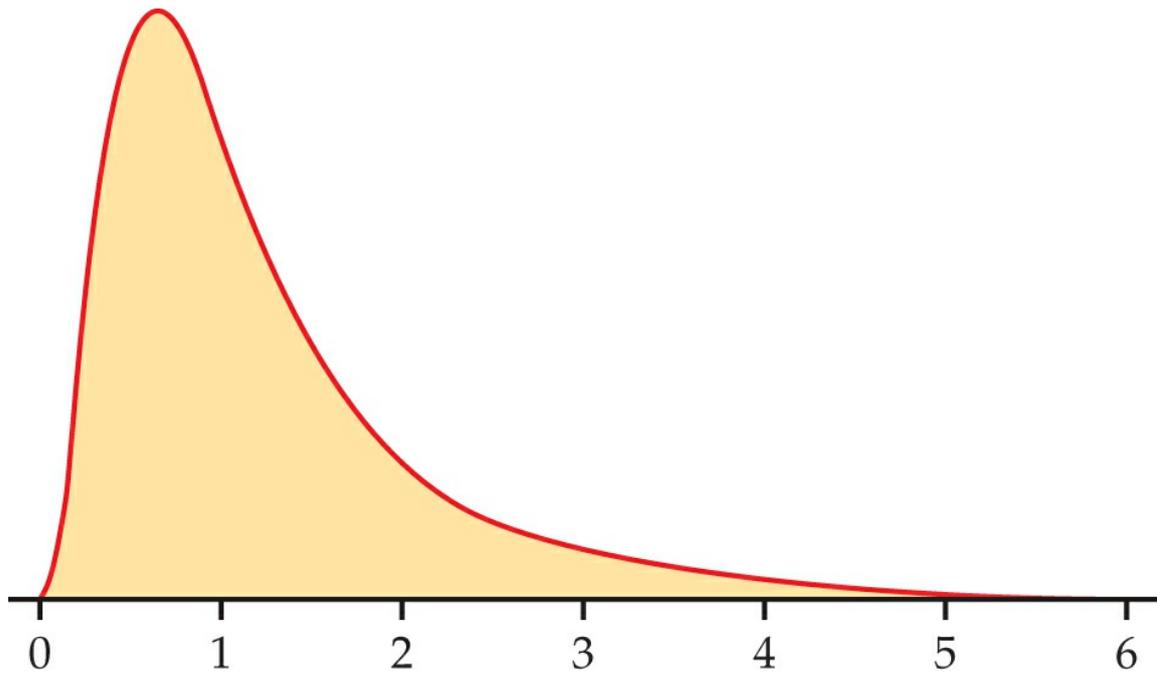
by a simple statistic, the ratio of the sample variances.

#### The $F$ Statistic and $F$ Distributions

When  $s_1^2$  and  $s_2^2$  are sample variances from independent SRSs of sizes  $n_1$  and  $n_2$  drawn from Normal populations, the  **$F$  statistic**

$$F = \frac{s_1^2}{s_2^2}$$

has the  **$F$  distribution** with  $n_1 - 1$  and  $n_2 - 1$  degrees of freedom when  $H_0: \sigma_1 = \sigma_2$  is true.



**FIGURE 7.18**

The density curve for the  $F(9, 10)$  distribution. The  $F$  distributions are skewed to the right.

The  **$F$  distributions** are a family of distributions with two parameters: the degrees of freedom of the sample variances in the numerator and denominator of the  $F$  statistic. The  $F$  distributions are another of R. A. Fisher's contributions to statistics and are called  $F$  in his honor. Fisher introduced  $F$  statistics for comparing several means. We will meet these useful statistics in later chapters.

### *F* distributions

Our brief notation will be  $F(j, k)$  for the  $F$  distribution with  $F$  degrees of freedom in the numerator and  $k$  degrees of freedom in the denominator. The numerator degrees of freedom are always mentioned first. Interchanging the degrees of freedom changes the distribution, so the order is important. The  $F$  distributions are not symmetric but are right-skewed. The density curve in Figure 7.18 illustrates the shape. Because sample variances cannot be negative, the  $F$  statistic takes only positive values and the  $F$  distribution has no probability below 0. The peak of the  $F$  density curve is near 1; values far from 1 in either direction provide evidence against the hypothesis of equal standard deviations.

Tables of  $F$  critical values are awkward because a separate table is needed for every pair of degrees of freedom  $j$  and  $k$ . Table E in the back of the book gives upper  $P$  critical values of the  $F$  distributions for  $P = 0.10, 0.05, 0.025, 0.01,$  and  $0.001$ . For example, these critical values for the  $F(9, 10)$  distribution shown in Figure 7.18 are

$p$	0.10	0.05	0.025	0.01	0.001
$F^*$	2.35	3.02	3.78	4.94	8.96

The skewness of  $F$  distributions causes additional complications. In the symmetric Normal and  $t$  distributions, the point with probability 0.05 below it is just the negative of the point with probability 0.05 above it. This is not true for  $F$  distributions. We therefore require either tables of both the upper and lower tails or a way to eliminate the need for lower-tail critical values. Statistical software that eliminates the need for tables is plainly very convenient. If you do not use statistical software, arrange the  $F$  test as follows:

1. Take the test statistic to be

$$F = \frac{\text{larger } s^2}{\text{smaller } s^2}$$

This amounts to naming the populations so that  $s_1^2$  is the larger of the observed sample variances. The resulting  $F$  is always 1 or greater.

2. Compare the value of  $F$  with the critical values from Table E. Then *double* the probabilities obtained from the table to get the  $P$ -value for the two-sided  $F$  test.

The idea is that we calculate the probability in the upper tail and double to obtain the probability of all ratios on either side of 1 that are at least as improbable as that observed. Remember that the order of the degrees of freedom is important in using Table E.

## Example

### 7.22 Comparing calcium and placebo groups

Example 7.19 (page 462) recounts a medical experiment comparing the effects of calcium and a placebo on the blood pressure of black men. The analysis (Example 7.20) employed the pooled two-sample  $t$  procedures. Because these procedures require equal population standard deviations, it is tempting to first test

$$H_0: \sigma_1 = \sigma_2$$

$$H_a: \sigma_1 \neq \sigma_2$$

The larger of the two sample standard deviations is  $s = 8.743$  from 10 observations. The other is  $s = 5.901$  from 11 observations. The two-sided test statistic is therefore

$$F = \frac{s_1^2}{s_2^2} = \frac{8.74325}{4.012} = 2.20$$

We compare the calculated value  $F = 2.20$  with critical points for the  $F(9, 10)$  distribution. Table E shows that 2.20 is *less* than the 0.10 critical value of the  $F(9, 10)$  distribution, which is  $F^* = 2.35$ . Doubling 0.10, we know that the observed  $F$  falls short of the 0.20 significance level. The results are not significant at the 20% level (or any lower level). Statistical software shows that the exact upper-tail probability is 0.118, and hence  $P = 0.236$ . *If* the populations were Normal, the observed standard deviations would give little reason to suspect unequal population standard deviations. Because one of the populations shows some non-Normality, however, we cannot be fully confident of this conclusion.

## USE YOUR KNOWLEDGE

### 7.99 The $F$ statistic

The  $F$  statistic  $F = s_1^2/s_2^2$  is calculated from samples of size  $n_1 = 13$  and  $n_2 = 22$ .

- What is the upper critical value for this  $F$  when using the 0.05 significance level?
- In a test of equality of standard deviations against the two-sided alternative, this statistic has the value  $F = 2.45$ . Is this value significant at the 5% level? Is it significant at the 10% level?

## Robustness of Normal inference procedures

We have claimed that

- The  $t$  procedures for inference about means are quite robust against non-Normal population distributions. These procedures are particularly robust when the population distributions are symmetric and (for the two-sample case) when the two sample sizes are equal.
- The  $F$  test and other procedures for inference about variances are so lacking in robustness as to be of little use in practice.

Simulations with a large variety of non-Normal distributions support these claims. One set of simulations was carried out with samples of size 25 and used significance tests with fixed level  $\alpha = 0.05$ . The three types of tests studied were the one-sample and pooled two-sample  $t$  tests and the  $F$  test for comparing two variances.

The robustness of the one-sample and two-sample  $t$  procedures is remarkable. The true significance level remains between about 4% and 6% for a large range of

populations. The  $t$  test and the corresponding confidence intervals are among the most reliable tools that statisticians use. Remember, however, that outliers can greatly disturb the  $t$  procedures. Also, two-sample procedures are less robust when the sample sizes are not similar.



*The lack of robustness of the tests for variances is equally remarkable.* The true significance levels depart rapidly from the target 5% as the population distribution departs from Normality. The two-sided  $F$  test carried out with 5% critical values can have a true level of less than 1% or greater than 11% even in symmetric populations with no outliers. Results such as these are the basis for our recommendation that these procedures not be used.

## The power of the two-sample $t$ test

The two-sample  $t$  test is one of the most used statistical procedures. Unfortunately, because of inadequate planning, users frequently fail to find evidence for the effects that they believe to be true. Power calculations should be part of the planning of any statistical study. Information from a pilot study or previous research is needed.

In Section 7.1, we learned how to find an approximation for the power of the one-sample  $t$  test. The basic concepts (three steps) for the two-sample case are the same. Here, we give the exact method, which involves a new distribution, the **noncentral  $t$  distribution**. To perform the calculations, we simply need software to calculate probabilities for this distribution.

### noncentral $t$ distribution

We first present the method for the pooled two-sample  $t$  test, where the parameters are  $\mu_1 - \mu_2$ , and the common standard deviation is  $\sigma$ . We then describe modifications to get approximate results when we do not pool.

To find the power for the pooled two-sample  $t$  test, use the following steps. We consider only the case where the null hypothesis is  $\mu_1 - \mu_2 = 0$ .

1. Specify
  - (a) an alternative value for  $\mu_1 - \mu_2$  that you consider important to detect;
  - (b) the sample sizes,  $n_1$  and  $n_2$ ;
  - (c) a fixed significance level,  $\alpha$ ;

- (d) a guess at the standard deviation,  $\sigma$ ;
2. Find the degrees of freedom  $df = n_1 + n_2 - 2$  and the value of  $t^*$  that will lead to rejection of  $H_0$ .
  3. (a) Calculate the **noncentrality parameter**

$$\Delta = |\mu_1 - \mu_2| \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

(b) Find the power as the probability that a noncentral  $t$  random variable with degrees of freedom  $df$  and noncentrality parameter  $\delta$  will be greater than  $t^*$ . In SAS the command is `1-PROBT(tstar,df,delta)`. In R the command is `1-pt(tstar,df,delta)`. If you do not have software that can perform this calculation, you can approximate the power as the probability that a standard Normal random variable is greater than  $t^* - \delta$ , that is,  $P(z > t^* - \delta)$  and use Table A.

Note that the denominator in the noncentrality parameter,

$$\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

is our guess at the standard deviation for the difference between the sample means. Therefore, if we wanted to assess a possible study in terms of the margin of error for the estimated difference, we would examine  $t^*$  times this quantity.

If we do not assume that the standard deviations are equal, we need to guess both standard deviations and then combine these for our guess at the standard deviation:

$$\sqrt{\sigma_1^2 \frac{1}{n_1} + \sigma_2^2 \frac{1}{n_2}}$$

This guess is then used in the denominator of the noncentrality parameter. For the degrees of freedom, the conservative approximation is appropriate.

## Example

### 7.23 Planning a new study of calcium versus placebo groups

In Example 7.20 (page 464) we examined the effect of calcium on blood pressure by comparing the means of a treatment group and a placebo group using a pooled two-sample  $t$  test. The  $P$ -value was 0.059, failing to achieve the usual standard of 0.05 for statistical significance. Suppose that we wanted to plan a new study that would provide convincing evidence—say, at the 0.01

level—with high probability. Let's examine a study design with 45 subjects in each group ( $n_1 = n_2 = 45$ ) to see if this meets our goals.

**Step 1.** Based on our previous results, we choose  $\mu_1 - \mu_2 = 5$  as an alternative that we would like to be able to detect with  $\alpha = 0.01$ . For  $\sigma$  we use 7.4, our pooled estimate from Example 7.20.

**Step 2.** The degrees of freedom are  $n_1 + n_2 = 88$  which leads to  $t^* = 2.37$  for the significance test.

**Step 3.** The noncentrality parameter is

$$\Delta = 57.4145 + 145 = 51.56 = 3.21$$

Software gives the power as 0.7965, or 80%. The Normal approximation gives 0.7983, a very accurate result.

With this choice of sample sizes, we are just barely below 80% power. If we judge this to be enough power, we can proceed to the recruitment of our samples. With  $n_1 = n_2 = 45$ , we would expect the margin of error for a 95% confidence interval ( $t^* = 1.99$ ) for the difference in means to be

$$t^* \times 7.4145 + 145 = 1.99 \times 1.56 = 3.1$$

With software it is very easy to examine the effects of variations in a study design. In the preceding example, we might want to examine the power for  $\alpha = 0.05$  and for smaller sample sizes.

## USE YOUR KNOWLEDGE

### 7.100 Power and $\mu_1 - \mu_2$

If you repeat the calculation in Example 7.23 for other values of  $\mu_1 - \mu_2$  that are larger than 5, would you expect the power to be higher or lower than 0.7965? Why?

### 7.101 Power and the standard deviation

If the true population standard deviation were 7.1 instead of the 7.4 hypothesized in Example 7.23, would the power for this new experiment be greater or smaller than 0.7965? Explain.

## Section 7.3 Summary

Inference procedures for comparing the standard deviations of two Normal populations are based on the **F statistic**, which is the ratio of sample variances:

$$F = \frac{s_1^2}{s_2^2}$$

If an SRS of size  $n_1$  is drawn from the  $x_1$  population and an independent SRS of size  $n_2$  is drawn from the  $x_2$  population, the  $F$  statistic has the **F distribution**  $F(n_1 - 1, n_2 - 1)$  if the two population standard deviations  $\sigma_1$  and  $\sigma_2$  are in fact equal.

The **F test for equality of standard deviations** tests  $H_0: \sigma_1 = \sigma_2$  versus  $H_a: \sigma_1 \neq \sigma_2$  using the statistic

$$F = \frac{\text{larger } s^2}{\text{smaller } s^2}$$

and doubles the upper-tail probability to obtain the  $P$ -value.

The  $t$  procedures are quite **robust** when the distributions are not Normal. The  $F$  tests and other procedures for inference about the spread of one or more Normal distributions are so strongly affected by non-Normality that we do not recommend them for regular use.

The **power** of the pooled two-sample  $t$  test is found by first computing the critical value for the significance test, the degrees of freedom, and the **noncentrality parameter** for the alternative of interest. These are used to find the power from the **noncentral  $t$  distribution**. A Normal approximation works quite well. Calculating margins of error for various study designs and assumptions is an alternative procedure for evaluating designs.

## SECTION 7.3 Exercises

For Exercise 7.99, see page 476; and for Exercises 7.100 and 7.101, see page 479.

In all exercises calling for use of the  $F$  test, assume that both population distributions are very close to Normal. The actual data are not always sufficiently Normal to justify use of the  $F$  test.

### 7.102 Comparison of standard deviations

Here are some summary statistics from two independent samples from Normal distributions:

Sample	$n$	$s^2$
1	11	3.5
2	16	9.1

You want to test the null hypothesis that the two population standard deviations are equal versus the two-sided alternative at the 5% significance level.

- Calculate the test statistic.
- Find the appropriate value from Table E that you need to perform the significance test.
- What do you conclude?

### 7.103 Revisiting the eating-group comparison

Compare the standard deviations of weight loss in Example 7.16 (page 456). Give the test statistic, the degrees of freedom, and the  $P$ -value. Write a short summary of your analysis, including comments on the assumptions for the test.

### 7.104 A fat intake comparison

Compare the standard deviations of fat intake in Exercise 7.75 (page 469).

(a) Give the test statistic, the degrees of freedom, and the  $P$ -value. Write a short summary of your analysis, including comments on the assumptions for the test.

(b) Assume that the sample standard deviation for the late-eaters group is the value 8.2 given in Exercise 7.75. How large would the standard deviation in the early-eaters group need to be to reject the null hypothesis of equal standard deviations at the 5% level?

### 7.105 Revisiting the dust exposure study

The two-sample problem in Exercise 7.77 (page 470) compares drill and blast workers with outdoor concrete workers with respect to the total dust that they are exposed to in the workplace. Here it may be useful to know whether or not the standard deviations differ in the two groups. Perform the  $F$  test and summarize the results. Are you concerned about the assumptions here? Explain why or why not.

### 7.106 More on the dust exposure study

Exercise 7.78 (page 470) is similar to Exercise 7.77, but the response variable here is exposure to dust particles that can enter and stay in the lungs. Compare the standard deviations with a significance test and summarize the results. Be sure to comment on the assumptions.

### 7.107 Revisiting the size of trees in the north and south

The diameters of trees in the Wade Tract for random samples selected from the north and south halves of the tract are compared in Exercise 7.85 (page 471). Is there a statistically significant difference between the standard deviations for these two parts of the tract? Perform the significance test and summarize the results. Does the Normal assumption appear reasonable for these data? 

### 7.108 Revisiting the size of trees in the east and west

Tree diameters for the east and west halves of the Wade Tract are compared in Exercise 7.86 (page 472). Using the questions in the previous exercise as a guide, analyze these data. 

### 7.109 Revisiting the small-sample example

In Example 7.17 (page 457), we addressed a study with only 5 observations per group. 

(a) Is there a statistically significant difference between the standard deviations of these two groups? Perform the test using a significance level of 0.05 and state your conclusion.

(b) Using Table E, state the value that the ratio of variances would need to exceed for us to reject the null hypothesis (at the 5% level) that the standard deviations are equal. Also, report this value for sample sizes of  $n = 4, 3,$  and  $2$ . What does this suggest about the power of this test when sample sizes are small?

### 7.110 Planning a study to compare tree size

In Exercise 7.85 (page 471) DBH data for longleaf pine trees in two parts of the Wade Tract are compared. Suppose that you are planning a similar study in which you will measure the diameters of longleaf pine trees. Based on Exercise 7.85, you are willing to assume that the standard deviation for both halves is 20 cm. Suppose that a difference in mean DBH of 10 cm or more would be important to detect. You will use a  $t$  statistic and a two-sided alternative for the comparison.

- (a) Find the power if you randomly sample 20 trees from each area to be compared.
- (b) Repeat the calculations for 60 trees in each sample.
- (c) If you had to choose between the 20 and 60 trees per sample, which would you choose? Give reasons for your answer.



### 7.111 More on planning a study to compare tree size

Refer to the previous exercise. Find the two standard deviations from Exercise 7.85. Do the same for the data in Exercise 7.86, which is a similar setting. These are somewhat smaller than the assumed value that you used in the previous exercise. Explain why it is generally a better idea to assume a standard deviation that is larger than you expect than one that is smaller. Repeat the power calculations for some other reasonable values of  $\sigma$  and comment on the impact of the size of  $\sigma$  for planning the new study.

### 7.112 Planning a study to compare ad placement

Refer to Exercise 7.84 (page 471), where we compared trustworthiness ratings for ads from two different publications. Suppose that you are planning a similar study using two different publications that are not expected to show the differences seen when comparing the *Wall Street Journal* with the *National Enquirer*. You would like to detect a difference of 1.5 points using a two-sided significance test with a 5% level of significance. Based on Exercise 7.84, it is reasonable to use 1.6 as the value of the common standard deviation for planning purposes.

- (a) What is the power if you use sample sizes similar to those used in the previous study—for example, 65 for each publication?
- (b) Repeat the calculations for 100 in each group.
- (c) What sample size would you recommend for the new study?

## CHAPTER 7 Exercises

### 7.113 LSAT scores

The scores of four senior roommates on the Law School Admission Test (LSAT) are

156 133 147 122

Find the mean, the standard deviation, and the standard error of the mean. Is it appropriate to calculate a confidence interval based on these data? Explain why or why not.  LSAT

### 7.114 Converting a two-sided $P$ -value

You use statistical software to perform a significance test of the null hypothesis that two means are equal. The software reports a  $P$ -value for the two-sided alternative. Your alternative is that the first mean is greater than the second mean.

(a) The software reports  $t = 2.08$  with a  $P$ -value of 0.068. Would you reject  $H_0$  at  $\alpha$ ? Explain your answer.

(b) The software reports  $t = -2.08$  with a  $P$ -value of 0.068. Would you reject  $H_0$  at  $\alpha = 0.05$ ? Explain your answer.

### 7.115 Degrees of freedom and confidence interval width

As the degrees of freedom increase, the  $t$  distributions get closer and closer to the  $z$  ( $N(0, 1)$ ) distribution. One way to see this is to look at how the value of  $t^*$  for a 95% confidence interval changes with the degrees of freedom. Make a plot with degrees of freedom from 2 to 100 on the  $x$  axis and  $t^*$  on the  $y$  axis. Draw a horizontal line on the plot corresponding to the value of  $z^* = 1.96$ . Summarize the main features of the plot.

### 7.116 Degrees of freedom and $t^*$

Refer to the previous exercise. Make a similar plot for a 90% confidence interval. How do the main features of this plot compare with those of the plot in the previous exercise?

### 7.117 Sample size and margin of error

The margin of error for a confidence interval depends on the confidence level, the standard deviation, and the sample size. Fix the confidence level at 95% and the standard deviation at 1 to examine the effect of the sample size. Find the margin of error for sample sizes of 5 to 100 by 5s—that is, let  $n = 5, 10, 15, \dots, 100$ . Plot the margins of error versus the sample size and summarize the relationship.

## 7.118 More on sample size and margin of error

Refer to the previous exercise. Make a similar plot and summarize its features for a 99% confidence interval.

## 7.119 Which design?

The following situations all require inference about a mean or means. Identify each as (1) a single sample, (2) matched pairs, or (3) two independent samples. Explain your answers.

- (a) Your customers are college students. You are interested in comparing the interest in a new product that you are developing between those students who live in the dorms and those who live elsewhere.
- (b) Your customers are college students. You are interested in finding out which of two new product labels is more appealing.
- (c) Your customers are college students. You are interested in assessing their interest in a new product.

## 7.120 Which design?

The following situations all require inference about a mean or means. Identify each as (1) a single sample, (2) matched pairs, or (3) two independent samples. Explain your answers.

- (a) You want to estimate the average age of your store's customers.
- (b) You do an SRS survey of your customers every year. One of the questions on the survey asks about customer satisfaction on a seven-point scale with the response 1 indicating "very dissatisfied" and 7 indicating "very satisfied." You want to see if the mean customer satisfaction has improved from last year.
- (c) You ask an SRS of customers their opinions on each of two new floor plans for your store.

## 7.121 Number of critical food violations

The results of a major city's restaurant inspections are available through its online newspaper.<sup>38</sup> Critical food violations are those that put patrons at risk of getting sick and must immediately be corrected by the restaurant. An SRS of  $n = 200$  inspections from the more than 16,000 inspections since January 2009 were collected, resulting in  $\bar{x} = 0.83$  violations and  $s = 0.95$  violations.

- (a) Test the hypothesis that the average number of critical violations is less than 1.5 using a significance level of 0.05. State the two hypotheses, the test statistic, and  $P$ -value.
- (b) Construct a 95% confidence interval for the average number of critical violations and summarize your result.
- (c) Which of the two summaries (significance test versus confidence interval) do you find more helpful in this case? Explain your answer.
- (d) These data are integers ranging from 0 to 9. The data are also skewed to the right, with 70% of the values either a 0 or a 1. Given this information, do you think use of the  $t$  procedures is appropriate? Explain your answer.

### 7.122 Two-sample $t$ test versus matched pairs $t$ test

Consider the following data set. The data were actually collected in pairs, and each row represents a pair.  **PAIRED**

Group 1	Group 2
48.86	48.88
50.60	52.63
51.02	52.55
47.99	50.94
54.20	53.02
50.66	50.66
45.91	47.78
48.79	48.44
47.76	48.92
51.13	51.63

- (a) Suppose that we ignore the fact that the data were collected in pairs and mistakenly treat this as a two-sample problem. Compute the sample mean and variance for each group. Then compute the two-sample  $t$  statistic, degrees of freedom, and  $P$ -value for the two-sided alternative.
- (b) Now analyze the data in the proper way. Compute the sample mean and variance of the differences. Then compute the  $t$  statistic, degrees of freedom, and  $P$ -value.
- (c) Describe the differences in the two test results.

### 7.123 Two-sample $t$ test versus matched pairs $t$ test, continued

Refer to the previous exercise. Perhaps an easier way to see the major difference in the two analysis approaches for these data is by computing 95% confidence intervals for the mean difference.

- (a) Compute the 95% confidence interval using the two-sample  $t$  confidence interval.
- (b) Compute the 95% confidence interval using the matched pairs  $t$  confidence interval.
- (c) Compare the estimates (that is, the centers of the intervals) and margins of error. What is the major difference between the two approaches for these data?

### 7.124 Average service time

Recall the drive-thru study in Exercise 7.73 (page 469). Another benchmark that was measured was the service time. A summary of the results (in seconds) for two of the chains is shown below.

Chain	$n$	$\bar{x}$	$s$
Taco Bell	307	149.69	35.7
McDonald's	362	188.83	42.8

- (a) Is there a difference in the average service time between these two chains? Test the null hypothesis that the chains' average service time is the same. Use a significance level of 0.05.
- (b) Construct a 95% confidence interval for the difference in average service time.

(c) Lex plans to go to Taco Bell and Sam to McDonald's. Does the interval in part (b) contain the difference in their service times that they're likely to encounter? Explain your answer.

### 7.125 Interracial friendships in college

A study utilized the random roommate assignment process of a small college to investigate the interracial mix of friends among students in college.<sup>39</sup> As part of this study, the researchers looked at 238 white students who were randomly assigned a roommate in their first year and recorded the proportion of their friends (not including the first-year roommate) who were black. The following table summarizes the results, broken down by roommate race, for the middle of the first and third years of college.

Middle of First Year			
Randomly assigned	$n$	$\bar{x}$	$s$
Black roommate	41	0.085	0.134
White roommate	197	0.063	0.112
Middle of Third Year			
Randomly assigned	$n$	$\bar{x}$	$s$
Black roommate	41	0.146	0.243
White roommate	197	0.062	0.154

- (a) Proportions are not Normally distributed. Explain why it may still be appropriate to use the  $t$  procedures for these data.
- (b) For each year, state the null and alternative hypotheses for comparing these two groups.
- (c) For each year, perform the significance test at the  $\alpha = 0.05$  level, making sure to report the test statistic, degrees of freedom, and  $P$ -value.
- (d) Write a one-paragraph summary of your conclusions from these two tests.

### 7.126 Interracial friendships in college, continued

Refer to the previous exercise. For each year, construct a 95% confidence interval for the difference in means  $\mu_1 - \mu_2$  and describe how these intervals can be used to test the null hypotheses in part (b) of the previous exercise.

### 7.127 Alcohol consumption and body composition

Individuals who consume large amounts of alcohol do not use the calories from this source as efficiently as calories from other sources. One study examined the effects of moderate alcohol consumption on body composition and the intake of other foods. Fourteen subjects participated in a crossover design where they either drank wine for the first 6 weeks and then abstained for the next 6 weeks or vice versa.<sup>40</sup> During the period when they drank wine, the subjects, on average, lost 0.4 kilograms (kg) of body weight; when they did not drink wine, they lost an average of 1.1 kg. The standard deviation of the difference between the weight lost under these two conditions is 8.6 kg. During the wine period, they consumed an average of 2589 calories; with no wine, the mean consumption was 2575. The standard deviation of the difference was 210.

- (a) Compute the differences in means and the standard errors for comparing body weight and caloric

intake under the two experimental conditions.

(b) A report of the study indicated that there were no significant differences in these two outcome measures. Verify this result for each measure, giving the test statistic, degrees of freedom, and the  $P$ -value.

(c) One concern with studies such as this, with a small number of subjects, is that there may not be sufficient power to detect differences that are potentially important. Address this question by computing 95% confidence intervals for the two measures and discuss the information provided by the intervals.

(d) Here are some other characteristics of the study. The study periods lasted for 6 weeks. All subjects were males between the ages of 21 and 50 years who weighed between 68 and 91 kg. They were all from the same city. During the wine period, subjects were told to consume two 135-milliliter (ml) servings of red wine per day and no other alcohol. The entire 6-week supply was given to each subject at the beginning of the period. During the other period, subjects were instructed to refrain from any use of alcohol. All subjects reported that they complied with these instructions except for three subjects, who said that they drank no more than three to four 12-ounce bottles of beer during the no-alcohol period. Discuss how these factors could influence the interpretation of the results.

## 7.128 Brain training

The assessment of computerized brain-training programs is a rapidly growing area of research. Researchers are now focusing on who this training benefits most, what brain functions can be best improved, and which products are most effective. One study looked at 487 community-dwelling adults aged 65 and older, each randomly assigned to one of two training groups. In one group, the participants used a computerized program for 1 hour per day. In the other, DVD-based educational programs were shown with quizzes following each video. The training period lasted 8 weeks. The response was the improvement in a composite score obtained from an auditory memory/attention survey given before and after the 8 weeks.<sup>41</sup> The results are summarized in the following table.

Group	$n$	$\bar{x}$	$s$
Computer program	242	3.9	8.28
DVD program	245	1.8	8.33

(a) Given that there are other studies showing a benefit of computerized brain training, state the null and alternative hypotheses.

(b) Report the test statistic, its degrees of freedom, and the  $P$ -value. What is your conclusion using significance level  $\alpha = 0.05$ ?

(c) Can you conclude that this computerized brain training always improves a person's auditory memory better than the DVD program? If not, explain why.

## 7.129 Can mockingbirds learn to identify specific humans?

A central question in urban ecology is why some animals adapt well to the presence of humans and others do not. The following results summarize part of a study of the northern mockingbird (*Mimus polyglottos*) that took place on a campus of a large university.<sup>42</sup> For 4 consecutive days, the same human approached a nest and stood 1 meter away for 30 seconds, placing his or her hand on the rim of the nest. On the 5th day, a new person did the same thing. Each day, the distance of the human from the nest when the bird flushed was recorded. This was repeated for 24 nests. The human intruder varied his or her appearance (that is, wore different clothes) over the 4 days. We report results for only Days 1, 4, and 5 here. The response variable is flush distance measured in meters.

Day	Mean	<i>s</i>
1	6.1	4.9
4	15.1	7.3
5	4.9	5.3

- (a) Explain why this should be treated as a matched design.
- (b) Unfortunately, the research article does not provide the standard error of the difference, only the standard error of the mean flush distance for each day. However, we can use the general addition rule for variances (page 275) to approximate it. If we assume that the correlation between the flush distance at Day 1 and Day 4 for each nest is  $\rho = 0.40$ , what is the standard deviation for the difference in distance?
- (c) Using your result in part (b), test the hypothesis that there is no difference in the flush distance across these two days. Use a significance level of 0.05.
- (d) Repeat parts (b) and (c) but now compare Day 1 and Day 5, assuming a correlation between flush distances for each nest of  $\rho = 0.30$ .
- (e) Write a brief summary of your conclusions.

### 7.130 The wine makes the meal?

In one study, 39 diners were given a free glass of cabernet sauvignon wine to accompany a French meal.<sup>43</sup> Although the wine was identical, half the bottle labels claimed the wine was from California and the other half claimed it was from North Dakota. The following table summarizes the grams of entrée and wine consumed during the meal.

	Wine label	<i>n</i>	Mean	St. dev.
Entrée	California	24	499.8	87.2
	North Dakota	15	439.0	89.2
Wine	California	24	100.8	23.3
	North Dakota	15	110.4	9.0

Did the patrons who thought that the wine was from California consume more? Analyze the data and write a report summarizing your work. Be sure to include details regarding the statistical methods you used, your assumptions, and your conclusions.

### 7.131 Study design information

In the previous study, diners were seated alone or in groups of two, three, four, and, in one case, nine (for a total of  $n = 16$  tables). Also, each table, not each patron, was randomly assigned a particular wine label. Does this information alter how you might do the analysis in the previous problem? Explain your answer.

### 7.132 Analysis of tree size using the complete data set

The data used in Exercises 7.31 (page 443), 7.85, and 7.86 (pages 471 and 472) were obtained by taking simple random samples from the 584 longleaf pine trees that were measured in the Wade Tract. The entire data set is given in the WADE data set. Find the 95% confidence interval for the mean DBH using the entire data set, and compare this interval with the one that you calculated in Exercise 7.31. Write a report about these data. Include comments on the effect of the sample size on

the margin of error, the distribution of the data, the appropriateness of the Normality-based methods for this problem, and the generalizability of the results to other similar stands of longleaf pine or other kinds of trees in this area of the United States and other areas. 

### 7.133 More on conditions for inference

Suppose that your state contains 85 school corporations and each corporation reports its expenditures per pupil. Is it proper to apply the one-sample  $t$  method to these data to give a 95% confidence interval for the average expenditure per pupil? Explain your answer.

### 7.134 A comparison of female high school students

A study was performed to determine the prevalence of the female athlete triad (low energy availability, menstrual dysfunction, and low bone mineral density) in high school students.<sup>44</sup> A total of 80 high school athletes and 80 sedentary students were assessed. The following table summarizes several measured characteristics:

Characteristic	Athletes		Sedentary	
	$\bar{x}$	$s$	$\bar{x}$	$s$
Body fat (%)	25.61	5.54	32.51	8.05
Body mass index	21.60	2.46	26.41	2.73
Calcium deficit (mg)	297.13	516.63	580.54	372.77
Glasses of milk/day	2.21	1.46	1.82	1.24

(a) For each of the characteristics, test the hypothesis that the means are the same in the two groups. Use a significance level of 0.05 for each test.

(b) Write a short report summarizing your results.

### 7.135 Competitive prices?

A retailer entered into an exclusive agreement with a supplier who guaranteed to provide all products at competitive prices. The retailer eventually began to purchase supplies from other vendors who offered better prices. The original supplier filed a legal action claiming violation of the agreement. In defense, the retailer had an audit performed on a random sample of invoices. For each audited invoice, all purchases made from other suppliers were examined and the prices were compared with those offered by the original supplier. For each invoice, the percent of purchases for which the alternate supplier offered a lower price than the original supplier was recorded.<sup>45</sup> Here are the data:

0	1000	10033	34	10048	78	10077	10038
68	10079	100100	100100	100100	10089	100100	

Report the average of the percents with a 95% margin of error. Do the sample invoices suggest that the original supplier's prices are not competitive on the average? 

### 7.136 Weight-loss programs

In a study of the effectiveness of weight-loss programs, 47 subjects who were at least 20% overweight took part in a group support program for 10 weeks. Private weighings determined each

subject's weight at the beginning of the program and 6 months after the program's end. The matched pairs  $t$  test was used to assess the significance of the average weight loss. The paper reporting the study said, "The subjects lost a significant amount of weight over time,  $t(46) = 4.68, p > 0.01$ ." It is common to report the results of statistical tests in this abbreviated style.<sup>46</sup>

- (a) Why was the matched pairs statistic appropriate?
- (b) Explain to someone who knows no statistics but is interested in weight-loss programs what the practical conclusion is.
- (c) The paper follows the tradition of reporting significance only at fixed levels such as  $\alpha = 0.01$ . In fact, the results are more significant than " $p > 0.01$ " suggests. What can you say about the  $P$ -value of the  $t$  test?

### 7.137 Do women perform better in school?

Some research suggests that women perform better than men in school, but men score higher on standardized tests. Table 1.3 (page 29) presents data on a measure of school performance, grade point average (GPA), and a standardized test, IQ, for 78 seventh-grade students. Do these data lend further support to the previously found gender differences? Give graphical displays of the data and describe the distributions. Use significance tests and confidence intervals to examine this question, and prepare a short report summarizing your findings.  **GRADES**

### 7.138 Self-concept and school performance

Refer to the previous exercise. Although self-concept in this study was measured on a scale with values in the data set ranging from 20 to 80, many prefer to think of this kind of variable as having only two possible values: low self-concept or high self-concept. Find the median of the self-concept scores in Table 1.3, and define those students with scores at or below the median to be low-self-concept students and those with scores above the median to be high-self-concept students. Do high-self-concept students have GPAs that differ from those of low-self-concept students? What about IQ? Prepare a report addressing these questions. Be sure to include graphical and numerical summaries and confidence intervals, and state clearly the details of significance tests.  **GRADES**

### 7.139 Behavior of pet owners

On the morning of March 5, 1996, a train with 14 tankers of propane derailed near the center of the small Wisconsin town of Weyauwega. Six of the tankers were ruptured and burning when the 1700 residents were ordered to evacuate the town. Researchers study disasters like this so that effective relief efforts can be designed for future disasters. About half the households with pets did not evacuate all their pets. A study conducted after the derailment focused on problems associated with retrieval of the pets after the evacuation and characteristics of the pet owners. One of the scales measured "commitment to adult animals," and the people who evacuated all or some of their pets were compared with those who did not evacuate any of their pets. Higher scores indicate that the pet owner is more likely to take actions that benefit the pet.<sup>47</sup> Here are the data summaries:

Group	$n$	$\bar{x}$	$s$
Evacuated all or some pets	116	7.95	3.62
Did not evacuate any pets	125	6.26	3.56

Analyze the data and prepare a short report describing the results.

## 7.140 Occupation and diet

Do various occupational groups differ in their diets? A British study of this question compared 98 drivers and 83 conductors of London double-decker buses.<sup>48</sup> The conductors' jobs require more physical activity. The article reporting the study gives the data as "Mean daily consumption ( $\pm$ se)." Here are some of the study results:

	Drivers	Conductors
Total calories	2821 $\pm$ 44	2844 $\pm$ 48
Alcohol (grams)	0.24 $\pm$ 0.06	0.39 $\pm$ 0.11

- What does "se" stand for? Give  $\bar{x}$  and  $s$  for each of the four sets of measurements.
- Is there significant evidence at the 5% level that conductors consume more calories per day than do drivers? Use the two-sample  $t$  method to give a  $P$ -value, and then assess significance.
- How significant is the observed difference in mean alcohol consumption? Use two-sample  $t$  methods to obtain the  $P$ -value.
- Give a 95% confidence interval for the mean daily alcohol consumption of London double-decker bus conductors.
- Give a 99% confidence interval for the difference in mean daily alcohol consumption between drivers and conductors.

## 7.141 Occupation and diet, continued

Use of the pooled two-sample  $t$  test is justified in part (b) of the previous exercise. Explain why. Find the  $P$ -value for the pooled  $t$  statistic, and compare it with your result in the previous exercise.

## 7.142 Conditions for inference

The report cited in Exercise 7.140 says that the distributions of alcohol consumption among the individuals studied are "grossly skew."

- Do you think that this skewness prevents the use of the two-sample  $t$  test for equality of means? Explain your answer.
- Do you think that the skewness of the distributions prevents the use of the  $F$  test for equality of standard deviations? Explain your answer.

## 7.143 Different methods of teaching reading

In the READ data set, the response variable Post3 is to be compared for three methods of teaching reading. The Basal method is the standard, or control, method, and the two new methods are DRTA and Strat. We can use the methods of this chapter to compare Basal with DRTA and Basal with Strat. Note that to make comparisons among three treatments it is more appropriate to use the procedures that we will learn in Chapter 12. 

- Is the mean reading score with the DRTA method higher than that for the Basal method? Perform an analysis to answer this question, and summarize your results.
- Answer part (a) for the Strat method in place of DRTA.

## 7.144 Sample size calculation

Example 7.13 (page 449) tells us that the mean height of 10-year-old girls is  $N(56.4, 2.7)$  and for boys it is  $N(55.7, 3.8)$ . The null hypothesis that the mean heights of 10-year-old boys and girls are equal is clearly false. The difference in mean heights is  $56.4 - 55.7 = 0.7$  inch. Small differences such as this can require large sample sizes to detect. To simplify our calculations, let's assume that the standard deviations are the same, say  $\sigma = 3.2$ , and that we will measure the heights of an equal number of girls and boys. How many would we need to measure to have a 90% chance of detecting the (true) alternative hypothesis?

# 8 Inference for Proportions

CHAPTER



**8.1 Inference for a Single Proportion**

**8.2 Comparing Two Proportions**

## Introduction

We frequently collect data on *categorical variables*, such as whether or not a person is employed, the brand name of a cell phone, or the country where a college student studies abroad. When we record categorical variables, our data consist of *counts* or of *percents* obtained from counts.

In these settings, our goal is to say something about the corresponding *population proportions*. Just as in the case of inference about population means, we may be concerned with a single population or with comparing two populations. Inference about one or two proportions is very similar to inference about means, which we discussed in Chapter 7. In particular, inference for both means and proportions is based on sampling distributions that are approximately Normal.

We begin in Section 8.1 with inference about a single population proportion. Section 8.2 concerns methods for comparing two proportions.

## 8.1 Inference for a Single Proportion

When you complete this section, you will be able to

- Identify the sample proportion, the sample size, and the count for a single proportion. Use this information to estimate the population proportion.
- Describe the relationship between the population proportion and the sample proportion.
- Identify the standard error for a sample proportion and the margin of error for confidence level  $C$ .
- Apply the guidelines for when to use the large-sample confidence interval for a population proportion.
- Find and interpret the large-sample confidence interval for a single proportion.
- Apply the guidelines for when to use the large-sample significance test for a population proportion.
- Use the large-sample significance test to test a null hypothesis about a population proportion.
- Find the sample size needed for a desired margin of error.

We want to estimate the proportion  $p$  of some characteristic in a large population. For example, we may want to know the proportion of likely voters who approve of the president's conduct in office. We select a simple random sample (SRS) of size  $n$  from the population and record the count  $X$  of "successes" (such as "Yes" answers to a question about the president). We will use "success" to represent the characteristic of interest. The sample proportion of successes  $\hat{p} = X/n$  estimates the unknown population proportion  $p$ . If the population is much larger than the sample (say, at least 20 times as large), the count  $X$  has approximately the binomial distribution  $B(n, p)$ .<sup>1</sup> In statistical terms, we are concerned with inference about the probability  $p$  of a success in the binomial setting.

 **LOOK BACK**  
sample proportion, p. 321

## Example

### 8.1 Take a break from Facebook



A Pew Internet survey reported that 61% of Facebook users have taken a voluntary break from Facebook of several weeks or more at one time or another. The survey contacted 1006 adults living in the United States by landline and cell phone. The 525 people who reported that they were Facebook users were asked, “Have you ever voluntarily taken a break from Facebook for a period of several weeks or more?” A total of 320 responded, “Yes, I have done this.”<sup>2</sup> Here,  $p$  is the proportion of adults in the population of Facebook users who have taken a break of several weeks or more, and the sample proportion  $\hat{p}$  is

$$\hat{p} = \frac{X}{n} = \frac{320}{525} = 0.6095$$

Pew uses the sample proportion  $\hat{p}$  to estimate the population proportion  $p$ . Pew estimates that 61% of all adult Facebook users in the United States have taken a break from using Facebook for several weeks or more.

## USE YOUR KNOWLEDGE

### 8.1 Smartphones and purchases

A Google research study asked 5013 smartphone users about how they used their phones. In response to a question about purchases, 2657 reported that they purchased an item after using their smartphone to search for information about the item.<sup>3</sup>

- What is the sample size  $n$  for this survey?
- In this setting, describe the population proportion  $p$  in a short sentence.
- What is the count  $X$ ? Describe the count in a short sentence.
- Find the sample proportion  $\hat{p}$ .

## 8.2 Past usage of Facebook

Refer to the Pew Internet survey described in Example 8.1. There were 334 Internet users who don't use Facebook. Of these, 67 reported that they have used Facebook in the past.

- What is the sample size  $n$  for the population of Internet users who don't use Facebook?
- In this setting, describe the population proportion  $p$  in a short sentence.
- What is the count  $X$  of Internet users who don't use Facebook but have used Facebook in the past?
- Find the sample proportion  $\hat{p}$ .

If the sample size  $n$  is very small, we must base tests and confidence intervals for  $p$  on the binomial distributions. These are awkward to work with because of the discreteness of the binomial distributions.<sup>4</sup> But we know that when the sample is large, both the count  $X$  and the sample proportion  $\hat{p}$  are approximately Normal. We will consider only inference procedures based on the Normal approximation. These procedures are similar to those for inference about the mean of a Normal distribution.

### Large-sample confidence interval for a single proportion

The unknown population proportion  $p$  is estimated by the sample proportion  $\hat{p}=X/n$ . If the sample size  $n$  is sufficiently large,  $\hat{p}$  has approximately the Normal distribution, with mean  $\mu_{\hat{p}}=p$  and standard deviation  $\sigma_{\hat{p}}=p(1-p)/n$ . This means that approximately 95% of the time  $\hat{p}$  will be within  $2p(1-p)/n$  of the unknown population proportion  $p$ .



**Normal approximation for proportions, p. 332**

Note that the standard deviation  $\sigma_{\hat{p}}$  depends upon the unknown parameter  $p$ . To estimate this standard deviation using the data, we replace  $p$  in the formula by the sample proportion  $\hat{p}$ . As we did in Chapter 7, we use the term *standard error* for the standard deviation of a statistic that is estimated from data. Here is a summary of the procedure.



**standard error, p. 418**

## LARGE-SAMPLE CONFIDENCE INTERVAL FOR A

## POPULATION PROPORTION

Choose an SRS of size  $n$  from a large population with an unknown proportion  $p$  of successes. The **sample proportion** is

$$\hat{p} = X/n$$

where  $X$  is the number of successes. The **standard error of  $\hat{p}$**  is

$$SE_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n}$$

and the **margin of error** for confidence level  $C$  is

$$m = z^* SE_{\hat{p}}$$

where the critical value  $z^*$  is the value for the standard Normal density curve with area  $C$  between  $-z^*$  and  $z^*$ .

An **approximate level  $C$  confidence interval** for  $p$  is

$$\hat{p} \pm m$$

Use this interval for 90%, 95%, or 99% confidence when the number of successes and the number of failures are both at least 10.

Table D includes a line at the bottom with values of  $z^*$  for selected values of  $C$ . Use Table A for other values of  $C$ .

### Example

#### 8.2 Inference for Facebook breaks

The sample survey in Example 8.1 found that 320 of a sample of 525 Facebook users took a break from Facebook for several weeks or more. In that example we calculated  $\hat{p} = 0.6095$ . The standard error is

$$SE_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{0.6095(1-0.6095)/525} = 0.02129$$

The  $z^*$  critical value for 95% confidence is  $z^* = 1.96$ , so the margin of error is

$$m = 1.96 SE_{\hat{p}} = (1.96)(0.02129) = 0.04173$$

The confidence interval is

$$\hat{p} \pm m = 0.61 \pm 0.04$$

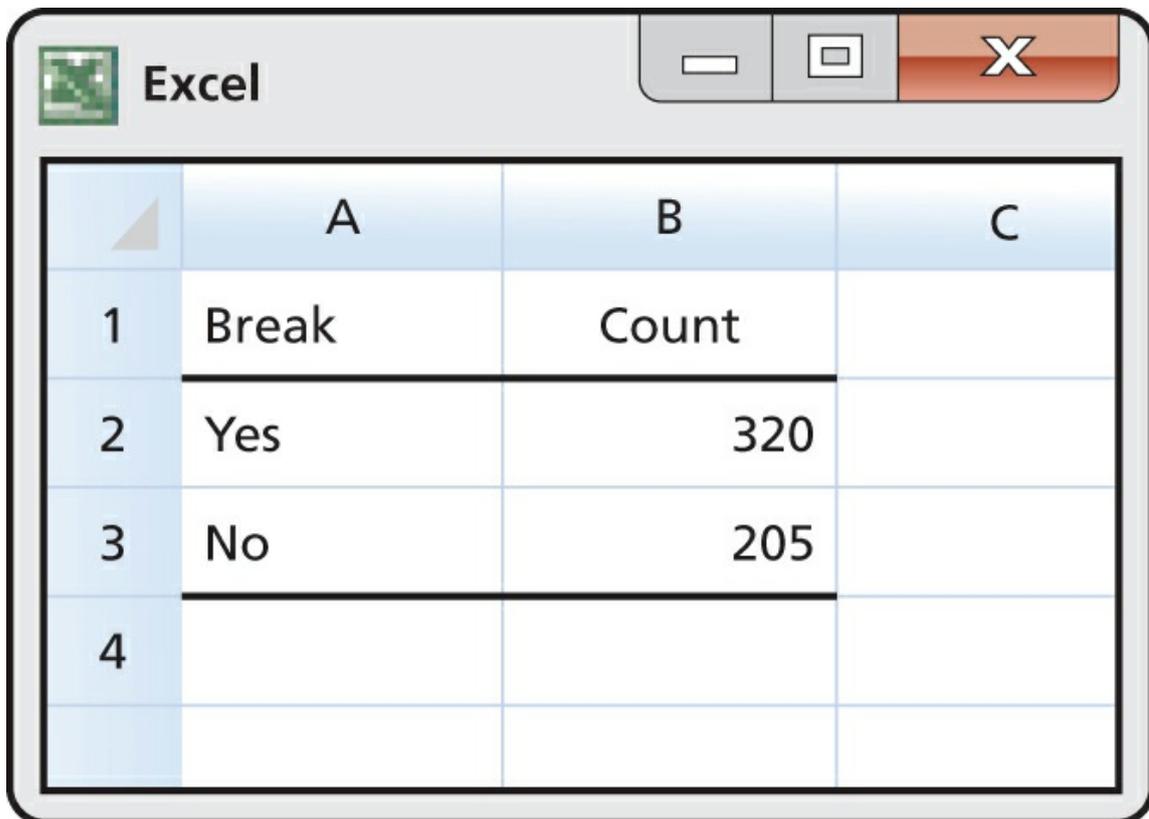
We are 95% confident that between 57% and 65% of Facebook users took a voluntary break of several weeks or more.

In performing these calculations, we have kept a large number of digits for our intermediate calculations. However, when reporting the results, we prefer to use rounded values: for example, 61% with a margin of error of 4%. In this way we focus attention on our major findings. There is no important information to be gained by reporting 0.6095 with a margin of error of 0.04173.

*Remember that the margin of error in any confidence interval includes only random sampling error.* If people do not respond honestly to the questions asked, for example, your estimate is likely to miss by more than the margin of error.



Although the calculations for statistical inference for a single proportion are relatively straightforward and can be done with a calculator or in a spreadsheet, we prefer to use software.



The image shows a screenshot of an Excel spreadsheet window. The window title is "Excel" and it has standard minimize, maximize, and close buttons. The spreadsheet contains the following data:

	A	B	C
1	Break	Count	
2	Yes	320	
3	No	205	
4			

**Figure 8.1**

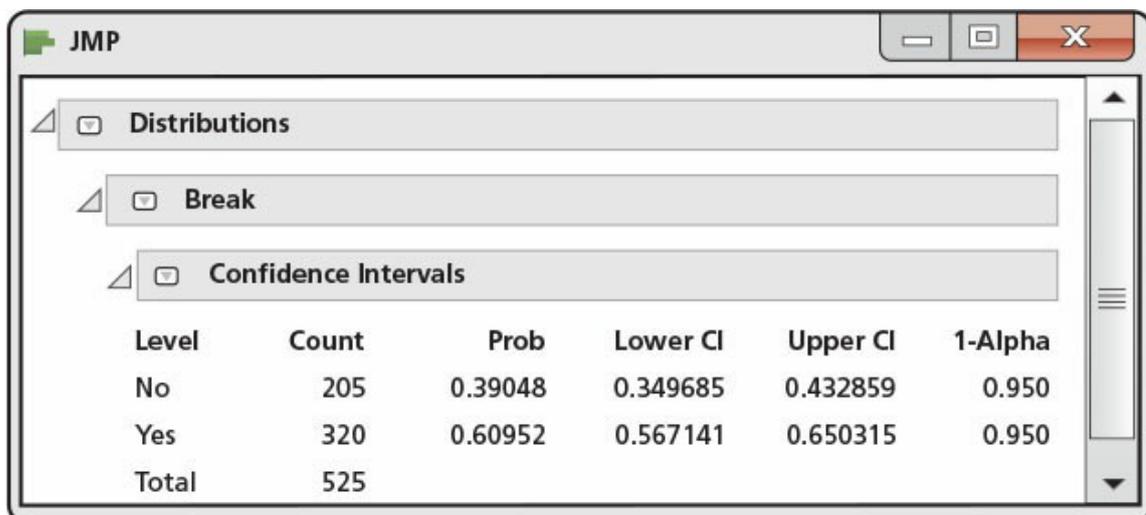
The Facebook break data in an Excel spreadsheet for the confidence interval in Example 8.3.

## Example

### 8.3 Facebook break confidence interval using software

Figure 8.1 shows a spreadsheet that could be used as input for statistical software that calculates a confidence interval for a proportion for our Facebook break example. Note that 525 is the number of cases for this example. The sheet specifies a value for each of these cases: there are 320 cases with the value “Yes” and 205 cases with the value “No.” An alternative sheet would list all 525 cases with the values for each case.

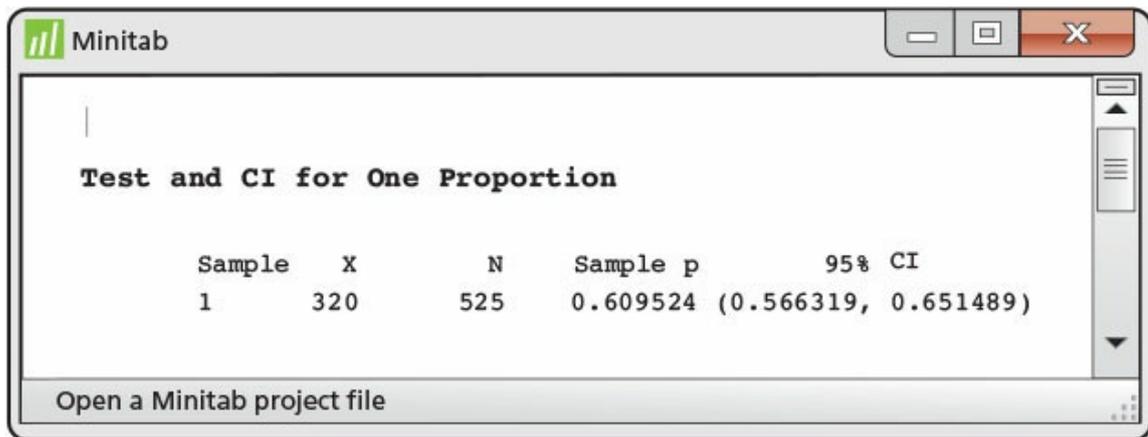
Figure 8.2 gives output from JMP, Minitab, and SAS for these data. Each is a little different but it is easy to find what we need. For JMP, the confidence interval is on the line with “Level” equal to “Yes” under the headings “Lower CL” and “Upper CL.” Minitab gives the output in the form of an interval under the heading “95% CI.” SAS reports the interval calculated in two different ways and uses the labels “95% Lower Conf Limit” and “95% Upper Conf Limit.”



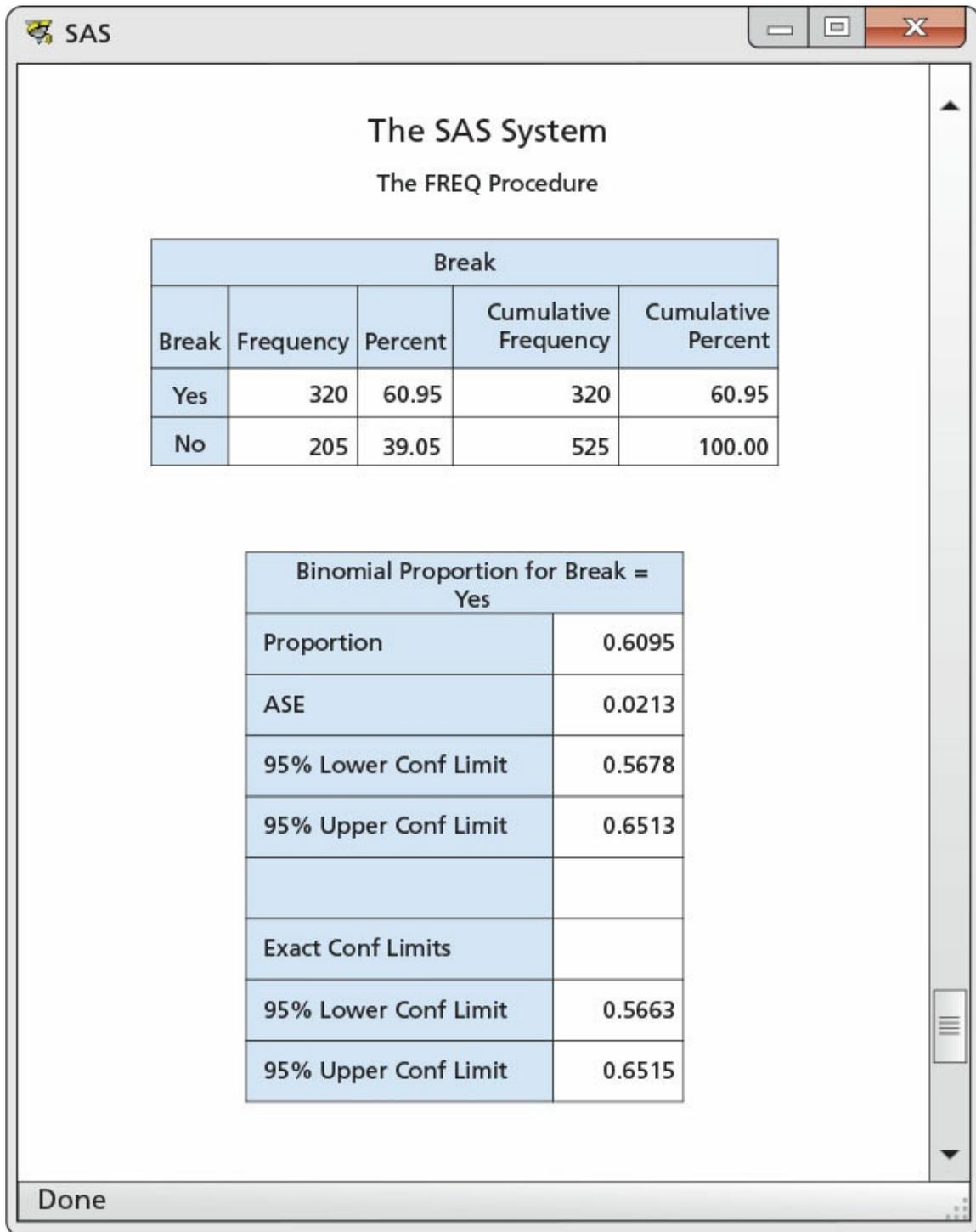
The screenshot shows the JMP software interface. The main window displays a table of results under the heading "Confidence Intervals". The table has six columns: Level, Count, Prob, Lower CI, Upper CI, and 1-Alpha. The rows represent the "No" and "Yes" levels, along with a "Total" row. The "Yes" row shows a Lower CI of 0.567141 and an Upper CI of 0.650315.

Level	Count	Prob	Lower CI	Upper CI	1-Alpha
No	205	0.39048	0.349685	0.432859	0.950
Yes	320	0.60952	0.567141	0.650315	0.950
Total	525				

(a) JMP



(b) Minitab



(c) SAS

**Figure 8.2**

(a) JMP, (b) Minitab, and (c) SAS output for the Facebook break confidence interval in Example 8.3.

As usual, the output reports more digits than are useful. *When you use software, be sure to think about how many digits are meaningful for your purposes. Do not clutter your report with information that is not meaningful.*



We recommend the large-sample confidence interval for 90%, 95%, and 99% confidence whenever the number of successes and the number of failures are both at least 10. For smaller sample sizes, we recommend exact methods that use the binomial distribution. These are available as the default or as options in many statistical software packages and we do not cover them here. There is also an intermediate case between large samples and very small samples where a slight modification of the large-sample approach works quite well.<sup>5</sup> This method is called the “plus four” procedure and is described next.

## USE YOUR KNOWLEDGE

### 8.3 Smartphones and purchases

Refer to Exercise 8.1 (page 489).

- Find  $SE_{\hat{p}}$ , the standard error of  $\hat{p}$ .
- Give the 95% confidence interval for  $p$  in the form of estimate plus or minus the margin of error.
- Give the confidence interval as an interval of percents.

### 8.4 Past usage of Facebook

Refer to Exercise 8.2 (page 489).

- Find  $SE_{\hat{p}}$ , the standard error of  $\hat{p}$ .
- Give the 95% confidence interval for  $p$  in the form of estimate plus or minus the margin of error.
- Give the confidence interval as an interval of percents.

## BEYOND THE BASICS

### The plus four confidence interval for a single proportion

Computer studies reveal that confidence intervals based on the large-sample approach can be quite inaccurate when the number of successes and the number of failures are not at least 10. When this occurs, a simple adjustment to the confidence interval works very well in practice. The adjustment is based on assuming that the sample contains 4 additional observations, 2 of which are successes and 2 of which are failures. The estimator of the population proportion based on this *plus four* rule is

$$\tilde{p} = \frac{X+2}{n+4}$$

This estimate was first suggested by Edwin Bidwell Wilson in 1927, and we call it the **plus four estimate**. The confidence interval is based on the  $z$  statistic obtained by standardizing the plus four estimate  $\tilde{p}$ . Because  $\tilde{p}$  is the sample proportion for our modified sample of size  $n+4$ , it isn't surprising that the distribution of  $\tilde{p}$  is close to the Normal distribution with mean  $p$  and standard deviation  $p(1-p)/(n+4)$ . To get a confidence interval, we estimate  $p$  by  $\tilde{p}$  in this standard deviation to get the standard error of  $\tilde{p}$ . Here is an example.

plus four estimate

## Example

### 8.4 Percent of equol producers



Research has shown that there are many health benefits associated with a diet that contains soy foods. Substances in soy called isoflavones are known to be responsible for these benefits. When soy foods are consumed, some subjects produce a chemical called equol, and it is thought that production of equol is a key factor in the health benefits of a soy diet. Unfortunately, not all people are equol producers; there appear to be two distinct subpopulations: equol producers and equol nonproducers.

A nutrition researcher planning some bone health experiments would like to include some equol producers and some nonproducers among her subjects. A preliminary sample of 12 female subjects were measured, and 4 were found

to be equol producers. We would like to estimate the proportion of equol producers in the population from which this researcher will draw her subjects.

The plus four estimate of the proportion of equol producers is

$$\tilde{p} = \frac{4+212+4}{616} = 0.375$$

For a 95% confidence interval, we use Table D to find  $z^* = 1.96$ . We first compute the standard error

$$\begin{aligned} SE_{\hat{p}} &= \sqrt{\tilde{p}(1-\tilde{p})/n+4} \\ &= \sqrt{(0.375)(1-0.375)/16} \\ &= 0.12103 \end{aligned}$$

and then the margin of error

$$\begin{aligned} m &= z^* SE_{\hat{p}} \\ &= (1.96)(0.12103) \\ &= 0.237 \end{aligned}$$

So the confidence interval is

$$\begin{aligned} \tilde{p} \pm m &= 0.375 \pm 0.237 \\ &= (0.138, 0.612) \end{aligned}$$

We estimate with 95% confidence that between 14% and 61% of women from this population are equol producers. Note that the interval is very wide because the sample size is very small.

If the true proportion of equol users is near 14%, the lower limit of this interval, there may not be a sufficient number of equol producers in the study if subjects are tested only after they are enrolled in the experiment. It may be necessary to determine whether or not a potential subject is an equol producer. The study could then be designed to have the same number of equol producers and nonproducers.

## Significance test for a single proportion

Recall that the sample proportion  $\hat{p} = X/n$  is approximately Normal, with mean  $\mu_{\hat{p}} = p$  and standard deviation  $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$ . For confidence intervals, we substitute  $\hat{p}$  for  $p$  in the last expression to obtain the standard error. When performing a significance test, however, the null hypothesis specifies a value for  $p$ , and we assume that this is the true value when calculating the  $P$ -value. Therefore, when we test  $H_0: p = p_0$ , we substitute  $p_0$  into the expression for  $\sigma_{\hat{p}}$  and then standardize  $\hat{p}$ . Here are the details.



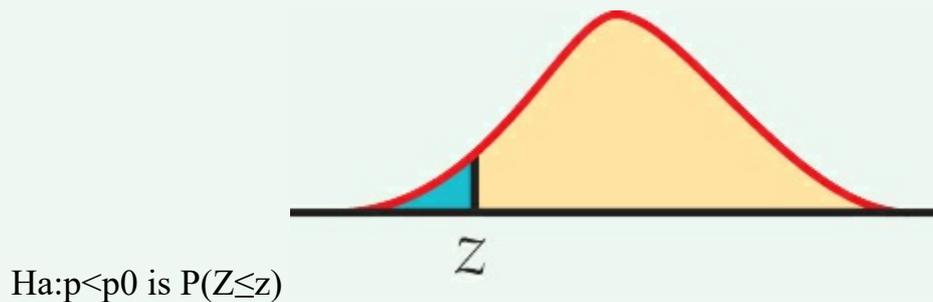
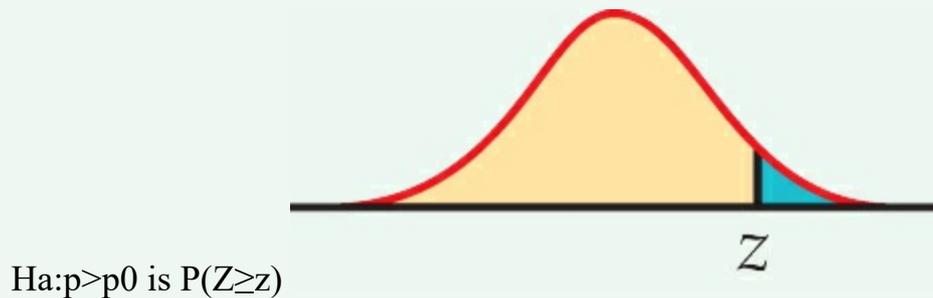
**Normal approximation for proportions, p. 332**

## LARGE-SAMPLE SIGNIFICANCE TEST FOR A POPULATION PROPORTION

Draw an SRS of size  $n$  from a large population with an unknown proportion  $p$  of successes. To test the hypothesis  $H_0: p=p_0$ , compute the  $z$  statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}} \sqrt{n}$$

In terms of a standard Normal random variable  $Z$ , the approximate  $P$ -value for a test of  $H_0$  against



We recommend the large-sample  $z$  significance test as long as the expected number of successes,  $np_0$ , and the expected number of failures,  $n(1-p_0)$ , are both greater than 10.

If the expected numbers of successes and failures are not both greater than 10, or if the population is less than 20 times as large as the sample, other procedures should be used. One such approach is to use the binomial distribution as we did with the sign test. Here is a large-sample example.

← LOOK BACK

## Example

### 8.5 Comparing two sunblock lotions



#### SUNBLOCK

Your company produces a sunblock lotion designed to protect the skin from both UVA and UVB exposure to the sun. You hire a company to compare your product with the product sold by your major competitor. The testing company exposes skin on the backs of a sample of 20 people to UVA and UVB rays and measures the protection provided by each product. For 13 of the subjects, your product provided better protection, while for the other 7 subjects, your competitor's product provided better protection.



Do you have evidence to support a commercial claiming that your product provides superior UVA and UVB protection? For the data we have  $n=20$  subjects and  $X=13$  successes. The parameter  $p$  is the proportion of people who would receive superior UVA and UVB protection from your product. To answer the claim question, we test

$$H_0:p=0.5$$

$$H_a:p\neq 0.5$$

The expected numbers of successes (your product provides better protection) and failures (your competitor's product provides better protection) are  $20 \times 0.5 = 10$  and  $20 \times 0.5 = 10$ . Both are at least 10, so we can use the  $z$  test. The sample proportion is

$$\hat{p} = X/n = 13/20 = 0.65$$

The test statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}} = \frac{0.65 - 0.5}{\sqrt{(0.5)(0.5)}} = 1.34$$

From Table A we find  $P(Z < 1.34) = 0.9099$ , so the probability in the upper tail is  $1 - 0.9099 = 0.0901$ . The  $P$ -value is the area in both tails,  $P = 2 \times 0.0901 = 0.1802$ .

We conclude that the sunblock testing data are compatible with the hypothesis of no difference between your product and your competitor's product ( $\hat{p} = 0.65$ ,  $z = 1.34$ ,  $P = 0.18$ ). The data do not support your proposed advertising claim.

Note that we have used the two-sided alternative for this example. In settings like this, we must start with the view that either product could be better if we want to prove a claim of superiority. Thinking or hoping that your product is superior cannot be used to justify a one-sided test.

Although these calculations are not particularly difficult to do using a calculator, we prefer to use software. Here are some details.

## Example

### 8.6 Sunblock significance tests using software

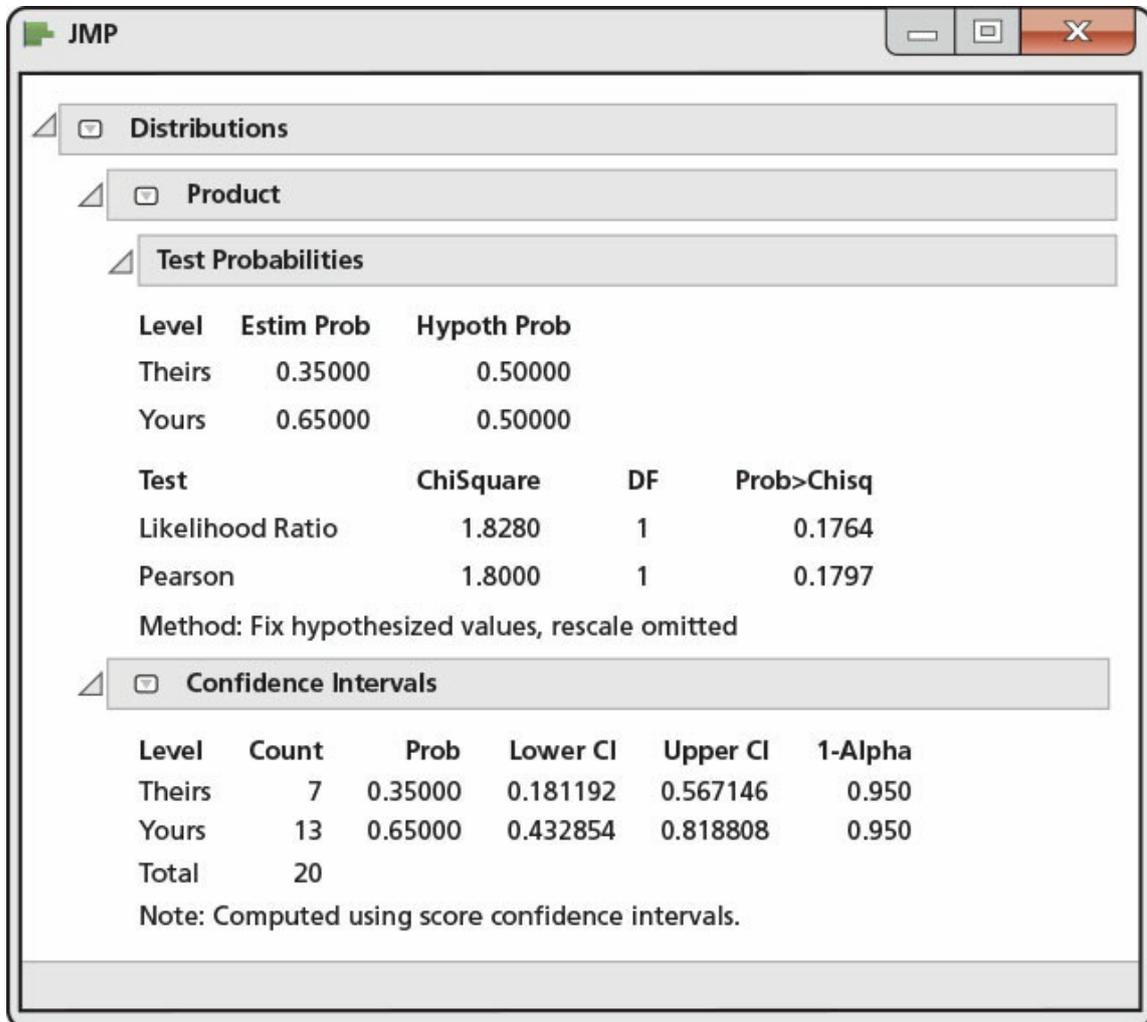


SUNBLOCK

JMP, Minitab, and SAS outputs for the analysis in Example 8.5 appear in Figure 8.3. JMP uses a slightly different way of reporting the results. Two ways of performing the significance test are labeled in the column "Test." The one that corresponds to the procedure that we have used is on the second line, labeled "Pearson." The  $P$ -value under the heading "Prob > Chisq" is 0.1797,

which is very close to the 0.1802 that we calculated using Table A.

Minitab reports the value of the test statistic  $z$  and the  $P$ -value is rounded to 0.180. SAS reports the  $P$ -value on the last line as 0.1797, the same as the value given in the JMP output.



Level	Estim Prob	Hypoth Prob
Theirs	0.35000	0.50000
Yours	0.65000	0.50000

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	1.8280	1	0.1764
Pearson	1.8000	1	0.1797

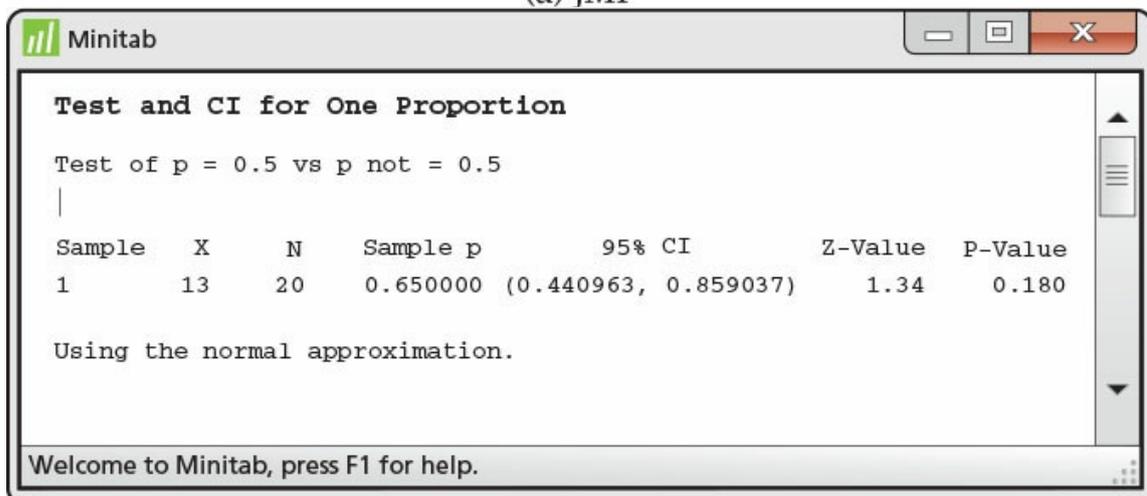
Method: Fix hypothesized values, rescale omitted

Level	Count	Prob	Lower CI	Upper CI	1-Alpha
Theirs	7	0.35000	0.181192	0.567146	0.950
Yours	13	0.65000	0.432854	0.818808	0.950
Total	20				

Note: Computed using score confidence intervals.

(a) JMP



**Test and CI for One Proportion**

Test of  $p = 0.5$  vs  $p \text{ not} = 0.5$

Sample	X	N	Sample p	95% CI	Z-Value	P-Value
1	13	20	0.650000	(0.440963, 0.859037)	1.34	0.180

Using the normal approximation.

Welcome to Minitab, press F1 for help.

(b) Minitab

SAS

### The SAS System

The FREQ Procedure

Product				
Product	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Yours	13	65.00	13	65.00
Theirs	7	35.00	20	100.00

Binomial Proportion for Product = Yours	
Proportion	0.6500
ASE	0.1067
95% Lower Conf Limit	0.4410
95% Upper Conf Limit	0.8590
Exact Conf Limits	
95% Lower Conf Limit	0.4078
95% Upper Conf Limit	0.8461

Test of H0: Proportion = 0.5	
ASE under H0	0.1118
Z	1.3416
One-sided Pr > Z	0.0899
Two-sided Pr >  Z	0.1797

Sample Size = 20

Done

(c) SAS

### Figure 8.3

(a) JMP, (b) Minitab, and (c) SAS output for the comparison of sunblock lotions in Example 8.5.

## USE YOUR KNOWLEDGE

### 8.5 Draw a picture

Draw a picture of a standard Normal curve and shade the tail areas to illustrate the calculation of the  $P$ -value for Example 8.5.

### 8.6 What does the confidence interval tell us?

Inspect the outputs in Figure 8.3. Report the confidence interval for the percent of people who would get better sun protection from your product than from your competitor's. Be sure to convert from proportions to percents and to round appropriately. Interpret the confidence interval and compare this way of analyzing data with the significance test.

### 8.7 The effect of $X$

In Example 8.5, suppose that your product provided better UVA and UVB protection for 15 of the 20 subjects. Perform the significance test and summarize the results.

### 8.8 The effect of $n$

In Example 8.5, consider what would have happened if you had paid for twice as many subjects to be tested. Assume that the results would be similar to those in Example 8.5: that is, 65% of the subjects had better UVA and UVB protection with your product. Perform the significance test and summarize the results.

In Example 8.5, we treated an outcome as a success whenever your product provided better sun protection. Would we get the same results if we defined success as an outcome where your competitor's product was superior? In this setting the null hypothesis is still  $H_0: p=0.5$ . You will find that the  $z$  test statistic is unchanged except for its sign and that the  $P$ -value remains the same.

## USE YOUR KNOWLEDGE

### 8.9 Redefining success

In Example 8.5 we performed a significance test to compare your product with your competitor's. Success was defined as the outcome where your product provided better protection. Now, take the viewpoint of your competitor where success is defined to be the outcome where your competitor's product provides better protection. In other words,  $n$  remains the same (20) but  $X$  is now 7.

- Perform the two-sided significance test and report the results. How do these compare with what we found in Example 8.5?
- Find the 95% confidence interval for this setting, and compare it with the interval calculated when success is defined as the outcome where your product provides better protection.



*We do not often use significance tests for a single proportion, because it is uncommon to have a situation where there is a precise  $p_0$  that we want to test. For physical experiments such as coin tossing or drawing cards from a well-shuffled deck, probability arguments lead to an ideal  $p_0$ . Even here, however, it can be argued, for example, that no real coin has a probability of heads *exactly* equal to 0.5. Data from past large samples can sometimes provide a  $p_0$  for the null hypothesis of a significance test. In some types of epidemiology research, for example, “historical controls” from past studies serve as the benchmark for evaluating new treatments. Medical researchers argue about the validity of these approaches, because the past never quite resembles the present. In general, we prefer comparative studies whenever possible.*

### Choosing a sample size

In Chapter 6, we showed how to choose the sample size  $n$  to obtain a confidence interval with specified margin of error  $m$  for a Normal mean. Because we are using a Normal approximation for inference about a population proportion, sample size selection proceeds in much the same way.



choosing sample size, p. 364

Recall that the margin of error for the large-sample confidence interval for a population proportion is

$$m = z^* SE_{\hat{p}} = z^* \sqrt{\hat{p}(1-\hat{p})/n}$$

Choosing a confidence level  $C$  fixes the critical value  $z^*$ . The margin of error also depends on the value of  $\hat{p}$  and the sample size  $n$ . Because we don't know the value of  $\hat{p}$  until we gather the data, we must guess a value to use in the calculations. We will call the guessed value  $p^*$ . There are two common ways to get  $p^*$ :

1. Use the sample estimate from a pilot study or from similar studies done earlier.
2. Use  $p^*=0.5$ . Because the margin of error is largest when  $\hat{p}=0.5$ , this choice gives a sample size that is somewhat larger than we really need for the confidence level we choose. It is a safe choice no matter what the data later show.

Once we have chosen  $p^*$  and the margin of error  $m$  that we want, we can find the  $n$  we need to achieve this margin of error. Here is the result.

### SAMPLE SIZE FOR DESIRED MARGIN OF ERROR

The level  $C$  confidence interval for a proportion  $p$  will have a margin of error approximately equal to a specified value  $m$  when the sample size satisfies

$$n = (z^* m)^2 / (p^* (1 - p^*))$$

Here  $z^*$  is the critical value for confidence level  $C$ , and  $p^*$  is a guessed value for the proportion of successes in the future sample.

The margin of error will be less than or equal to  $m$  if  $p^*$  is chosen to be 0.5. Substituting  $p^*=0.5$  into the formula above gives

$$n = 4(z^* m)^2$$

The value of  $n$  obtained by this method is not particularly sensitive to the choice of  $p^*$  when  $p^*$  is fairly close to 0.5. However, if the value of  $p$  is likely to be smaller than about 0.3 or larger than about 0.7, use of  $p^*=0.5$  may result in a sample size that is much larger than needed.

### Example

## 8.7 Planning a survey of students

A large university is interested in assessing student satisfaction with the overall campus environment. The plan is to distribute a questionnaire to an SRS of students, but before proceeding, the university wants to determine how many students to sample. The questionnaire asks about a student's degree of satisfaction with various student services, each measured on a five-point scale. The university is interested in the proportion  $p$  of students who are satisfied (that is, who choose either "satisfied" or "very satisfied," the two highest levels on the five-point scale).

The university wants to estimate  $p$  with 95% confidence and a margin of error less than or equal to 3%, or 0.03. For planning purposes, it is willing to use  $p^*=0.5$ . To find the sample size required,

$$n=14(z^*m)^2=14(1.960.03)^2=1067.1$$

Round up to get  $n=1068$ . (Always round up. Rounding down would give a margin of error slightly greater than 0.03.)

Similarly, for a 2.5% margin of error, we have (after rounding up)

$$n=14(1.960.025)^2=1537$$

and for a 2% margin of error,

$$n=14(1.960.02)^2=2401$$

News reports frequently describe the results of surveys with sample sizes between 1000 and 1500 and a margin of error of about 3%. These surveys generally use sampling procedures more complicated than simple random sampling, so the calculation of confidence intervals is more involved than what we have studied in this section. The calculations in Example 8.7 show in principle how such surveys are planned.

In practice, many factors influence the choice of a sample size. The following example illustrates one set of factors.

### Example

## 8.8 Assessing interest in Pilates classes



The Division of Recreational Sports (Rec Sports) at a major university is responsible for offering comprehensive recreational programs, services, and facilities to the students. Rec Sports is continually examining its programs to determine how well it is meeting the needs of the students. Rec Sports is considering adding some new programs and would like to know how much interest there is in a new exercise program based on the Pilates method.<sup>6</sup> They will take a survey of undergraduate students. In the past, they emailed short surveys to all undergraduate students. The response rate obtained in this way was about 5%. This time they will send emails to a simple random sample of the students and will follow up with additional emails and eventually a phone call to get a higher response rate. Because of limited staff and the work involved with the follow-up, they would like to use a sample size of about 200 responses. They assume that the new procedures will improve the response rate to 90%, so they will contact 225 students in the hope that these will provide at least 200 valid responses. One of the questions they will ask is “Have you ever heard about the Pilates method of exercise?”

The primary purpose of the survey is to estimate various sample proportions for undergraduate students. Will the proposed sample size of  $n=200$  be adequate to provide Rec Sports with the needed information? To address this question, we calculate the margins of error of 95% confidence intervals for various values of  $p^{\wedge}$ .

## Example

### 8.9 Margins of error

In the Rec Sports survey, the margin of error of a 95% confidence interval for any value of  $p^\wedge$  and  $n=200$  is

$$\begin{aligned} m &= z * SE_{p^\wedge} \\ &= 1.96 p^\wedge (1 - p^\wedge) / 200 \\ &= 0.139 p^\wedge (1 - p^\wedge) \end{aligned}$$

The results for various values of  $p^\wedge$  are

$p^\wedge$	$m$
0.05	0.030
0.10	0.042
0.20	0.056
0.30	0.064
0.40	0.068
0.50	0.070
0.60	0.068
0.70	0.064
0.80	0.056
0.90	0.042
0.95	0.030

Rec Sports judged these margins of error to be acceptable, and they used a sample size of 200 in their survey.

The table in Example 8.9 illustrates two points. First, the margins of error for  $p^\wedge=0.05$  and  $p^\wedge=0.95$  are the same. The margins of error will always be the same for  $p^\wedge$  and  $1-p^\wedge$ . This is a direct consequence of the form of the confidence interval. Second, the margin of error varies between only 0.064 and 0.070 as  $p^\wedge$  varies from 0.3 to 0.7, and the margin of error is greatest when  $p^\wedge=0.5$ , as we claimed earlier (page 500). It is true in general that the margin of error will vary relatively little for values of  $p^\wedge$  between 0.3 and 0.7. Therefore, when planning a study, it is not necessary to have a very precise guess for  $p$ . If  $p^\wedge=0.5$  is used and the observed  $p^\wedge$  is between 0.3 and 0.7, the actual interval will be a little shorter than needed, but the difference will be small.



Again it is important to emphasize that these calculations consider only the effects of sampling variability that are quantified in the margin of error. Other sources of error, such as nonresponse and possible misinterpretation of questions, are not included in the table of margins of error for Example 8.9. Rec Sports is trying to minimize these kinds of errors. They did a pilot study using a small group of current users of their facilities to check the wording of the questions, and for the final survey they devised a careful plan to follow up with the students who did not respond to the initial email.

## USE YOUR KNOWLEDGE

### 8.10 Confidence level and sample size

Refer to Example 8.7 (page 501). Suppose that the university was interested in a 90% confidence interval with margin of error 0.03. Would the required sample size be smaller or larger than 1068 students? Verify this by performing the calculation.

### 8.11 Make a plot

Use the values for  $\hat{p}$  and  $m$  given in Example 8.9 to draw a plot of the sample proportion versus the margin of error. Summarize the major features of your plot.

## SECTION 8.1 Summary

Inference about a population proportion  $p$  from an SRS of size  $n$  is based on the **sample proportion**  $\hat{p} = X/n$ . When  $n$  is large,  $\hat{p}$  has approximately the Normal distribution with mean  $p$  and standard deviation  $\sqrt{p(1-p)/n}$ .

For large samples, the **margin of error for confidence level  $C$**  is

$$m = z^* SE_{\hat{p}}$$

where the critical value  $z^*$  is the value for the standard Normal density curve with area  $C$  between  $-z^*$  and  $z^*$ , and the **standard error of  $\hat{p}$**  is

$$SE_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n}$$

The **level  $C$  large-sample confidence interval** is

$$\hat{p} \pm m$$

We recommend using this interval for 90%, 95%, and 99% confidence whenever

the number of successes and the number of failures are both at least 10. When sample sizes are smaller, alternative procedures such as the **plus four estimate of the population proportion** are recommended.

The **sample size** required to obtain a confidence interval of approximate margin of error  $m$  for a proportion is found from

$$n = (z^*m)^2 p^*(1-p^*)$$

where  $p^*$  is a guessed value for the proportion, and  $z^*$  is the standard Normal critical value for the desired level of confidence. To ensure that the margin of error of the interval is less than or equal to  $m$  no matter what  $p^*$  may be, use

$$n = 14(z^*m)^2$$

Tests of  $H_0: p = p_0$  are based on the  **$z$  statistic**

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}} \sqrt{n}$$

with  $P$ -values calculated from the  $N(0,1)$  distribution. Use this procedure when the expected number of successes,  $np_0$ , and the expected number of failures,  $n(1-p_0)$ , are both greater than 10.

## SECTION 8.1 Exercises

*For Exercises 8.1 and 8.2, see page 489; for Exercises 8.3 and 8.4, see page 493; for Exercises 8.5 to 8.8, see page 499; for Exercises 8.9, see page 499; and for Exercises 8.10 and 8.11, see page 503.*

### 8.12 How did you use your cell phone?

A Pew Internet poll asked cell phone owners about how they used their cell phones. One question asked whether or not during the past 30 days they had used their phone while in a store to call a friend or family member for advice about a purchase they were considering. The poll surveyed 1003 adults living in the United States by telephone. Of these, 462 responded that they had used their cell phone while in a store within the last 30 days to call a friend or family member for advice about a purchase they were considering.<sup>7</sup>

- Identify the sample size and the count.
- Calculate the sample proportion.
- Explain the relationship between the population proportion and the sample proportion.

### 8.13 Do you eat breakfast?

A random sample of 200 students from your college are asked if they regularly eat breakfast. Eighty-four students responded that they did eat breakfast regularly.

- Identify the sample size and the count.
- Calculate the sample proportion.

- (c) Explain the relationship between the population proportion and the sample proportion.

### **8.14 Would you recommend the service to a friend?**

An automobile dealership asks all its customers who used their service department in a given two-week period if they would recommend the service to a friend. A total of 230 customers used the service during the two-week period, and 180 said that they would recommend the service to a friend.

- (a) Identify the sample size and the count.  
(b) Calculate the sample proportion.  
(c) Explain the relationship between the population proportion and the sample proportion.

### **8.15 How did you use your cell phone?**

Refer to Exercise 8.12.

- (a) Report the sample proportion, the standard error of the sample proportion, and the margin of error for 95% confidence.  
(b) Are the guidelines for when to use the large-sample confidence interval for a population proportion satisfied in this setting? Explain your answer.  
(c) Find the 95% large-sample confidence interval for the population proportion.  
(d) Write a short statement explaining the meaning of your confidence interval.

### **8.16 Do you eat breakfast?**

Refer to Exercise 8.13.

- (a) Report the sample proportion, the standard error of the sample proportion, and the margin of error for 95% confidence.  
(b) Are the guidelines for when to use the large-sample confidence interval for a population proportion satisfied in this setting? Explain your answer.  
(c) Find the 95% large-sample confidence interval for the population proportion.  
(d) Write a short statement explaining the meaning of your confidence interval.

### **8.17 Would you recommend the service to a friend?**

Refer to Exercise 8.14.

- (a) Report the sample proportion, the standard error of the sample proportion, and the margin of error for 95% confidence.  
(b) Are the guidelines for when to use the large-sample confidence interval for a population proportion satisfied in this setting? Explain your answer.  
(c) Find the 95% large-sample confidence interval for the population proportion.  
(d) Write a short statement explaining the meaning of your confidence interval.

## 8.18 Whole grain versus regular grain?

A study of young children was designed to increase their intake of whole-grain, rather than regular-grain, snacks. At the end of the study the 76 children who participated in the study were presented with a choice between a regular-grain snack and a whole-grain alternative. The whole-grain alternative was chosen by 52 children. You want to examine the possibility that the children are equally likely to choose each type of snack.

- Formulate the null and alternative hypotheses for this setting.
- Are the guidelines for using the large-sample significance test satisfied for testing this null hypothesis? Explain your answer.
- Perform the significance test and summarize your results in a short paragraph.

## 8.19 Find the sample size.

You are planning a survey similar to the one about cell phone use described in Exercise 8.12. You will report your results with a large-sample confidence interval. How large a sample do you need to be sure that the margin of error will not be greater than 0.04? Show your work.

## 8.20 What's wrong?

Explain what is wrong with each of the following:

- An approximate 90% confidence interval for an unknown proportion  $p$  is  $\hat{p}$  plus or minus its standard error.
- You can use a significance test to evaluate the hypothesis  $H_0: p = 0.3$  versus the one-sided alternative.
- The large-sample significance test for a population proportion is based on a  $t$  statistic.

## 8.21 What's wrong?

Explain what is wrong with each of the following:

- A student project used a confidence interval to describe the results in a final report. The confidence level was 115%.
- The margin of error for a confidence interval used for an opinion poll takes into account the fact that people who did not answer the poll questions may have had different responses from those who did answer the questions.
- If the  $P$ -value for a significance test is 0.50, we can conclude that the null hypothesis has a 50% chance of being true.

## 8.22 Draw some pictures.

Consider the binomial setting with  $n=100$  and  $p=0.4$ .

- The sample proportion  $\hat{p}$  will have a distribution that is approximately Normal. Give the mean and the standard deviation of this Normal distribution.

(b) Draw a sketch of this Normal distribution. Mark the location of the mean.

(c) Find a value of  $x$  for which the probability is 95% that  $\hat{p}$  is within  $x$  of 0.4. Mark the corresponding interval on your plot.

### 8.23 Country food and Inuits.

Country food includes seals, caribou, whales, ducks, fish, and berries and is an important part of the diet of the aboriginal people called Inuits who inhabit Inuit Nunangat, the northern region of what is now called Canada. A survey of Inuits in Inuit Nunangat reported that 3274 out of 5000 respondents said that at least half of the meat and fish that they eat is country food.<sup>8</sup> Find the sample proportion and a 95% confidence interval for the population proportion of Inuits whose meat and fish consumption consists of at least half country food.

### 8.24 Soft drink consumption in New Zealand.

A survey commissioned by the Southern Cross Healthcare Group reported that 16% of New Zealanders consume five or more servings of soft drinks per week. The data were obtained by an online survey of 2006 randomly selected New Zealanders over 15 years of age.<sup>9</sup>

(a) What number of survey respondents reported that they consume five or more servings of soft drinks per week? You will need to round your answer. Why?

(b) Find a 95% confidence interval for the proportion of New Zealanders who report that they consume five or more servings of soft drinks per week.

(c) Convert the estimate and your confidence interval to percents.

(d) Discuss reasons why the estimate might be biased.

### 8.25 Violent video games.

A 2013 survey of 1050 parents who have a child under the age of 18 living at home asked about their opinions regarding violent video games. A report describing the results of the survey stated that 89% of parents say that violence in today's video games is a problem.<sup>10</sup>

(a) What number of survey respondents reported that they thought that violence in today's video games is a problem? You will need to round your answer. Why?

(b) Find a 95% confidence interval for the proportion of parents who think that violence in today's video games is a problem.

(c) Convert the estimate and your confidence interval to percents.

(d) Discuss reasons why the estimate might be biased.

### 8.26 Bullying.

Refer to the previous exercise. The survey also reported that 93% of the parents surveyed said that bullying contributes to violence in the United States. Answer the questions in the previous exercise for this item on the survey.

### 8.27 $p^{\wedge}$ and the Normal distribution.

Consider the binomial setting with  $n=50$ . You are testing the null hypothesis that  $p=0.3$  versus the two-sided alternative with a 5% chance of rejecting the null hypothesis when it is true.

- (a) Find the values of the sample proportion  $p^{\wedge}$  that will lead to rejection of the null hypothesis.
- (b) Repeat part (a) assuming a sample size of  $n=100$ .
- (c) Make a sketch illustrating what you have found in parts (a) and (b). What does your sketch show about the effect of the sample size in this setting?

### 8.28 Students doing community service.

In a sample of 159, 949 first-year college students, the National Survey of Student Engagement reported that 39% participated in community service or volunteer work.<sup>11</sup>

- (a) Find the margin of error for 99% confidence.
- (b) Here are some facts from the report that summarizes the survey. The students were from 617 four-year colleges and universities. The response rate was 36%. Institutions paid a participation fee of between \$1800 and \$7800 based on the size of their undergraduate enrollment. Discuss these facts as possible sources of error in this study. How do you think these errors would compare with the margin of error that you calculated in part (a)?

### 8.29 Plans to study abroad.

The survey described in the previous exercise also asked about items related to academics. In response to one of these questions, 42% of first-year students reported that they plan to study abroad.

- (a) Based on the information available, how many students plan to study abroad?
- (b) Give a 99% confidence interval for the population proportion of first-year college students who plan to study abroad.

### 8.30 Student credit cards.

In a survey of 1430 undergraduate students, 1087 reported that they had one or more credit cards.<sup>12</sup> Give a 95% confidence interval for the proportion of all college students who have at least one credit card.

### 8.31 How many credit cards?

The summary of the survey described in the previous exercise reported that 43% of undergraduates had four or more credit cards. Give a 95% confidence interval for the proportion of all college students who have four or more credit cards.

### 8.32 How would the confidence interval change?

Refer to Exercise 8.31.

- (a) Would a 99% confidence interval be wider or narrower than the one that you found in Exercise 8.31?

Verify your results by computing the interval.

(b) Would a 90% confidence interval be wider or narrower than the one that you found in that exercise? Verify your results by computing the interval.

### 8.33 Do students report Internet sources?

The National Survey of Student Engagement found that 87% of students report that their peers at least “sometimes” copy information from the Internet in their papers without reporting the source.<sup>13</sup> Assume that the sample size is 430,000.

(a) Find the margin of error for 99% confidence.

(b) Here are some items from the report that summarizes the survey. More than 430,000 students from 730 four-year colleges and universities participated. The average response rate was 43% and ranged from 15% to 89%. Institutions pay a participation fee of between \$3000 and \$7500 based on the size of their undergraduate enrollment. Discuss these facts as possible sources of error in this study. How do you think these errors would compare with the error that you calculated in part (a)?

### 8.34 Can we use the $z$ test?

In each of the following cases state whether or not the Normal approximation to the binomial should be used for a significance test on the population proportion  $p$ . Explain your answers.

(a)  $n=40$  and  $H_0:p=0.2$ .

(b)  $n=30$  and  $H_0:p=0.4$ .

(c)  $n=100$  and  $H_0:p=0.15$ .

(d)  $n=200$  and  $H_0:p=0.04$ .



### 8.35 Long sermons

The National Congregations Study collected data in a one-hour interview with a key informant—that is, a minister, priest, rabbi, or other staff person or leader.<sup>14</sup> One question concerned the length of the typical sermon. For this question 390 out of 1191 congregations reported that the typical sermon lasted more than 30 minutes.

(a) Use the large-sample inference procedures to estimate the true proportion for this question with a 95% confidence interval.

(b) The respondents to this question were not asked to use a stopwatch to record the lengths of a random sample of sermons at their congregations. They responded based on their impressions of the sermons. Do you think that ministers, priests, rabbis, or other staff persons or leaders might perceive sermon lengths differently from the people listening to the sermons? Discuss how your ideas would influence your interpretation of the results of this study.

### 8.36 Confidence level and interval width.

Refer to the previous exercise. Would a 99% confidence interval be wider or narrower than the one that you found in that exercise? Verify your results by computing the interval.

### 8.37 Instant versus fresh-brewed coffee.

A matched pairs experiment compares the taste of instant and fresh-brewed coffee. Each subject tastes two unmarked cups of coffee, one of each type, in random order and states which he or she prefers. Of the 50 subjects who participate in the study, 15 prefer the instant coffee. Let  $p$  be the probability that a randomly chosen subject prefers fresh-brewed coffee to instant coffee. (In practical terms,  $p$  is the proportion of the population who prefer fresh-brewed coffee.)

- (a) Test the claim that a majority of people prefer the taste of fresh-brewed coffee. Report the large-sample  $z$  statistic and its  $P$ -value.
- (b) Draw a sketch of a standard Normal curve and mark the location of your  $z$  statistic. Shade the appropriate area that corresponds to the  $P$ -value.
- (c) Is your result significant at the 5% level? What is your practical conclusion?

### 8.38 Annual income of older adults.

In a study of older adults, 1444 subjects out of a total of 2733 reported that their annual income was \$30,000 or more.

- (a) Give a 95% confidence interval for the true proportion of subjects in this population with incomes of at least \$30,000.
- (b) Do you think that some respondents might not give truthful answers to a question about their income? Discuss the possible effects on your estimate and confidence interval.

### 8.39 Tossing a coin 10,000 times!

The South African mathematician John Kerrich, while a prisoner of war during World War II, tossed a coin 10,000 times and obtained 5067 heads.

- (a) Is this significant evidence at the 5% level that the probability that Kerrich's coin comes up heads is not 0.5? Use a sketch of the standard Normal distribution to illustrate the  $P$ -value.
- (b) Use a 95% confidence interval to find the range of probabilities of heads that would not be rejected at the 5% level.

### 8.40 Is there interest in a new product?

One of your employees has suggested that your company develop a new product. You decide to take a random sample of your customers and ask whether or not there is interest in the new product. The response is on a 1 to 5 scale with 1 indicating "definitely would not purchase"; 2, "probably would not purchase"; 3, "not sure"; 4, "probably would purchase"; and 5, "definitely would purchase." For an initial analysis, you will record the responses 1, 2, and 3 as "No" and 4 and 5 as "Yes." What sample size would you use if you wanted the 95% margin of error to be 0.2 or less?

### 8.41 More information is needed.

Refer to the previous exercise. Suppose that after reviewing the results of the previous survey, you proceeded with preliminary development of the product. Now you are at the stage where you need to decide whether or not to make a major investment to produce and market it. You will use another random sample of your customers, but now you want the margin of error to be smaller. What sample size would you use if you wanted the 95% margin of error to be 0.01 or less?

### 8.42 Sample size needed for an evaluation.

You are planning an evaluation of a semester-long alcohol awareness campaign at your college. Previous evaluations indicate that about 20% of the students surveyed will respond “Yes” to the question “Did the campaign alter your behavior toward alcohol consumption?” How large a sample of students should you take if you want the margin of error for 95% confidence to be about 0.08?



### 8.43 Sample size needed for an evaluation, continued

The evaluation in the previous exercise will also have questions that have not been asked before, so you do not have previous information about the possible value of  $p$ . Repeat the preceding calculation for the following values of  $p^*$ : 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. Summarize the results in a table and graphically. What sample size will you use?

### 8.44 Are the customers dissatisfied?

An automobile manufacturer would like to know what proportion of its customers are dissatisfied with the service received from their local dealer. The customer relations department will survey a random sample of customers and compute a 95% confidence interval for the proportion who are dissatisfied. From past studies, they believe that this proportion will be about 0.2. Find the sample size needed if the margin of error of the confidence interval is to be no more than 0.02.

## 8.2 Comparing Two Proportions

When you complete this section, you will be able to

- Identify the counts and sample sizes for a comparison between two proportions; compute the proportions and find their difference.
- Apply the guidelines for when to use the large-sample confidence interval for a difference between two proportions.
- Apply the large-sample method to find the confidence interval for a difference between two proportions and interpret the confidence interval.
- Apply the guidelines for when to use the large-sample significance test for a difference between two proportions.
- Apply the large-sample method to perform a significance test for comparing two proportions and interpret the results of the significance test.
- Calculate and interpret the relative risk.

Because comparative studies are so common, we often want to compare the proportions of two groups (such as men and women) that have some characteristic. In the previous section we learned how to estimate a single proportion. Our problem now concerns the comparison of two proportions.

We call the two groups being compared Population 1 and Population 2, and the two population proportions of “successes”  $p_1$  and  $p_2$ . The data consist of two independent SRSs, of size  $n_1$  from Population 1 and size  $n_2$  from Population 2. The proportion of successes in each sample estimates the corresponding population proportion. Here is the notation we will use in this section:

Population	Population proportion	Sample size	Count of successes	Sample proportion
1	$p_1$	$n_1$	$X_1$	$\hat{p}_1 = X_1/n_1$
2	$p_2$	$n_2$	$X_2$	$\hat{p}_2 = X_2/n_2$

To compare the two populations, we use the difference between the two sample proportions:

$$D = \hat{p}_1 - \hat{p}_2$$

When both sample sizes are sufficiently large, the sampling distribution of the difference  $D$  is approximately Normal.

Inference procedures for comparing proportions are  $z$  procedures based on the Normal approximation and on standardizing the difference  $D$ . The first step is to

obtain the mean and standard deviation of  $D$ . By the addition rule for means, the mean of  $D$  is the difference of the means:

$$\mu_D = \mu_{p^1} - \mu_{p^2} = p_1 - p_2$$

 **LOOK BACK**

**addition rule for means, p. 272**

That is, the difference  $D = p^1 - p^2$  between the sample proportions is an unbiased estimator of the population difference  $p_1 - p_2$ . Similarly, the addition rule for variances tells us that the variance of  $D$  is the *sum* of the variances:

$$\begin{aligned}\sigma_D^2 &= \sigma_{p^1}^2 + \sigma_{p^2}^2 \\ &= p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2\end{aligned}$$

 **LOOK BACK**

**addition rule for variances, p. 275**

Therefore, when  $n_1$  and  $n_2$  are large,  $D$  is approximately Normal with mean  $\mu_D = p_1 - p_2$  and standard deviation

$$\sigma_D = \sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$$

## USE YOUR KNOWLEDGE

### 8.45 Rules for means and variances

Suppose that  $p_1 = 0.3$ ,  $n_1 = 20$ ,  $p_2 = 0.6$ ,  $n_2 = 30$ . Find the mean and the standard deviation of the sampling distribution of  $p_1 - p_2$ .

### 8.46 Effect of the sample sizes

Suppose that  $p_1 = 0.3$ ,  $n_1 = 80$ ,  $p_2 = 0.6$ ,  $n_2 = 120$ .

- Find the mean and the standard deviation of the sampling distribution of  $p_1 - p_2$ .
- The sample sizes here are four times as large as those in the previous exercise while the population proportions are the same. Compare the results for this exercise with those that you found in the previous exercise. What is the effect of multiplying the sample sizes by 4?

### 8.47 Rules for means and variances

It is quite easy to verify the formulas for the mean and standard

deviation of the difference  $D$ .

- (a) What are the means and standard deviations of the two sample proportions  $p^1$  and  $p^2$ ?
- (b) Use the addition rule for means of random variables: what is the mean of  $D=p^1-p^2$ ?
- (c) The two samples are independent. Use the addition rule for variances of random variables: what is the variance of  $D$ ?

## Large-sample confidence interval for a difference in proportions

To obtain a confidence interval for  $p_1-p_2$ , we once again replace the unknown parameters in the standard deviation by estimates to obtain an estimated standard deviation, or standard error. Here is the confidence interval we want.

### LARGE-SAMPLE CONFIDENCE INTERVAL FOR COMPARING TWO PROPORTIONS

Choose an SRS of size  $n_1$  from a large population having proportion  $p_1$  of successes and an independent SRS of size  $n_2$  from another population having proportion  $p_2$  of successes. The estimate of the difference in the population proportions is

$$D=p^1-p^2$$

The **standard error of  $D$**  is

$$SED=p^1(1-p^1)n_1+p^2(1-p^2)n_2$$

and the **margin of error** for confidence level  $C$  is

$$m=z^*SED$$

where the critical value  $z^*$  is the value for the standard Normal density curve with area  $C$  between  $-z^*$  and  $z^*$ . An **approximate level  $C$  confidence interval** for  $p_1-p_2$  is

$$D\pm m$$

Use this method for 90%, 95%, or 99% confidence when the number of successes and the number of failures in each sample are both at least 10.

### Example

## 8.10 Are you spending more time on Facebook?



### FACEBOOK

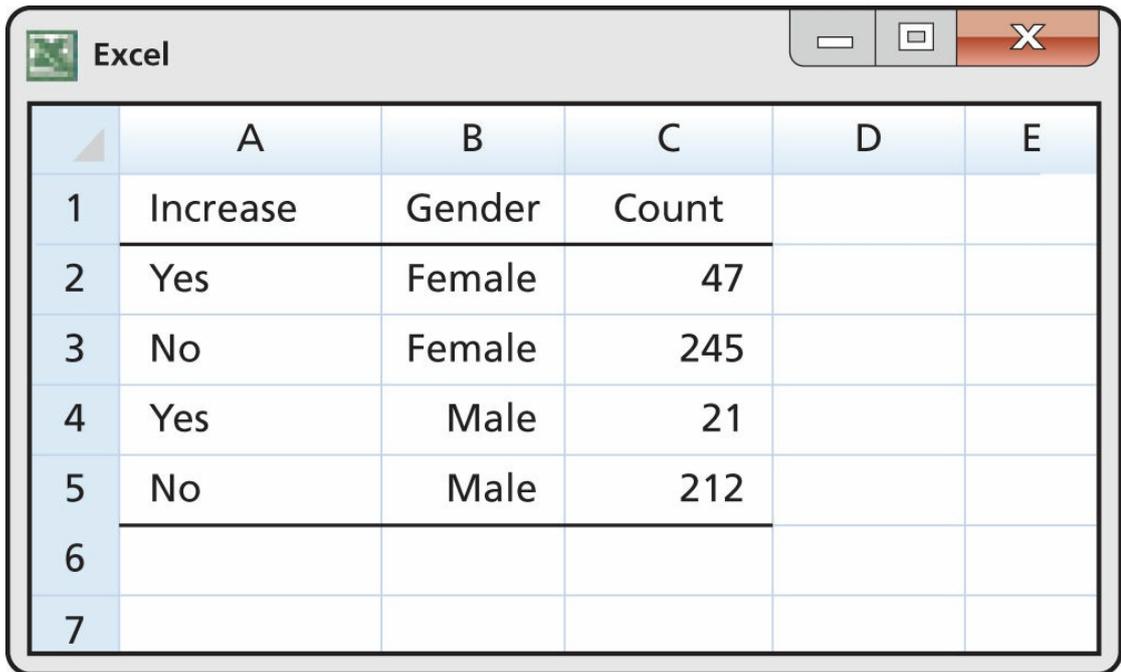
A Pew Internet survey asked 525 Facebook users about changes in the amount of time spent using Facebook over the past year. Here are the data for the response variable, Increase, with values “Yes” and “No,” classified by the explanatory variable, Gender, with values “Men” and “Women.” The cases are the 525 Facebook users who participated in the survey.<sup>15</sup> Here are the data:

Population	<i>n</i>	<i>X</i>	$p^{\wedge}=X/n$
1 (women)	292	47	0.1610
2 (men)	233	21	0.0901
Total	525	68	0.1295

In this table the  $p^{\wedge}$  column gives the sample proportions of Facebook users who increased their use of Facebook over the past year.

Let’s find a 95% confidence interval for the difference between the proportions of women and of men who increased their time spent on Facebook over the past year. Figure 8.4 shows a spreadsheet that can be used as input to software that can compute the confidence interval. Output from JMP, Minitab, and SAS is given in Figure 8.5. To perform the computations using our formulas, we first find the difference in the proportions:

$$\begin{aligned} D &= p^{\wedge}1 - p^{\wedge}2 \\ &= 0.1610 - 0.0901 \\ &= 0.0709 \end{aligned}$$

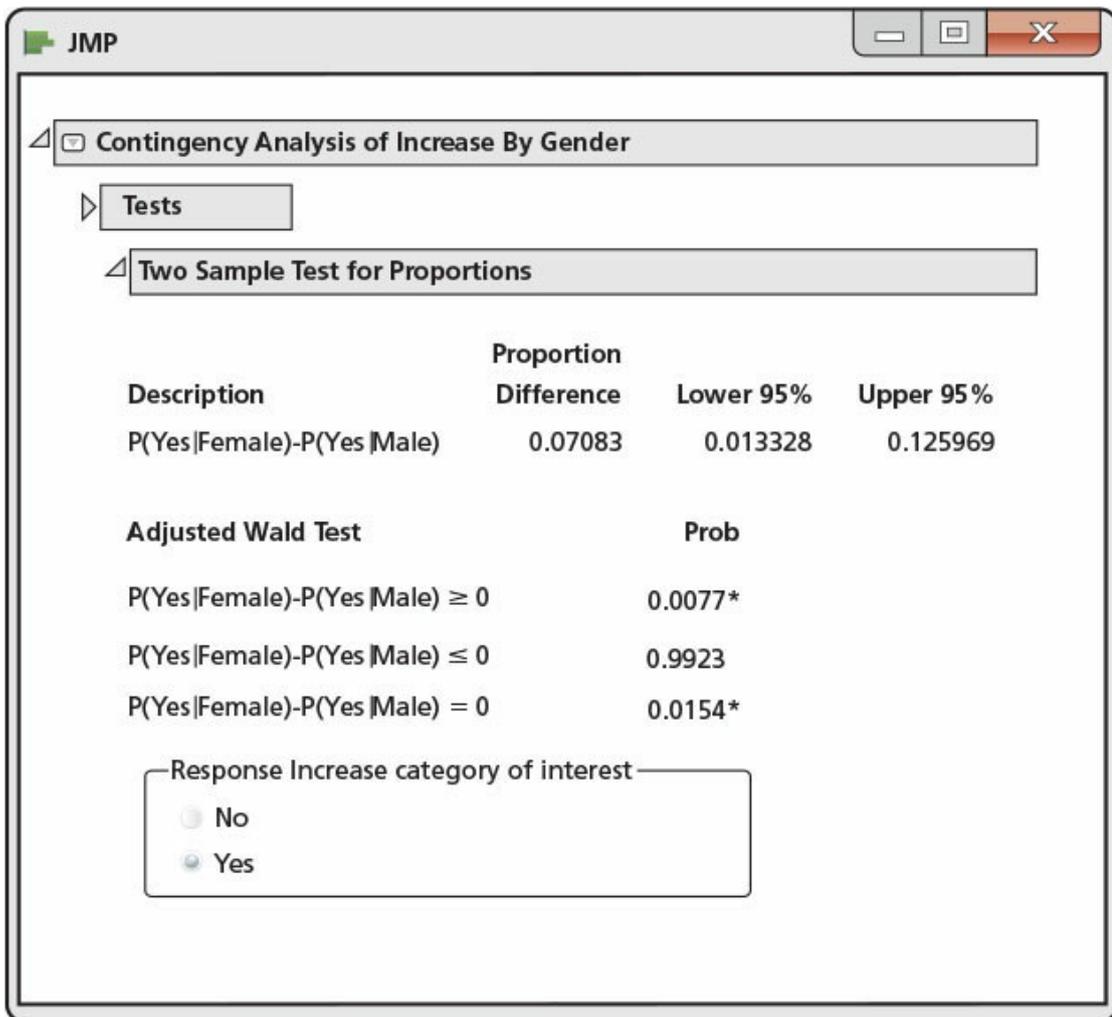


The image shows a screenshot of an Excel spreadsheet window. The window title is "Excel" and it has standard minimize, maximize, and close buttons. The spreadsheet contains a table with 7 rows and 6 columns (A-F). The first row (row 1) contains headers: "Increase" in column A, "Gender" in column B, and "Count" in column C. The following rows contain data:

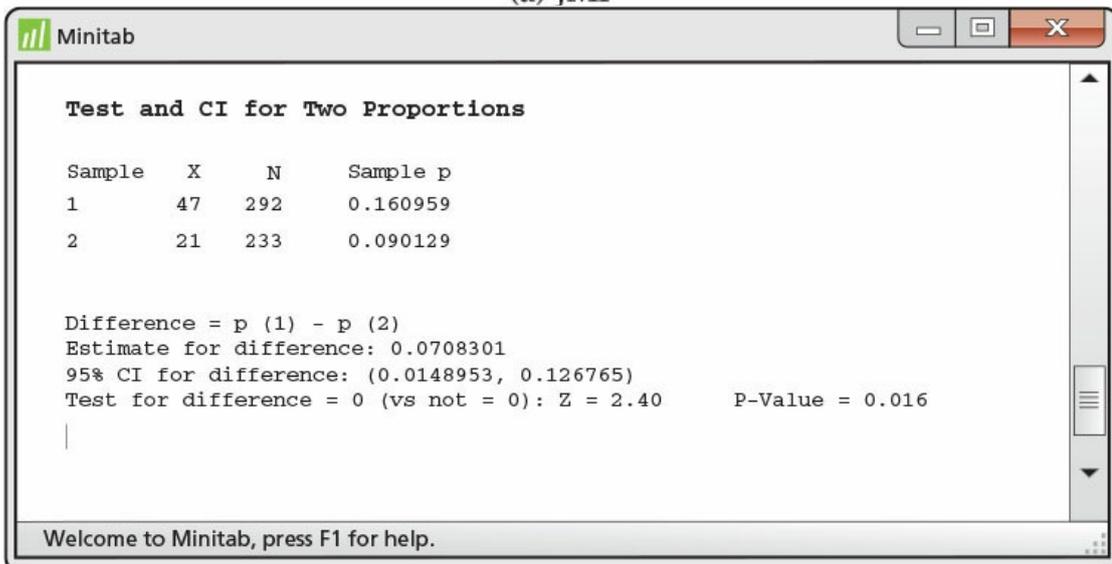
	A	B	C	D	E
1	Increase	Gender	Count		
2	Yes	Female	47		
3	No	Female	245		
4	Yes	Male	21		
5	No	Male	212		
6					
7					

**Figure 8.4**

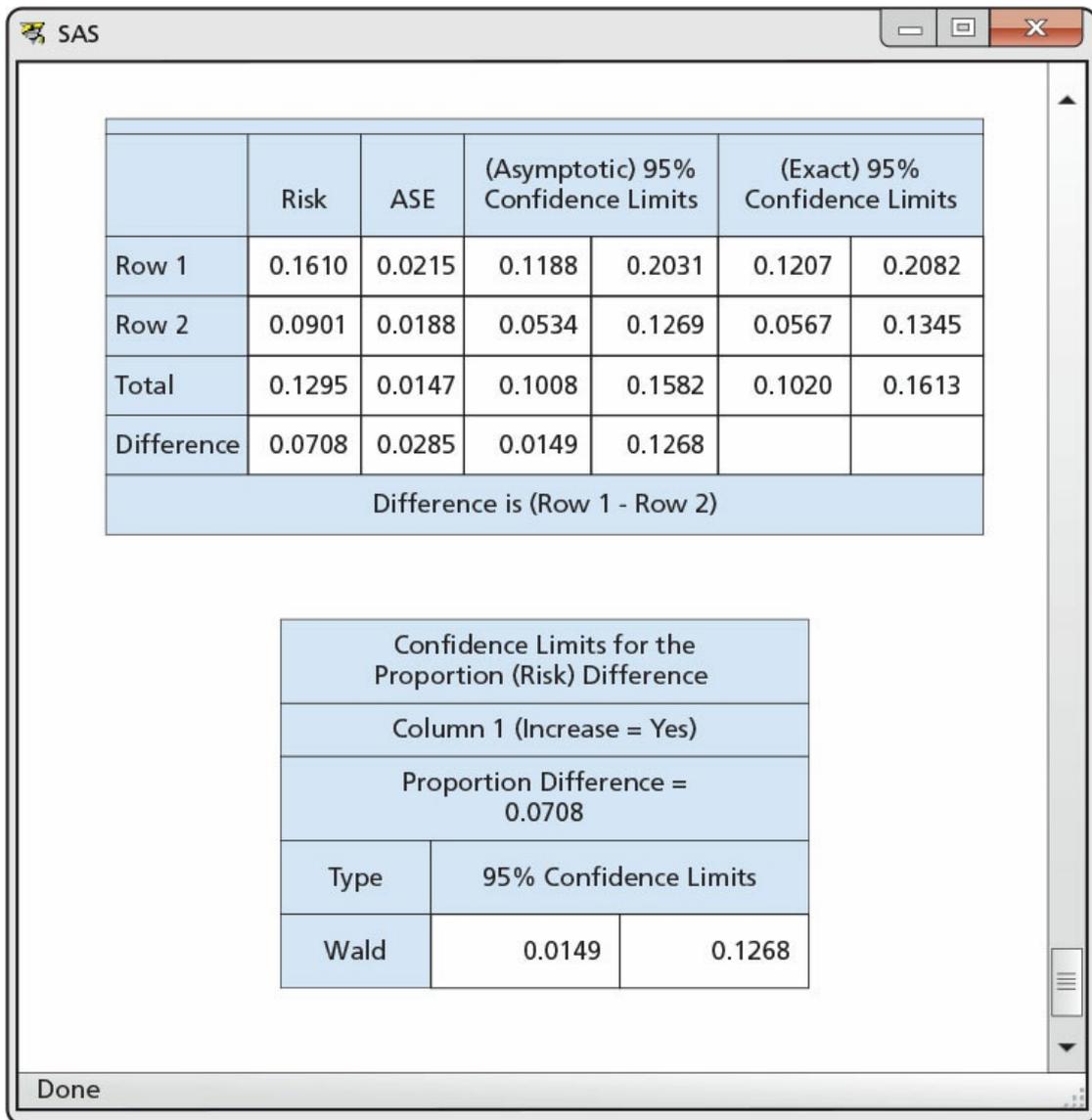
Spreadsheet that can be used as input to software that computes the confidence interval for the Facebook data in Example 8.10.



(a) JMP



(b) Minitab



(c) SAS

### Figure 8.5

(a) JMP, (b) Minitab, and (c) SAS output for the Facebook time confidence interval in Example 8.10.

Then we calculate the standard error of  $D$ :

$$\begin{aligned}
 SED &= p^1(1-p^1)n1 + p^2(1-p^2)n2 \\
 &= (0.1610)(0.8390)292 + (0.0901)(0.9099)233 \\
 &= 0.0285
 \end{aligned}$$

For 95% confidence, we have  $z^*=1.96$ , so the margin of error is

$$\begin{aligned}
 m &= z^*SED \\
 &= (1.96)(0.0285) \\
 &= 0.0559
 \end{aligned}$$

The 95% confidence interval is

$$\begin{aligned} D \pm m &= 0.0709 \pm 0.0559 \\ &= (0.0150, 0.1268) \end{aligned}$$

With 95% confidence we can say that the difference in the proportions is between 0.0150 and 0.1268. Alternatively, we can report that the difference between the percent of women who increased their time spent on Facebook over the past year and the percent of men who did so is 7.1%, with a 95% margin of error of 5.6%.

In this example men and women were not sampled separately. The sample sizes are, in fact, random and reflect the gender distributions of the subjects who responded to the survey. Two-sample significance tests and confidence intervals are still approximately correct in this situation.

In the example above we chose women to be the first population. Had we chosen men to be the first population, the estimate of the difference would be negative ( $-0.0709$ ). Because it is easier to discuss positive numbers, we generally choose the first population to be the one with the higher proportion.

## USE YOUR KNOWLEDGE

### 8.48 Gender and commercial preference

A study was designed to compare two energy drink commercials. Each participant was shown the commercials in random order and asked to select the better one. Commercial A was selected by 44 out of 100 women and 79 out of 140 men. Give an estimate of the difference in gender proportions that favored Commercial A. Also construct a large-sample 95% confidence interval for this difference.

### 8.49 Gender and commercial preference, revisited

Refer to Exercise 8.48. Construct a 95% confidence interval for the difference in proportions that favor Commercial B. Explain how you could have obtained these results from the calculations you did in Exercise 8.48.

## BEYOND THE BASICS

### The plus four confidence interval for a difference in proportions

Just as in the case of estimating a single proportion, a small modification of the sample proportions can greatly improve the accuracy of confidence intervals.<sup>16</sup> As before, we add 2 successes and 2 failures to the actual data, but now we divide them equally between the two samples. That is, we *add 1 success and 1 failure to each sample*. We will again call the estimates produced by adding hypothetical observations plus four estimates. The plus four estimates of the two population proportions are

$$\tilde{p}_1 = \frac{X_1 + 1}{n_1 + 2} \text{ and } \tilde{p}_2 = \frac{X_2 + 1}{n_2 + 2}$$

The estimated difference between the populations is

$$\tilde{D} = \tilde{p}_1 - \tilde{p}_2$$

and the standard deviation of  $\tilde{D}$  is approximately

$$\sigma_{\tilde{D}} = \sqrt{\tilde{p}_1(1-\tilde{p}_1)/(n_1+2) + \tilde{p}_2(1-\tilde{p}_2)/(n_2+2)}$$

This is similar to the formula for  $\sigma_D$ , adjusted for the sizes of the modified samples.

To obtain a confidence interval for  $p_1 - p_2$ , we once again replace the unknown parameters in the standard deviation by estimates to obtain an estimated standard deviation, or standard error. Here is the confidence interval we want.

## PLUS FOUR CONFIDENCE INTERVAL FOR COMPARING TWO PROPORTIONS

Choose an SRS of size  $n_1$  from a large population having proportion  $p_1$  of successes and an independent SRS of size  $n_2$  from another population having proportion  $p_2$  of successes. The **plus four estimate of the difference in proportions** is

$$\tilde{D} = \tilde{p}_1 - \tilde{p}_2$$

where

$$\tilde{p}_1 = \frac{X_1 + 1}{n_1 + 2} \quad \tilde{p}_2 = \frac{X_2 + 1}{n_2 + 2}$$

The **standard error of  $\tilde{D}$**  is

$$SED_{\tilde{D}} = \sqrt{\tilde{p}_1(1-\tilde{p}_1)/(n_1+2) + \tilde{p}_2(1-\tilde{p}_2)/(n_2+2)}$$

and the **margin of error** for confidence level  $C$  is

$$m = z^* \text{SED}^{\sim}$$

where  $z^*$  is the value for the standard Normal density curve with area  $C$  between  $-z^*$  and  $z^*$ . An **approximate level  $C$  confidence interval** for  $p_1 - p_2$  is

$$\tilde{D} \pm m$$

Use this method for 90%, 95%, or 99% confidence when both sample sizes are at least 5.

## Example

### 8.11 Gender and sexual maturity

In studies that look for a difference between genders, a major concern is whether or not apparent differences are due to other variables that are associated with gender. Because boys mature more slowly than girls, a study of adolescents that compares boys and girls of the same age may confuse a gender effect with an effect of sexual maturity. The “Tanner score” is a commonly used measure of sexual maturity.<sup>17</sup> Subjects are asked to determine their score by placing a mark next to a rough drawing of an individual at their level of sexual maturity. There are five different drawings, so the score is an integer between 1 and 5.

A pilot study included 12 girls and 12 boys from a population that will be used for a large experiment. Four of the boys and three of the girls had Tanner scores of 4 or 5, a high level of sexual maturity. Let’s find a 95% confidence interval for the difference between the proportions of boys and girls who have high (4 or 5) Tanner scores in this population. The numbers of successes and failures in both groups are not all at least 10, so the large-sample approach is not recommended. On the other hand, the sample sizes are both at least 5, so the plus four method is appropriate.

The plus four estimate of the population proportion for boys is

$$\tilde{p}_1 = \frac{X_1 + 1}{n_1 + 2} = \frac{4 + 1}{12 + 2} = 0.3571$$

For girls, the estimate is

$$\tilde{p}_2 = \frac{X_2 + 1}{n_2 + 2} = \frac{3 + 1}{12 + 2} = 0.2857$$

Therefore, the estimate of the difference is

$$\tilde{D} = \tilde{p}_1 - \tilde{p}_2 = 0.3571 - 0.2857 = 0.071$$

The standard error of  $\tilde{D}$  is

$$\begin{aligned} \text{SED}_{\tilde{D}} &= \sqrt{\tilde{p}_1(1-\tilde{p}_1)/n_1 + \tilde{p}_2(1-\tilde{p}_2)/n_2} \\ &= \sqrt{(0.3571)(1-0.3571)/12 + (0.2857)(1-0.2857)/12} \\ &= 0.1760 \end{aligned}$$

For 95% confidence,  $z^* = 1.96$  and the margin of error is

$$m = z^* \text{SED}_{\tilde{D}} = (1.96)(0.1760) = 0.345$$

The confidence interval is

$$\begin{aligned} \tilde{D} \pm m &= 0.071 \pm 0.345 \\ &= (-0.274, 0.416) \end{aligned}$$

With 95% confidence we can say that the difference in the proportions is between  $-0.274$  and  $0.416$ . Alternatively, we can report that the difference in the proportions of boys and girls with high Tanner scores in this population is 7.1% with a 95% margin of error of 34.5%.

The very large margin of error in this example indicates that either boys or girls could be more sexually mature in this population and that the difference could be quite large. *Although the interval includes the possibility that there is no difference, corresponding to  $p_1 = p_2$  or  $p_1 - p_2 = 0$ , we should not conclude that there is **no** difference in the proportions.* With small sample sizes such as these, the data do not provide us with a lot of information for our inference. This fact is expressed quantitatively through the very large margin of error.



## Significance test for a difference in proportions

Although we prefer to compare two proportions by giving a confidence interval for the difference between the two population proportions, it is sometimes useful to test the null hypothesis that the two population proportions are the same.

We standardize  $D = p_1 - p_2$  by subtracting its mean  $p_1 - p_2$  and then dividing by its standard deviation

$$\sigma_D = \sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$$

If  $n_1$  and  $n_2$  are large, the standardized difference is approximately  $N(0,1)$ . For the large-sample confidence interval we used sample estimates in place of the unknown population values in the expression for  $\sigma_D$ . Although this approach would lead to a valid significance test, we instead adopt the more common practice

of replacing the unknown  $\sigma_D$  with an estimate that takes into account our null hypothesis  $H_0: p_1 = p_2$ . If these two proportions are equal, then we can view all the data as coming from a single population. Let  $p$  denote the common value of  $p_1$  and  $p_2$ ; then the standard deviation of  $D = \hat{p}_1 - \hat{p}_2$  is

$$\begin{aligned}\sigma_D &= p(1-p)n_1 + p(1-p)n_2 \\ &= p(1-p)(n_1 + n_2)\end{aligned}$$

We estimate the common value of  $p$  by the overall proportion of successes in the two samples:

$$\hat{p} = \frac{\text{number of successes in both samples}}{\text{number of observations in both samples}} = \frac{X_1 + X_2}{n_1 + n_2}$$

This estimate of  $p$  is called the **pooled estimate** because it combines, or pools, the information from both samples.

### pooled estimate of $p$

To estimate  $\sigma_D$  under the null hypothesis, we substitute  $\hat{p}$  for  $p$  in the expression for  $\sigma_D$ . The result is a standard error for  $D$  that assumes  $H_0: p_1 = p_2$ :

$$SE_D = \hat{p}(1-\hat{p})(n_1 + n_2)$$

The subscript on  $SE_D$  reminds us that we pooled data from the two samples to construct the estimate.

## SIGNIFICANCE TEST FOR COMPARING TWO PROPORTIONS

To test the hypothesis

$$H_0: p_1 = p_2$$

compute the **z statistic**

$$z = \frac{\hat{p}_1 - \hat{p}_2}{SE_D}$$

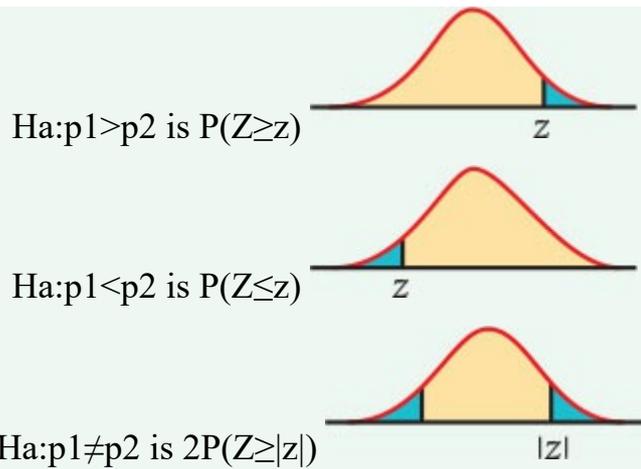
where the **pooled standard error** is

$$SE_D = \hat{p}(1-\hat{p})(n_1 + n_2)$$

and where the **pooled estimate** of the common value of  $p_1$  and  $p_2$  is

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

In terms of a standard Normal random variable  $Z$ , the approximate  $P$ -value for a test of  $H_0$  against



This  $z$  test is based on the Normal approximation to the binomial distribution. As a general rule, we will use it when the number of successes and the number of failures in each of the samples are at least 5.

## Example

### 8.12 Gender and Facebook time: the $z$ test



#### FACEBOOK TIME

Are men and women equally likely to say that they increased the amount of time that they spend on Facebook over the past year? We examine the data in Example 8.10 (page 510) to answer this question. Here is the data summary:

Population	$n$	$X$	$\hat{p} = X/n$
1 (women)	292	47	0.1610
2 (men)	233	21	0.0901
Total	525	68	0.1295

The sample proportions are certainly quite different, but we will perform a significance test to see if the difference is large enough to lead us to believe that the population proportions are not equal. Formally, we test the hypotheses

$$H_0: p_1 = p_2$$

$$H_a: p_1 \neq p_2$$

The pooled estimate of the common value of  $p$  is

$$\hat{p} = \frac{47 + 21292 + 233}{68525} = 0.1295$$

Note that this is the estimate on the bottom line of the preceding data summary.

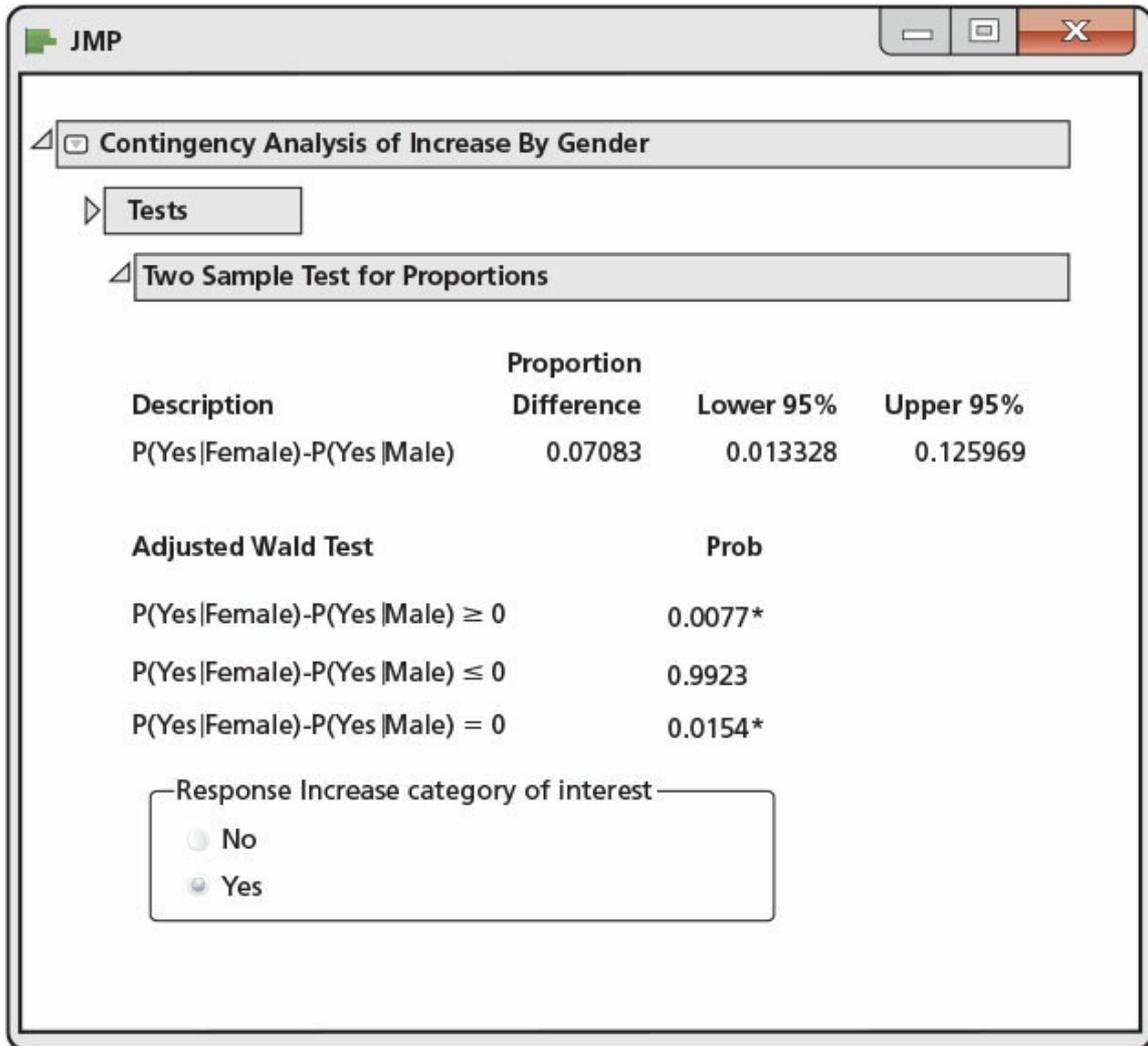
The test statistic is calculated as follows:

$$SE_{\hat{p}} = (0.1295)(0.8705) \sqrt{\frac{1}{1292} + \frac{1}{1233}} = 0.02949$$

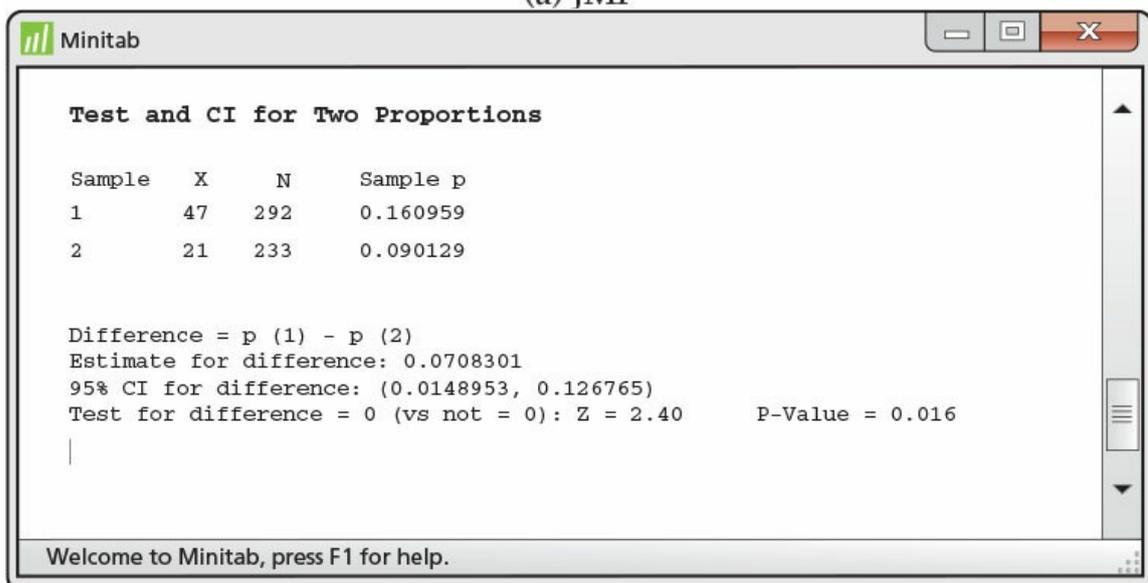
$$z = \frac{\hat{p}_1 - \hat{p}_2}{SE_{\hat{p}}} = \frac{0.1610 - 0.0901}{0.02949} = 2.40$$

The  $P$ -value is  $2P(Z \geq 2.40)$ . We can conclude that  $P < 2(1 - 0.9918) = 0.0164$ . Output from JMP, Minitab, and SAS is given in Figure 8.6. JMP reports the  $P$ -value as 0.0154, Minitab reports 0.016, and SAS reports 0.0163. Here is our summary: among the Facebook users in the study, 16.1% of the women and 9.0% of the men said that they increased the time they spent on Facebook last year; the difference is statistically significant ( $z = 2.40$ ,  $P < 0.02$ ).

Do you think that we could have argued that the proportion would be higher for women than for men before looking at the data in this example? This would allow us to use the one-sided alternative  $H_a: p_1 > p_2$ . The  $P$ -value would be half of the value obtained for the two-sided test. Do you think that this approach is justified?



(a) JMP



(b) Minitab

Proportion (Risk) Difference Test	
H0 : P1 - P2 = 0	
Proportion Difference	0.0708
ASE (H0)	0.0295
Z	2.4013
One-sided Pr > Z	0.0082
Two-sided Pr >  Z	0.0163

Done

(c) SAS

**Figure 8.6**

(a) JMP, (b) Minitab, and (c) SAS output for the Facebook time significance test in Example 8.10.

**USE YOUR KNOWLEDGE**

**8.50 Gender and commercial preference: the z test**

Refer to Exercise 8.48 (page 513). Test whether the proportions of women and men who liked Commercial A are the same versus the two-sided alternative at the 5% level.

**8.51 Changing the alternative hypothesis**

Refer to the previous exercise. Does your conclusion change if you test

whether the proportion of men who favor Commercial A is larger than the proportion of females? Explain.

## BEYOND THE BASICS

### Relative risk

We summarized the comparison of the increased Facebook time during the past year for women and men by reporting the difference in the proportions with a confidence interval. Another way to compare two proportions is to take the ratio. This approach can be used in any setting and it is particularly common in medical settings.

We think of each proportion as a **risk** that something (usually bad) will happen. We then compare these two risks with the ratio of the two proportions, which is called the **relative risk** (RR). Note that a relative risk of 1 means that the two proportions,  $p^1$  and  $p^2$ , are equal. The procedure for calculating confidence intervals for relative risk is based on the same kind of principles that we have studied, but the details are somewhat more complicated. Fortunately, we can leave the details to software and concentrate on interpretation and communication of the results.

risk

relative risk

## Example

### 8.13 Aspirin and blood clots: relative risk

A study of patients who had blood clots (venous thromboembolism) and had completed the standard treatment were randomly assigned to receive a low-dose aspirin or a placebo treatment. The 822 patients in the study were randomized to the treatments, 411 to each. Patients were monitored for several years for the occurrence of several related medical conditions. Counts of

patients who experienced one or more of these conditions were reported for each year after the study began.<sup>18</sup> The following table gives the data for a composite of events, termed “major vascular events.” Here,  $X$  is the number of patients who had a major event.

Population	$n$	$X$	$\hat{p}=X/n$
1 (aspirin)	411	45	0.1095
2 (placebo)	411	73	0.1776
Total	822	118	0.1436

The relative risk is

$$RR = \hat{p}^1 / \hat{p}^2 = 0.1095 / 0.1776 = 0.6164$$

Software gives the 95% confidence interval as 0.4364 to 0.8707. Taking aspirin has reduced the occurrence of major events to 62% of what it is for patients taking the placebo. The 95% confidence interval is 44% to 87%.

Note that the confidence interval is not symmetric about the estimate. Relative risk is one of many situations where this occurs.

## SECTION 8.2 Summary

The **large-sample estimate of the difference in two population proportions** is

$$D = \hat{p}^1 - \hat{p}^2$$

where  $\hat{p}^1$  and  $\hat{p}^2$  are the sample proportions:

$$\hat{p}^1 = X_1/n_1 \text{ and } \hat{p}^2 = X_2/n_2$$

The **standard error of the difference**  $D$  is

$$SED = \sqrt{\hat{p}^1(1-\hat{p}^1)/n_1 + \hat{p}^2(1-\hat{p}^2)/n_2}$$

The **margin of error for confidence level**  $C$  is

$$m = z^* SED$$

where  $z^*$  is the value for the standard Normal density curve with area  $C$  between  $-z^*$  and  $z^*$ . The **large-sample level**  $C$  **confidence interval** is

$$D \pm m$$

We recommend using this interval for 90%, 95%, or 99% confidence when the number of successes and the number of failures in both samples are all at least 10. When sample sizes are smaller, alternative procedures such as the **plus four estimate of the difference in two population proportions** are recommended.

Significance tests of  $H_0: p_1 = p_2$  use the  **$z$  statistic**

$$z = \frac{\hat{p}^1 - \hat{p}^2}{SED_p}$$

with  $P$ -values from the  $N(0,1)$  distribution. In this statistic,

$$SEDP = \hat{p}(1 - \hat{p})(\frac{1}{n_1} + \frac{1}{n_2})$$

and  $\hat{p}$  is the **pooled estimate** of the common value of  $p_1$  and  $p_2$ :

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

Use this test when the number of successes and the number of failures in each of the samples are at least 5.

**Relative risk** is the ratio of two sample proportions:

$$RR = \frac{\hat{p}_1}{\hat{p}_2}$$

Confidence intervals for relative risk are often used to summarize the comparison of two proportions.

## SECTION 8.2 Exercises

*For Exercise 8.45 to 8.47, see page 509; for Exercise 8.48 and 8.49, see page 513; and for Exercise 8.50 and 8.51, see page 520.*

### 8.52 Identify the key elements.

For each of the following scenarios, identify the populations, the counts, and the sample sizes; compute the two proportions and find their difference.

(a) Two website designs are being compared. Fifty students have agreed to be subjects for the study, and they are randomly assigned to visit one or the other of the websites for as long as they like. For each student the study directors record whether or not the visit lasts for more than a minute. For the first design, 12 students visited for more than a minute; for the second, 5 visited for more than a minute.

(b) Samples of first-year students and fourth-year students were asked if they were in favor of a new proposed core curriculum. Among the first-year students, 85 said “Yes” and 276 said “No.” For the fourth-year students, 117 said “Yes” and 104 said “No.”

### 8.53 Apply the confidence interval guidelines.

Refer to the previous exercise. For each of the scenarios, determine whether or not the guidelines for using the large-sample method for a 95% confidence interval are satisfied. Explain your answers.

### 8.54 Find the 95% confidence interval.

Refer to Exercise 8.52. For each scenario, find the large-sample 95% confidence interval for the difference in proportions, and use the scenario to explain the meaning of the confidence interval.

### 8.55 Apply the significance test guidelines.

Refer to Exercise 8.52. For each of the scenarios, determine whether or not the guidelines for using the large-sample significance test are satisfied. Explain your answers.

### 8.56 Perform the significance test.

Refer to Exercise 8.52. For each scenario, perform the large-sample significance test, and use the scenario to explain the meaning of the significance test.

### **8.57 Find the relative risk.**

Refer to Exercise 8.52. For each scenario, find the relative risk. Be sure to give a justification for your choice of proportions to use in the numerator and the denominator of the ratio. Use the scenarios to explain the meaning of the relative risk.

### **8.58 Teeth and military service.**

In 1898 the United States and Spain fought a war over the U.S. intervention in the Cuban War of Independence. At that time the U.S. military was concerned about the nutrition of its recruits. Many did not have a sufficient number of teeth to chew the food provided to soldiers. As a result, it was likely that they would be undernourished and unable to fulfill their duties as soldiers. The requirements at that time specified that a recruit must have “at least four sound double teeth, one above and one below on each side of the mouth, and so opposed” so that they could chew food. Of the 58,952 recruits who were under the age of 20,68 were rejected for this reason. For the 43,786 recruits who were 40 or over, 3801 were rejected.<sup>19</sup>

- (a) Find the proportion of rejects for each age group.
- (b) Find a 99% confidence interval for the difference in the proportions.
- (c) Use a significance test to compare the proportions. Write a short paragraph describing your results and conclusions.
- (d) Are the guidelines for the use of the large-sample approach satisfied for your work in parts (b) and (c)? Explain your answers.

### **8.59 Physical education requirements.**

In the 1920s, about 97% of U.S. colleges and universities required a physical education course for graduation. Today, about 40% require such a course. A recent study of physical education requirements included 354 institutions: 225 private and 129 public. Among the private institutions, 60 required a physical education course, while among the public institutions, 101 required a course.<sup>20</sup>

- (a) What are the explanatory and response variables for this exercise? Justify your answers.
- (b) What are the populations?
- (c) What are the statistics?
- (d) Use a 95% confidence interval to compare the private and the public institutions with regard to the physical education requirement.
- (e) Use a significance test to compare the private and the public institutions with regard to the physical education requirement.
- (f) For parts (d) and (e), verify that the guidelines for using the large-sample methods are satisfied.
- (g) Summarize your analysis of these data in a short paragraph.

### **8.60 Exergaming in Canada.**

Exergames are active video games such as rhythmic dancing games, virtual bicycles, balance board simulators, and virtual sports simulators that require a screen and a console. A study of exergaming practiced by students from grades 10 and 11 in Montreal, Canada, examined many factors related to participation in exergaming.<sup>21</sup> Of the 358 students who reported that they stressed about their health, 29.9% said that they were exergamers. Of the 851 students who reported that they did not stress about their health, 20.8% said that they were exergamers.

- (a) Define the two populations to be compared for this exercise.
- (b) What are the counts, the sample sizes, and the proportions?
- (c) Are the guidelines for the use of the large-sample confidence interval satisfied?
- (d) Are the guidelines for the use of the large-sample significance test satisfied?

### **8.61 Confidence interval for exergaming in Canada.**

Refer to the previous exercise. Find the 95% confidence interval for the difference in proportions. Write a short statement interpreting this result.

### **8.62 Significance test for exergaming in Canada.**

Refer to Exercise 8.60. Use a significance test to compare the proportions. Write a short statement interpreting this result.

### **8.63 Adult gamers versus teen gamers.**

A Pew Internet Project Data Memo presented data comparing adult gamers with teen gamers with respect to the devices on which they play. The data are from two surveys. The adult survey had 1063 gamers while the teen survey had 1064 gamers. The memo reports that 54% of adult gamers played on game consoles (Xbox, PlayStation, Wii, etc.) while 89% of teen gamers played on game consoles.<sup>22</sup>

- (a) Refer to the table that appears on page 508. Fill in the numerical values of all quantities that are known.
- (b) Find the estimate of the difference between the proportion of teen gamers who played on game consoles and the proportion of adults who played on these devices.
- (c) Is the large-sample confidence interval for the difference between two proportions appropriate to use in this setting? Explain your answer.
- (d) Find the 95% confidence interval for the difference.
- (e) Convert your estimated difference and confidence interval to percents.
- (f) The adult survey was conducted between October and December 2008, whereas the teen survey was conducted between November 2007 and February 2008. Do you think that this difference should have any effect on the interpretation of the results? Be sure to explain your answer.

### **8.64 Significance test for gaming on consoles.**

Refer to the previous exercise. Test the null hypothesis that the two proportions are equal. Report the test statistic with the  $P$ -value and summarize your conclusion.

### 8.65 Gamers on computers.

The report described in Exercise 8.63 also presented data from the same surveys for gaming on computers (desktops or laptops). These devices were used by 73% of adult gamers and by 76% of teen gamers. Answer the questions given in Exercise 8.63 for gaming on computers.

### 8.66 Significance test for gaming on computers.

Refer to the previous exercise. Test the null hypothesis that the two proportions are equal. Report the test statistic with the  $P$ -value and summarize your conclusion.

### 8.67 Can we compare gaming on consoles with gaming on computers?

Refer to the previous four exercises. Do you think that you can use the large-sample confidence intervals for a difference in proportions to compare teens' use of computers with teens' use of consoles? Write a short paragraph giving the reason for your answer. (*Hint*: Look carefully in the box giving the assumptions needed for this procedure.)

### 8.68 Draw a picture.

Suppose that there are two binomial populations. For the first, the true proportion of successes is 0.3; for the second, it is 0.5. Consider taking independent samples from these populations, 40 from the first and 60 from the second.

- Find the mean and the standard deviation of the distribution of  $p^1 - p^2$ .
- This distribution is approximately Normal. Sketch this Normal distribution and mark the location of the mean.
- Find a value  $d$  for which the probability is 0.95 that the difference in sample proportions is within  $\pm d$ . Mark these values on your sketch.

### 8.69 What's wrong?

For each of the following, explain what is wrong and why.

- A  $z$  statistic is used to test the null hypothesis that  $p^1 = p^2$ .
- If two sample proportions are equal, then the sample counts are equal.
- A 95% confidence interval for the difference in two proportions includes errors due to nonresponse.



### 8.70 $p^1 - p^2$ and the Normal distribution

Refer to Exercise 8.68. Assume that all the conditions for that exercise remain the same, with the exception that  $n_2 = 1200$

- Find the mean and the standard deviation of the distribution of  $p^1 - p^2$ .
- Find the mean and the standard deviation of the distribution of  $p^1 - 0.5$ .
- Because  $n_2$  is very large, we expect  $p^2$  to be very close to 0.5. How close?

(d) Summarize what you have found in parts (a), (b), and (c) of this exercise. Interpret your results in terms of inference for comparing two proportions when the sample size of one of the samples is much larger than the sample size of the other.

### 8.71 Gender bias in textbooks.

To what extent do syntax textbooks, which analyze the structure of sentences, illustrate gender bias? A study of this question sampled sentences from 10 texts.<sup>23</sup> One part of the study examined the use of the words “girl,” “boy,” “man,” and “woman.” We will call the first two words *juvenile* and the last two *adult*. Is the proportion of female references that are juvenile (girl) equal to the proportion of male references that are juvenile (boy)? Here are data from one of the texts:

Gender	$n$	$X$ (juvenile)
Female	60	48
Male	132	52

- (a) Find the proportion of juvenile references for females and its standard error. Do the same for the males.
- (b) Give a 90% confidence interval for the difference and briefly summarize what the data show.
- (c) Use a test of significance to examine whether the two proportions are equal.

## CHAPTER 8 Exercises

### 8.72 The future of gamification.

Gamification is an interactive design that includes rewards such as points, payments, and gifts. A Pew survey of 1021 technology stakeholders and critics was conducted to predict the future of gamification. A report on the survey said that 42% of those surveyed thought that there would be no major increases in gamification by 2020. On the other hand, 53% said that they believed that there would be significant advances in the adoption and use of gamification by 2020.<sup>24</sup> Analyze these data using the methods that you learned in this chapter and write a short report summarizing your work.

### 8.73 Where do you get your news?

A report produced by the Pew Research Center's Project for Excellence in Journalism summarized the results of a survey on how people get their news. Of the 2342 people in the survey who own a desktop or laptop, 1639 reported that they get their news from the desktop or laptop.<sup>25</sup>

- Identify the sample size and the count.
- Find the sample proportion and its standard error.
- Find and interpret the 95% confidence interval for the population proportion.
- Are the guidelines for use of the large-sample confidence interval satisfied? Explain your answer.

### 8.74 Is the calcium intake adequate?

Young children need calcium in their diet to support the growth of their bones. The Institute of Medicine provides guidelines for how much calcium should be consumed by people of different ages.<sup>26</sup> One study examined whether or not a sample of children consumed an adequate amount of calcium based on these guidelines. Since there are different guidelines for children aged 5 to 10 years and those aged 11 to 13 years, the children were classified into these two age groups. Each student's calcium intake was classified as meeting or not meeting the guideline. There were 2029 children in the study. Here are the data:<sup>27</sup>

	Age (years)	
Met requirement	5 to 10	11 to 13
No	194	557
Yes	861	417

Identify the populations, the counts, and the sample sizes for comparing the extent to which the two age groups of children met the calcium intake requirement.

### 8.75 Use a confidence interval for the comparison.

Refer to the previous exercise. Use a 95% confidence interval for the comparison and explain what the confidence interval tells us. Be sure to include a justification for the use of the large-sample procedure for this comparison.

### **8.76 Use a significance test for the comparison.**

Refer to Exercise 8.74. Use a significance test to make the comparison. Interpret the result of your test. Be sure to include a justification for the use of the large-sample procedure for this comparison.

### **8.77 Confidence interval or significance test?**

Refer to Exercise 8.74 to 8.76. Do you prefer to use the confidence interval or the significance test for this comparison? Give reasons for your answer.

### **8.78 Punxsutawney Phil.**

There is a gathering every year on February 2 at Gobbler's Knob in Punxsutawney, Pennsylvania. A groundhog, always named Phil, is the center of attraction. If Phil sees his shadow when he emerges from his burrow, tradition says that there will be six more weeks of winter. If he does not see his shadow, spring has arrived. How well has Phil predicted the arrival of spring for the past several years? The National Oceanic and Atmospheric Administration has collected data for the 25 years from 1988 to 2012. For each year, whether or not Phil saw his shadow is recorded. This is compared with the February temperature for that year, classified as above or below normal. For 18 of the 25 years, Phil saw his shadow, and for 6 of these years, the temperature was below normal. For the years when Phil did not see his shadow, 2 of these years had temperatures below normal.<sup>28</sup> Analyze the data and write a report on how well Phil predicts whether or not winter is over.

### **8.79 Facebook users.**

A Pew survey of 1802 Internet users found that 67% use Facebook.<sup>29</sup>

- (a) How many of those surveyed used Facebook?
- (b) Give a 95% confidence interval for the proportion of Internet users who use Facebook.
- (c) Convert the confidence interval that you found in part (b) to a confidence interval for the percent of Internet users who use Facebook.

### **8.80 Twitter users.**

Refer to the previous exercise. The same survey reported that 16% of Internet users use Twitter. Answer the questions in the previous exercise for Twitter use.

### **8.81 Facebook versus Twitter.**

Refer to Exercise 8.79 and 8.80. Can you use the data provided in these two exercises to compare the proportion of Facebook users with the proportion of Twitter users? If your answer is yes, do the comparison. If your answer is no, explain why you cannot make the comparison.

## 8.82 Video game genres.

U.S. computer and video game software sales were \$13.26 billion in 2012.<sup>30</sup> A survey of 1102 teens collected data about video game use by teens. According to the survey, the following are the most popular game genres.<sup>31</sup>

Genre	Examples	Percent who play
Racing	NASCAR, Mario Kart, Burnout	74
Puzzle	Bejeweled, Tetris, Solitaire	72
Sports	Madden, FIFA, Tony Hawk	68
Action	Grand Theft Auto, Devil May Cry, Ratchet and Clank	67
Adventure	Legend of Zelda, Tomb Raider	66
Rhythm	Guitar Hero, Dance Dance Revolution, Lumines	61

Give a 95% confidence interval for the proportion who play games in each of these six genres.

## 8.83 Too many errors.

Refer to the previous exercise. The chance that each of the six intervals that you calculated includes the true proportion for that genre is approximately 95%. In other words, the chance that your interval misses the true value is approximately 5%.

(a) Explain why the chance that at least one of your intervals does not contain the true value of the parameter is greater than 5%.

(b) One way to deal with this problem is to adjust the confidence level for each interval so that the overall probability of at least one miss is 5%. One simple way to do this is to use a **Bonferroni procedure**. Here is the basic idea: You have an error budget of 5% and you choose to spend it equally on six intervals. Each interval has a budget of  $0.05/6=0.008$ . So, each confidence interval should have a 0.8% chance of missing the true value. In other words, the confidence level for each interval should be  $1-0.008=0.992$ . Use Table A to find the value of  $z^*$  for a large-sample confidence interval for a single proportion corresponding to 99.2% confidence.

(c) Calculate the six confidence intervals using the Bonferroni procedure.

## 8.84 Changes in credit card usage by undergraduates.

In Exercise 8.31 (page 506) we looked at data from a survey of 1430 undergraduate students and their credit card use. In the sample, 43% said that they had four or more credit cards. A similar study performed four years earlier by the same organization reported that 32% of the sample said that they had four or more credit cards.<sup>32</sup> Assume that the sample sizes for the two studies are the same. Find a 95% confidence interval for the change in the percent of undergraduates who report having four or more credit cards.

## 8.85 Do the significance test for the change.

Refer to the previous exercise. Perform the significance test for comparing the two proportions. Report your test statistic, the  $P$ -value, and summarize your conclusion.

## 8.86 We did not know the sample size.

Refer to the previous two exercises. We did not report the sample size for the earlier study, but it is reasonable to assume that it is close to the sample size for the later study.

(a) Suppose that the sample size for the earlier study was only 800. Redo the confidence interval and significance test calculations for this scenario.

(b) Suppose that the sample size for the earlier study was 2500. Redo the confidence interval and significance test calculations for this scenario.

(c) Compare your results for parts (a) and (b) of this exercise with the results that you found in the previous two exercises. Write a short paragraph about the effects of assuming a value for the sample size on your conclusions.

### **8.87 Student employment during the school year.**

A study of 1530 undergraduate students reported that 1006 work 10 or more hours a week during the school year. Give a 95% confidence interval for the proportion of all undergraduate students who work 10 or more hours a week during the school year.

### **8.88 Examine the effect of the sample size.**

Refer to the previous exercise. Assume a variety of different scenarios where the sample size changes, but the proportion in the sample who work 10 or more hours a week during the school year remains the same. Write a short report summarizing your results and conclusions. Be sure to include numerical and graphical summaries of what you have found.

### **8.89 Gender and soft drink consumption.**

Refer to Exercise 8.24 (page 505). This survey found that 16% of the 2006 New Zealanders surveyed reported that they consumed five or more servings of soft drinks per week. The corresponding percents for men and women were 17% and 15%, respectively. Assuming that the numbers of men and women in the survey are approximately equal, do the data suggest that the proportions vary by gender? Explain your methods, assumptions, results, and conclusions.

### **8.90 Examine the effect of the sample size.**

Refer to the previous exercise. Assume the following values for the total sample size: 1000, 4000, 10,000. Also assume that the sample proportions do not change. For each of these scenarios, redo the calculations that you performed in the previous exercise. Write a short paragraph summarizing the effect of the sample size on the results.

### **8.91 Gallup Poll study.**

Go to the Gallup Poll website [gallup.com](http://gallup.com) and find a poll that has several questions of interest to you. Summarize the results of the poll giving margins of error and comparisons of interest. (For this exercise, you may assume that the data come from an SRS.)



### **8.92 More on gender bias in textbooks**

Refer to the study of gender bias and stereotyping described in Exercise 8.71 (page 524). Here are

the counts of “girl,” “woman,” “boy,” and “man” for all the syntax texts studied. The one we analyzed in Exercise 8.71 was number 6.  **GENDERS**

	Text Number									
	1	2	3	4	5	6	7	8	9	10
Girl	2	5	25	11	2	48	38	5	48	13
Woman	3	2	31	65	1	12	2	13	24	5
Boy	7	18	14	19	12	52	70	6	128	32
Man	27	45	51	138	31	80	2	27	48	95

For each text perform the significance test to compare the proportions of juvenile references for females and males. Summarize the results of the significance tests for the 10 texts studied. The researchers who conducted the study note that the authors of the last 3 texts are women, while the other 7 texts were written by men. Do you see any pattern that suggests that the gender of the author is associated with the results?

### 8.93 Even more on gender bias in textbooks

Refer to the previous exercise. Let us now combine the categories “girl” with “woman” and “boy” with “man.” For each text calculate the proportion of male references and test the hypothesis that male and female references are equally likely (that is, the proportion of male references is equal to 0.5). Summarize the results of your 10 tests. Is there a pattern that suggests a relation with the gender of the author?

### 8.94 Changing majors during college

In a random sample of 975 students from a large public university, it was found that 463 of the students changed majors during their college years.

- Give a 95% confidence interval for the proportion of students at this university who change majors.
- Express your results from (a) in terms of the *percent* of students who change majors.
- University officials concerned with counseling students are interested in the number of students who change majors rather than the proportion. The university has 37,500 undergraduate students. Convert the confidence interval you found in (a) to a confidence interval for the *number* of students who change majors during their college years.

### 8.95 Sample size and the *P*-value

In this exercise we examine the effect of the sample size on the significance test for comparing two proportions. In each case suppose that  $p^1=0.55$  and  $p^2=0.45$ , and take  $n$  to be the common value of  $n_1$  and  $n_2$ . Use the  $z$  statistic to test  $H_0:p_1=p_2$  versus the alternative  $H_a:p_1 \neq p_2$ . Compute the statistic and the associated *P*-value for the following values of  $n$ : 60, 70, 80, 100, 400, 500, and 1000. Summarize the results in a table. Explain what you observe about the effect of the sample size on statistical significance when the sample proportions  $p^1$  and  $p^2$  are unchanged.

### 8.96 Sample size and the margin of error

In Section 8.1, we studied the effect of the sample size on the margin of error of the confidence

interval for a single proportion. In this exercise we perform some calculations to observe this effect for the two-sample problem. Suppose that  $p^1=0.8$  and  $p^2=0.6$  and  $n$  represents the common value of  $n_1$  and  $n_2$ . Compute the 95% margins of error for the difference between the two proportions for  $n = 60, 70, 80, 100, 400, 500,$  and  $1000$ . Present the results in a table and with a graph. Write a short summary of your findings.

## 8.97 Calculating sample sizes for the two-sample problem

For a single proportion, the margin of error of a confidence interval is largest for any given sample size  $n$  and confidence level  $C$  when  $p^*=0.5$ . This led us to use  $p^*=0.5$  for planning purposes. The same kind of result is true for the two-sample problem. The margin of error of the confidence interval for the difference between two proportions is largest when  $p^1=p^2=0.5$ . You are planning a survey and will calculate a 95% confidence interval for the difference between two proportions when the data are collected. You would like the margin of error of the interval to be less than or equal to 0.06. You will use the same sample size  $n$  for both populations.

- How large a value of  $n$  is needed?
- Give a general formula for  $n$  in terms of the desired margin of error  $m$  and the critical value  $z^*$ .

## 8.98 A corporate liability trial.

A major court case on the health effects of drinking contaminated water took place in the town of Woburn, Massachusetts. A town well in Woburn was contaminated by industrial chemicals. During the period that residents drank water from this well, there were 16 birth defects among 414 births. In years when the contaminated well was shut off and water was supplied from other wells, there were 3 birth defects among 228 births. The plaintiffs suing the firm responsible for the contamination claimed that these data show that the rate of birth defects was higher when the contaminated well was in use.<sup>33</sup> How statistically significant is the evidence? What assumptions does your analysis require? Do these assumptions seem reasonable in this case?

## 8.99 Statistics and the law

*Castaneda v. Partida* is an important court case in which statistical methods were used as part of a legal argument.<sup>34</sup> When reviewing this case, the Supreme Court used the phrase “two or three standard deviations” as a criterion for statistical significance. This Supreme Court review has served as the basis for many subsequent applications of statistical methods in legal settings. (The two or three standard deviations referred to by the Court are values of the  $z$  statistic and correspond to  $P$ -values of approximately 0.05 and 0.0026.) In *Castaneda* the plaintiffs alleged that the method for selecting juries in a county in Texas was biased against Mexican Americans. For the period of time at issue, there were 181,535 persons eligible for jury duty, of whom 143,611 were Mexican Americans. Of the 870 people selected for jury duty, 339 were Mexican Americans.

- What proportion of eligible jurors were Mexican Americans? Let this value be  $p_0$ .
- Let  $p$  be the probability that a randomly selected juror is a Mexican American. The null hypothesis to be tested is  $H_0:p=p_0$ . Find the value of  $p^{\wedge}$  for this problem, compute the  $z$  statistic, and find the  $P$ -value. What do you conclude? (A finding of statistical significance in this circumstance does not constitute proof of discrimination. It can be used, however, to establish a prima facie case. The burden of proof then shifts to the defense.)
- We can reformulate this exercise as a two-sample problem. Here we wish to compare the proportion of Mexican Americans among those selected as jurors with the proportion of Mexican

Americans among those not selected as jurors. Let  $p_1$  be the probability that a randomly selected juror is a Mexican American, and let  $p_2$  be the probability that a randomly selected nonjuror is a Mexican American. Find the  $z$  statistic and its  $P$ -value. How do your answers compare with your results in part (b)?

### 8.100 Home court advantage

In many sports there is a home field or home court advantage. This means that the home team is more likely to win when playing at home than when playing at an opponent's field or court, all other things being equal. Go to the website of your favorite sports team and find the proportion of wins for home games and the proportion of wins for away games. Now consider these games to be a random sample of the process that generates wins and losses. A complete analysis of data like these requires methods that are beyond what we have studied, but the methods discussed in this chapter will give us a reasonable approximation. Examine the home court advantage for your team and write a summary of your results. Be sure to comment on the effect of the sample size.

### 8.101 Attitudes toward student loan debt

The National Student Loan Survey asked the student loan borrowers in their sample about attitudes toward debt.<sup>35</sup> Below are some of the questions they asked, with the percent who responded in a particular way. Assume that the sample size is 1280 for all these questions. Compute a 95% confidence interval for each of the questions, and write a short report about what student loan borrowers think about their debt.

- (a) "Do you feel burdened by your student loan payments?" 55.5% said they felt burdened.
- (b) "If you could begin again, taking into account your current experience, what would you borrow?" 54.4% said they would borrow less.
- (c) "Since leaving school, my education loans have not caused me more financial hardship than I had anticipated at the time I took out the loans." 34.3% disagreed.
- (d) "Making loan payments is unpleasant, but I know that the benefits of education loans are worth it." 58.9% agreed.
- (e) "I am satisfied that the education I invested in with my student loan(s) was worth the investment for career opportunities." 58.9% agreed.
- (f) "I am satisfied that the education I invested in with my student loan(s) was worth the investment for personal growth." 71.5% agreed.

# 9 Analysis of Two-Way Tables

CHAPTER



9.1 Inference for Two-Way Tables

9.2 Goodness of Fit

## Introduction

We continue our study of methods for analyzing categorical data in this chapter. Inference about proportions in one-sample and two-sample settings was the focus of Chapter 8. We now study how to compare two or more populations when the response variable has two or more categories and how to test whether two categorical variables are independent. A single statistical test handles both of these cases.

The first section of this chapter gives the basics of statistical inference that are appropriate in this setting. A goodness-of-fit test is presented in the second section. The methods in this chapter answer questions such as

- Are men and women equally likely to suffer lingering fear symptoms after watching scary movies like *Jaws* and *Poltergeist* at a young age?
- Is there an association between texting while driving and automobile accidents?
- Does political preference predict whether a person makes contributions online?

## 9.1 Inference for Two-Way Tables

When you complete this section, you will be able to

- Translate a problem from a comparison of two proportions to an analysis of a  $2 \times 2$  table.
- Find the joint distribution, the marginal distributions, and the conditional distributions for a two-way table of counts.
- Identify the joint distribution, the marginal distributions, and the conditional distributions for a two-way table from software output.
- Distinguish between settings where the goal is to describe a relationship between an explanatory variable and a response variable or to just explain the relationship between two categorical variables. If there are explanatory and response variables, identify them.
- Choose appropriate conditional distributions to describe relationships in a two-way table.
- Compute expected counts from the counts in a two-way table.
- Compute the chi-square statistic and the  $P$ -value from the expected counts in a two-way table. Use the  $P$ -value to draw your conclusion.
- For a  $2 \times 2$  table, explain the relationship between the chi-square test and the  $z$  test for comparing two proportions.
- Distinguish between two models for two-way tables.

When we studied inference for two proportions in Chapter 8, we started summarizing the raw data by giving the number of observations in each population ( $n$ ) and how many of these were classified as “successes” ( $X$ ).

### Example

#### 9.1 Are you spending more time on Facebook?



FACE

In Example 8.10 (page 510), we compared the proportions of women and men who said that they increased the amount of time that they spent on Facebook during the past year. The following table summarizes the data used in this comparison:

Population	$n$	$X$	$\hat{p} = X/n$
1 (women)	292	47	0.1610
2 (men)	233	21	0.0901
Total	525	68	0.1295

These data suggest that the percent of women who increased the amount of time spent on Facebook is 7.1% larger than the percent of men, with a 95% margin of error of 5.6%.

In this chapter we consider a different summary of the data. Rather than recording just the count of those who spent more time on Facebook during the past year, we record counts of all the outcomes in a two-way table.



two-way table, p. 139

## Example

### 9.2 Two-way table for time spent on Facebook.



Here is the two-way table classifying Facebook users by gender and whether or not they increased the amount of time that they spent on Facebook during the past year:

Increased	Gender		Total
	Women	Men	
Yes	47	21	68
No	245	212	457
Total	292	233	525

We use the term  $r \times c$  table to describe a two-way table of counts with  $r$  rows and  $c$  columns. The two categorical variables in the  $2 \times 2$  table of Example 9.2 are “Increased” and “Gender.” “Increased” is the row variable, with values “Yes” and “No,” and “Gender” is the column variable, with values “Men” and “Women.” Since the objective in this example is to compare the genders, we view “Gender” as an explanatory variable, and therefore, we make it the column variable. The next example presents another two-way table.

$r \times c$  table

## Example

### 9.3 Lingering symptoms from frightening movies.



There is a growing body of literature demonstrating that early exposure to frightening movies is associated with lingering fright symptoms. As part of a class on media effects, college students were asked to write narrative accounts of their exposure to frightening movies before the age of 13. More than one-fourth of the respondents said that some of the fright symptoms were still present in waking life.<sup>1</sup> The following table breaks down these results by gender:

Observed numbers of students			
Ongoing fright symptoms	Gender		Total
	Female	Male	
No	50	31	81
Yes	29	7	36
Total	79	38	117

The two categorical variables in Example 9.3 are “Ongoing fright symptoms,” with values “Yes” and “No,” and “Gender,” with values “Female” and “Male.” Again we view “Gender” as an explanatory variable and “Ongoing fright symptoms” as a categorical response variable.

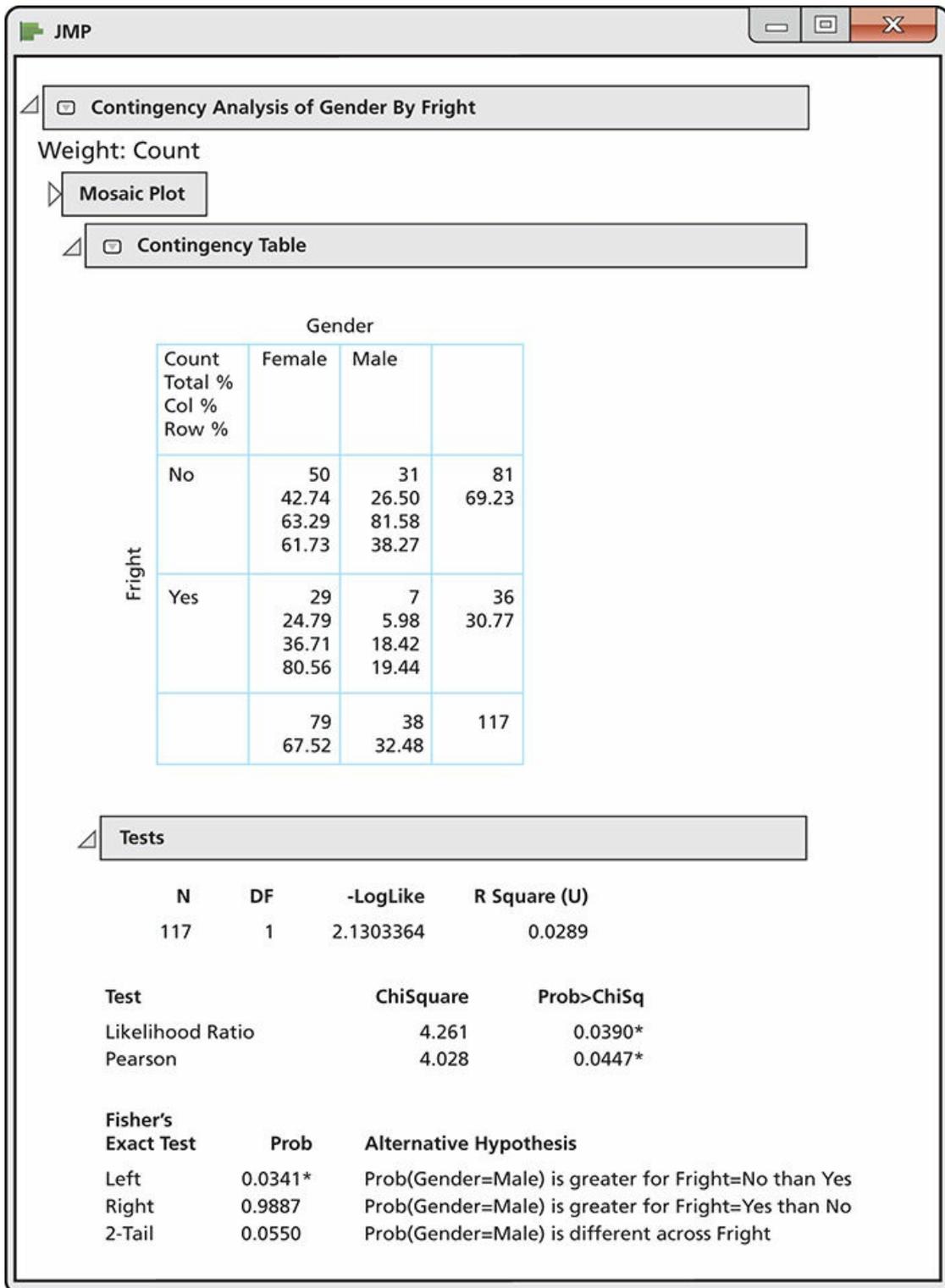
In Chapter 2 we discussed two-way tables and the basics about joint, marginal, and conditional distributions. We now view those sample distributions as estimates of the corresponding population distributions. Let’s look at some software output that gives these distributions.

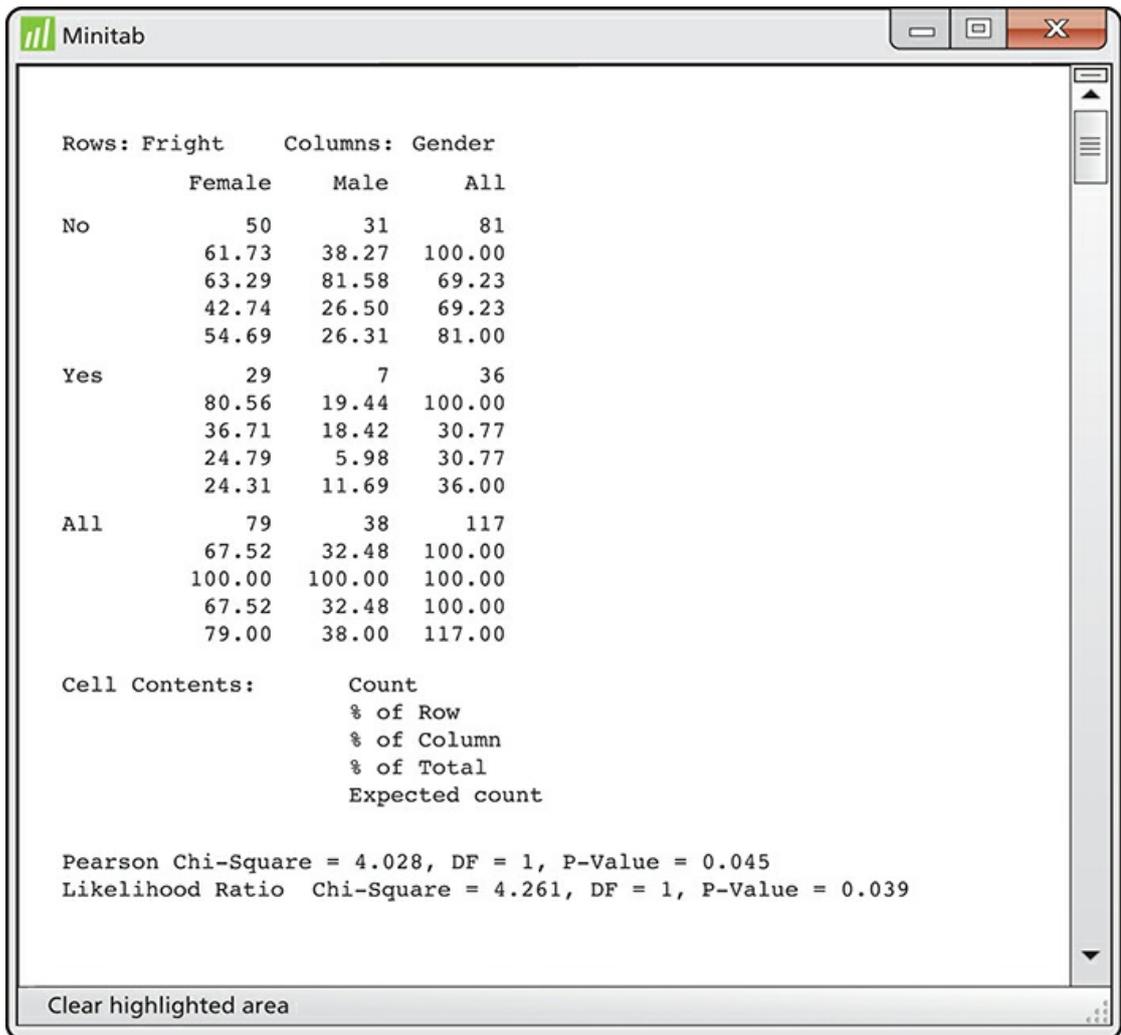
## Example

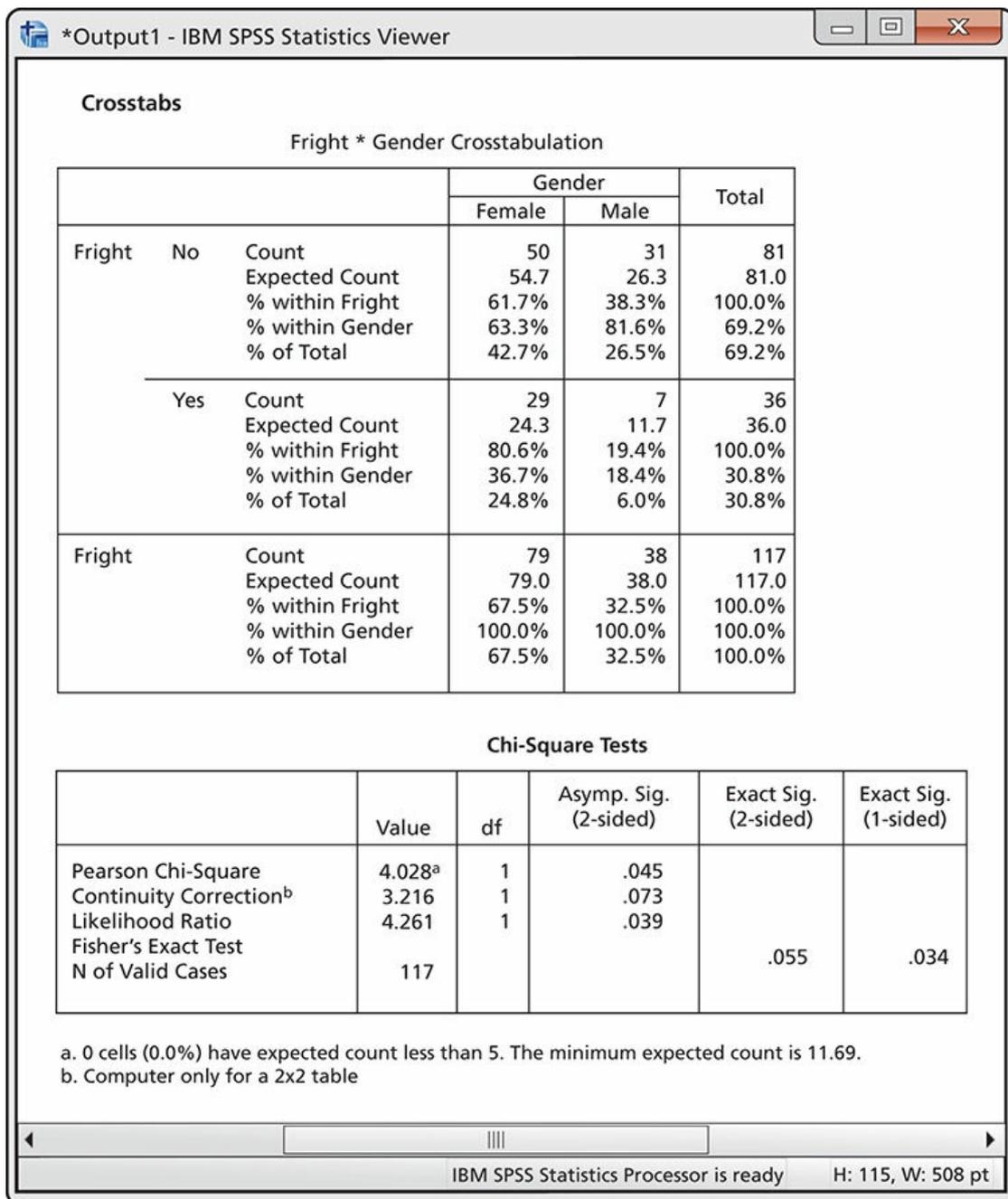
### 9.4 Software output for ongoing fright symptoms.



Figure 9.1 shows the output from JMP, Minitab, and SPSS for the fright symptoms data of Example 9.3. For now, we will just concentrate on the different distributions. Later, we will explore other parts of the output.







**FIGURE 9.1**

Computer output for Examples 9.3 and 9.4.

The three packages use similar displays for the distributions. In the cells of the  $2 \times 2$  table we find the counts, the conditional distributions of the column variable for each value of the row variable, the conditional distributions of the row variable for each value of the column variable, and the joint distribution. All of these are expressed as percents rather than proportions.

Let's look at the entries in the upper-left cell of the JMP output. We see that there are 50 females whose response is "No" to the fright symptoms question. These 50 represent 42.74% of the study participants. They represent 63.29% of the females in the study. And they represent 61.73% of the people

who responded “No” to the fright symptoms question. The marginal distributions are in the rightmost column and the bottom row. Minitab and SPSS give the same information but not necessarily in the same order.

In Chapter 2, we learned that the key to examining the relationship between two categorical variables is to look at conditional distributions. Let’s do that for the fright symptoms data.



conditional distributions, p. 144

## Example

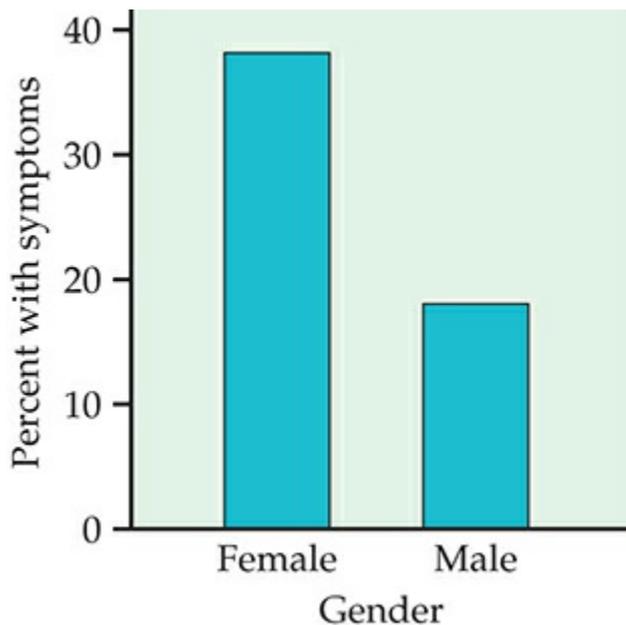
### 9.5 Two-way table of ongoing fright symptoms and gender.



To compare the frequency of lingering fright symptoms across genders, we examine column percents. Here they are, rounded from the output in Figure 9.1 for clarity:

Column percents for gender		
Ongoing fright symptoms	Gender	
	Male	Female
Yes	18%	37%
No	82%	63%
Total	100%	100%

The “Total” row reminds us that 100% of the male and female students have been classified as having ongoing fright symptoms or not. (The sums sometimes differ slightly from 100% because of roundoff error.) The bar graph in Figure 9.2 compares the percents. The data reveal a clear relationship: 37% of the women have ongoing fright symptoms, as opposed to only 18% of the men.



**FIGURE 9.2**

Bar graph of the percents of male and female students with ongoing fright symptoms.

The difference between the percents of students with lingering fears is reasonably large. A statistical test will tell us whether or not this difference can be plausibly attributed to chance. Specifically, if there is no association between gender and having ongoing fright symptoms, how likely is it that a sample would show a difference as large or larger than that displayed in Figure 9.2? In the remainder of this section we discuss the significance test to examine this question.

## USE YOUR KNOWLEDGE

### 9.1 Find two conditional distributions.



Use the output in Figure 9.3 (page 536) to answer the following questions.

- Find the conditional distribution of increased Facebook time for females.
- Do the same for males.
- Graphically display the two conditional distributions.

(d) Write a short summary interpreting the two conditional distributions.

## 9.2 Condition on Facebook time.

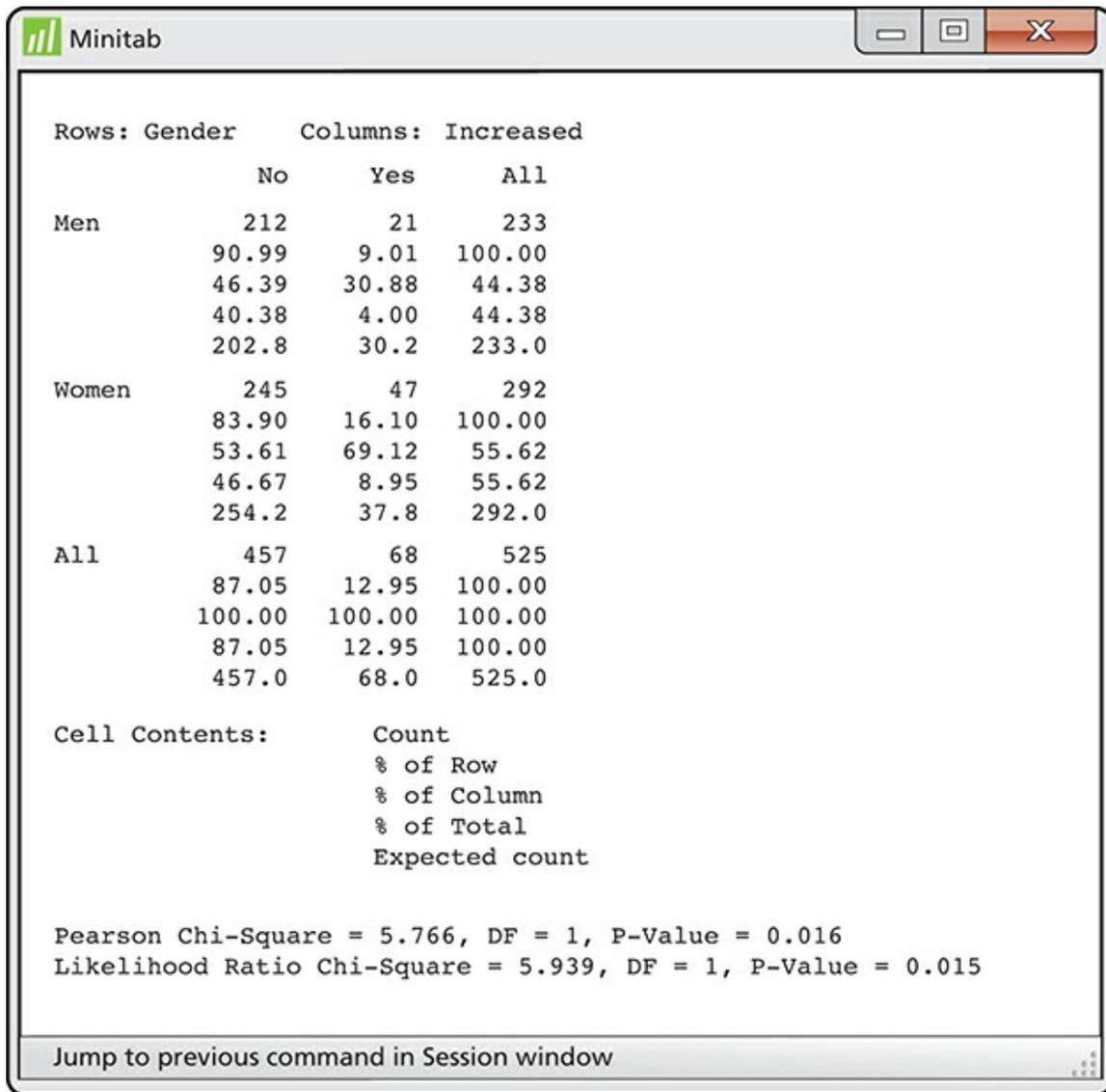


Refer to Exercise 9.1 (page 530). Use the output in Figure 9.3 (page 536) to answer the following questions.

- (a) Find the conditional distribution of gender for those who have increased their Facebook time in the past year.
- (b) Do the same for those who did not increase their Facebook time.
- (c) Graphically display the two conditional distributions.
- (d) Write a short summary interpreting the two conditional distributions.

## 9.3 Which conditional distributions should you use?

Refer to your answers to the two previous exercises. Which of these distributions do you prefer for interpreting these data? Give reasons for your answer.



**FIGURE 9.3**  
 Computer output for Exercises 9.1 to 9.3.

### The hypothesis: no association

The null hypothesis  $H_0$  of interest in a two-way table is “There is *no association* between the row variable and the column variable.” In Example 9.3, this null hypothesis says that gender and having ongoing fright symptoms are not related. The alternative hypothesis  $H_a$  is that there is an association between these two variables. The alternative  $H_a$  does not specify any particular direction for the association. For two-way tables in general, the alternative includes many different possibilities. Because it includes all sorts of possible associations, we cannot describe  $H_a$  as either one-sided or two-sided.

In our example, the hypothesis  $H_0$  that there is no association between gender and having ongoing fright symptoms is equivalent to the statement that the

variables “ongoing fright symptoms” and “gender” are independent. For other two-way tables, where the columns correspond to independent samples from  $c$  distinct populations, there are  $c$  distributions for the row variable, one for each population. The null hypothesis then says that the  $c$  distributions of the row variable are identical. The alternative hypothesis is that the distributions are not all the same.

## Expected cell counts

To test the null hypothesis in  $r \times c$  tables, we compare the observed cell counts with **expected cell counts** calculated under the assumption that the null hypothesis is true. A numerical summary of the comparison will be our test statistic.

### expected cell counts

## Example

### 9.6 Expected counts from software.

The observed and expected counts for the ongoing fright symptoms example appear in the Minitab and SPSS computer outputs shown in Figure 9.1 (pages 532–534). The expected counts are given as the last entry in each cell for Minitab and as the second entry in each cell for SPSS. For example, in the cell for males with fright symptoms, the observed count is 7 and the expected count is 11.69 (Minitab) or 11.7 (SPSS).

How is this expected count obtained? Look at the percents in the right margin of the tables in Figure 9.1. We see that 30.77% of all students had ongoing fright symptoms. If the null hypothesis of no relation between gender and ongoing fright is true, we expect this overall percent to apply to both men and women. In particular, we expect 30.77% of the men to have lingering fright symptoms. Since there are 38 men, the expected count is 30.77% of 38, or 11.69. The other expected counts are calculated in the same way.

The reasoning of Example 9.6 leads to a simple formula for calculating expected cell counts. To compute the expected count of men with ongoing fright symptoms, we multiplied the proportion of students with fright symptoms ( $36/117$ ) by the number of men (38). From Figure 9.1 we see that the numbers 36 and 38 are the row and column totals for the cell of interest and that 117 is  $n$  the total number of observations for the table. The expected cell count is therefore the product of the row and column totals divided by the table total.

## EXPECTED CELL COUNTS

$$\text{expected count} = \text{row total} \times \text{column total} / \text{total}$$

### The chi-square test

To test the  $H_0$  that there is no association between the row and column classifications, we use a statistic that compares the entire set of observed counts with the set of expected counts. To compute this statistic,

- First, take the difference between each observed count and its corresponding expected count, and square these values so that they are all 0 or positive.
- Since a large difference means less if it comes from a cell that is expected to have a large count, divide each squared difference by the expected count. This is a type of standardization.
- Finally, sum over all cells.

The result is called the *chi-square statistic*  $X^2$ . The chi-square statistic was proposed by the English statistician Karl Pearson (1857–1936) in 1900. It is the oldest inference procedure still used in its original form.

## CHI-SQUARE STATISTIC

The **chi-square statistic** is a measure of how much the observed cell counts in a two-way table diverge from the expected cell counts. The formula for the statistic is

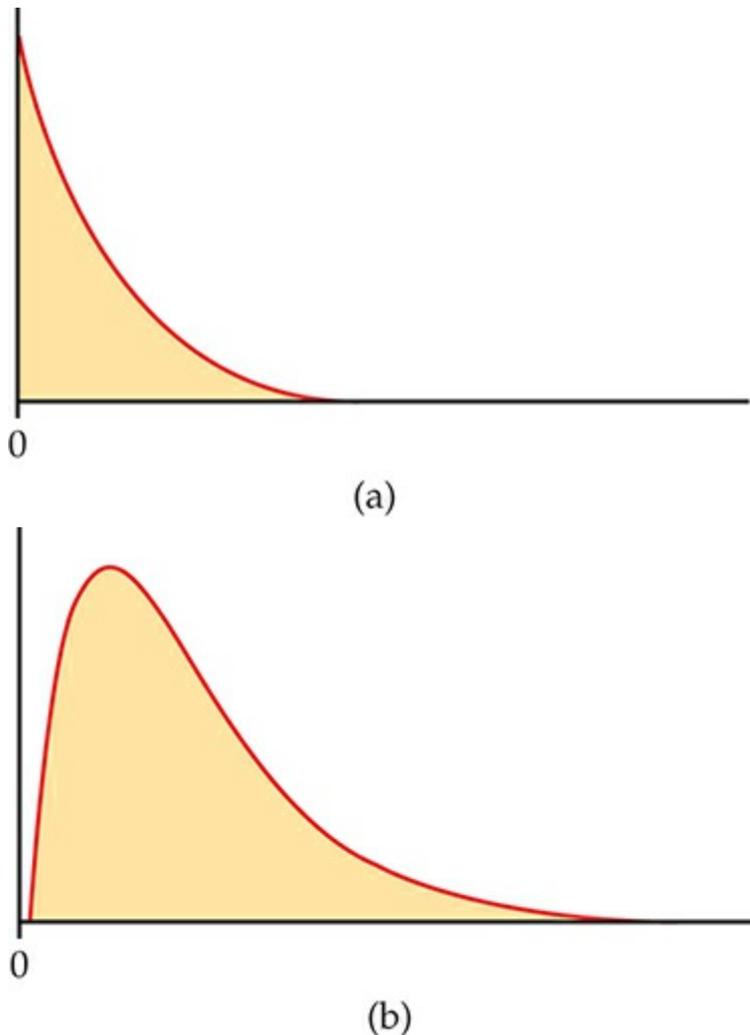
$$X^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

where “observed” represents an observed cell count, “expected” represents the expected count for the same cell, and the sum is over all  $r \times c$  cells in the table.

If the expected counts and the observed counts are very different, a large value of  $X^2$  will result. Large values of  $X^2$  provide evidence against the null hypothesis. To obtain a  $P$ -value for the test, we need the sampling distribution of  $X^2$  under the assumption that  $H_0$  (no association between the row and column variables) is true. The distribution is called the **chi-square distribution**, which we denote by  $\chi^2$  ( $\chi$  is the lowercase Greek letter chi).

chi-square distribution  $\chi^2$

Like the  $t$  distributions, the  $\chi^2$  distributions form a family described by a single parameter, the degrees of freedom. We use  $\chi^2(\text{df})$  to indicate a particular member of this family. Figure 9.4 displays the density curves of the  $\chi^2(2)$  and  $\chi^2(4)$  distributions. As you can see in the figure,  $\chi^2$  distributions take only positive values and are skewed to the right. Table F in the back of the book gives upper critical values for the  $\chi^2$  distributions.



**FIGURE 9.4**

(a) The  $\chi^2(2)$  density curve. (b) The  $\chi^2(4)$  density curve.

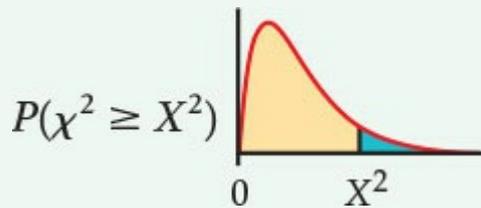
## CHI-SQUARE TEST FOR TWO-WAY TABLES

The null hypothesis  $H^0$  is that there is no association between the row and column variables in a two-way table. The alternative hypothesis is that these variables are related.

If  $H^0$  is true, the chi-square statistic  $X^2$  has approximately a  $\chi^2$  distribution with  $(r - 1)(c - 1)$

degrees of freedom.

The  $P$ -value for the chi-square test is



where  $\chi^2$  is a random variable having the distribution with

For tables larger than  $2 \times 2$ , we will use this approximation whenever the average of the expected counts is 5 or more and the smallest expected count is 1 or more. For  $2 \times 2$  tables, we require all four expected cell counts to be 5 or more.<sup>2</sup>

The chi-square test always uses the upper tail of the  $\chi^2$  distribution, because any deviation from the null hypothesis makes the statistic larger. The approximation of the distribution of  $X^2$  by  $\chi^2$  becomes more accurate as the cell counts increase. Moreover, it is more accurate for tables larger than  $2 \times 2$  tables.

## Example

### 9.7 Chi-square significance test from software.



#### FRIGHT

The results of the chi-square significance test for the ongoing fright symptoms example appear in the computer outputs in Figures 9.1 (pages 532–534), labeled Pearson or Pearson Chi-Square. Because all the expected cell counts are moderately large (5 or more), the  $\chi^2$  distribution provides an accurate  $P$ -value. We see that  $X^2 = 4.03$ ,  $df = 1$ , and  $P = 0.045$ . As a check we verify that the degrees of freedom are correct for a  $2 \times 2$  table:

$$df = (r-1)(c-1) = (2-1)(2-1) = 1$$

The chi-square test confirms that the data provide evidence against the null hypothesis that there is no relationship between gender and ongoing fright symptoms. Under  $H^0$ , the chance of obtaining a value of  $X^2$  greater than or

equal to the calculated value of 4.03 is small, 0.045—fewer than 5 times in 100.

The test does not provide insight into the nature of the relationship between the variables. It is up to us to see that the data show that women are more likely to have lingering fright symptoms. You should always accompany a chi-square test by percents such as those in Example 9.5 and Figure 9.2 and by a description of the nature of the relationship.

The observational study of Example 9.3 cannot tell us whether gender is a *cause* of lingering fright symptoms. The association may be explained by confounding with other variables. For example, other research has shown that there are gender differences in the social desirability of admitting fear.<sup>3</sup> *Our data don't allow us to investigate possible confounding variables.* Often a randomized comparative experiment can settle the issue of causation, but we cannot randomly assign gender to each student. The researcher who published the data of our example states merely that women are more likely to report lingering fright symptoms and that this conclusion is consistent with other studies.



confounding, p. 173



## Computations

The calculations required to analyze a two-way table are straightforward but tedious. In practice, we recommend using software, but it is possible to do the work with a calculator, and some insight can be gained by examining the details. Here is an outline of the steps required.

### COMPUTATIONS FOR TWO-WAY TABLES

1. Calculate descriptive statistics that convey the important information in the table. Usually these will be column or row percents.
2. Find the expected counts and use these to compute the  $X^2$  statistic.
3. Use chi-square critical values from Table F to find the approximate  $P$ -value.
4. Draw a conclusion about the association between the row and column variables.

The following examples illustrate these steps.

## Example

### 9.8 Health habits of college students.

Physical activity generally declines when students leave high school and enroll in college. This suggests that college is an ideal setting to promote physical activity. One study examined the level of physical activity and other health-related behaviors in a sample of 1184 college students.<sup>4</sup> Let's look at the data for physical activity and consumption of fruits. We categorize physical activity as low, moderate, or vigorous and fruit consumption as low, medium, or high. Here is the two-way table that summarizes the data:

Fruit consumption	Physical activity			Total
	Low	Moderate	Vigorous	
Low	69	206	294	569
Medium	25	126	170	321
High	14	111	169	294
Total	108	443	633	1184

The table in Example 9.8 is a  $3 \times 3$  table, to which we have added the marginal totals obtained by summing across rows and columns. For example, the first-row total is  $69 + 206 + 294 = 569$ . The grand total, the number of students in the study, can be computed by summing the row totals ( $569 + 321 + 294 = 1184$ ) or the column totals ( $108 + 443 + 633 = 1184$ ). *It is easy to make an error in these calculations, so it is a good idea to do both as a check on your arithmetic.*



## Computing conditional distributions

First, we summarize the observed relation between physical activity and fruit consumption. We expect a positive association, but there is no clear distinction between an explanatory variable and a response variable in this setting. If we have such a distinction, then the clearest way to describe the relationship is to compare

the conditional distributions of the response variable for each value of the explanatory variable. Otherwise, we can compute the conditional distribution each way and then decide which gives a better description of the data.

## Example

### 9.9 Health habits of college students: conditional distributions.



HEALTH

Let's look at the data in the first column of the table in Example 9.8. There were 108 students with low physical activity. Of these, there were 69 with low fruit consumption. Therefore, the column proportion for this cell is

$$\frac{69}{108}=0.639$$

That is, 63.9% of the low physical activity students had low fruit consumption. Similarly, 25 of the low physical activity students has moderate fruit consumption. This percent is 23.1%.

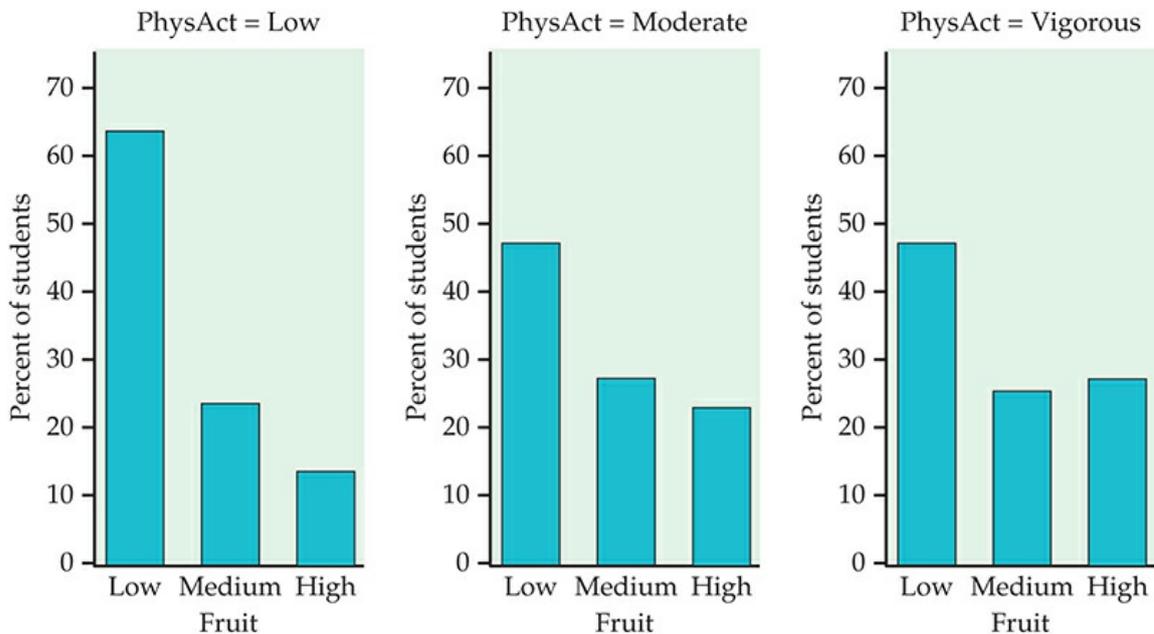
$$\frac{25}{108}=0.231$$

In all, we calculate nine percents. Here are the results:

Column percents for fruit consumption and physical activity				
Fruit consumption	Physical activity			Total
	Low	Moderate	Vigorous	
Low	63.9	46.5	46.4	48.1
Medium	23.1	28.4	26.9	27.1
High	13.0	25.1	26.7	24.8

Total	100.0	100.0	100.0	100.0
-------	-------	-------	-------	-------

In addition to the conditional distributions of fruit consumption for each level of physical activity, the table also gives the marginal distribution of fruit consumption. These percents appear in the rightmost column, labeled “Total.”



**FIGURE 9.5**

Comparison of the distribution of fruit consumption for different levels of physical activity, for Example 9.9.

The sum of the percents in each column should be 100, except for possible small roundoff errors. *It is good practice to calculate each percent separately and then sum each column as a check.* In this way we can find arithmetic errors that would not be uncovered if, for example, we calculated the column percent for the “High” row by subtracting the sum of the percents for “Low” and “Medium” from 100.



Figure 9.5 compares the distributions of fruit consumption for each of the three physical activity levels. For each activity level, the highest percent is for students who consume low amounts of fruit. For low physical activity, there is a clear decrease in the percent when moving from low to medium to high fruit consumption. The patterns for moderate physical activity and vigorous physical activity are similar. Low fruit consumption is still dominant, but the percents for medium and high fruit consumption are about the same for the moderate and vigorous activity levels. The percent of low fruit consumption is highest for the low physical activity students compared with those who have moderate or vigorous

physical activity. These plots suggest that there is an association between these two variables.

## USE YOUR KNOWLEDGE

### 9.4 Examine the row percents.



Refer to the health habits data that we examined in Example 9.8 (page 540). For the row percents, make a table similar to the one in Example 9.9 (page 541).

### 9.5 Make some plots.



Refer to the previous exercise. Make plots of the row percents similar to those in Figure 9.5.

### 9.6 Compare the conditional distributions.



Compare the plots you made in the previous exercise with those given in Figure 9.5. Which set of plots do you think gives a better graphical summary of the relationship between these two categorical variables? Give reasons for your answer. Note that there is not a clear right or wrong answer for this exercise. You need to make a choice and to explain your reasons for making it.

We observe a clear relationship between physical activity and fruit consumption in this study. The chi-square test assesses whether this observed association is statistically significant, that is, too strong to occur often just by chance. The test confirms only that there is some relationship. The percents we have compared describe the nature of the relationship.



*The chi-square test does not in itself tell us what population our conclusion describes.* The subjects in this study were college students from four midwestern universities. The researchers could argue that these findings apply to college students in general. This type of inference is important, but it is based on expert judgment and is beyond the scope of the statistical inference that we have been studying.

## Example

### 9.10 The chi-square significance test for health habits of college students.



The first step in performing the significance test is to calculate the expected cell counts. Let's start with the cell for students with low fruit consumption and low physical activity. Using the formula on page 537, we need three quantities: (1) the corresponding row total, 569, the number of students who have low fruit consumption, (2) the column total, 108, the number of students who have low physical activity, and (3) the total number of students, 1184. The expected cell count is therefore

$$(108)(569)1184=51.90$$

Note that although any observed count of the number of students must be a whole number, an expected count need not be.

Calculations for the other eight cells in the  $3 \times 3$  table are performed in the same way. With these nine expected counts we are now ready to use the formula for the  $X^2$  statistic on page 538. The first term in the sum comes from

the cell for students with low fruit consumption and low physical activity. The observed count is 69 and the expected count is 51.90. Therefore, the contribution to the  $X^2$  statistic for this cell is

$$(69-51.90)^2/51.90=5.63$$

When we add the terms for each of the nine cells, the result is

$$X^2=14.15$$

Because there are  $r = 3$  levels of fruit consumption and  $c = 3$  levels of physical activity, the degrees of freedom for this statistic are

$$df=(r-1)(c-1)=(3-1)(3-1)=4$$

Under the null hypothesis that fruit consumption and physical activity are independent, the test statistic  $X^2$  has a  $\chi^2(4)$  distribution. To obtain the  $P$ -value, look at the  $df = 4$  row in Table F. The calculated value  $X^2 = 14.15$  lies between the critical points for probabilities 0.01 and 0.005. The  $P$ -value is therefore between 0.01 and 0.005. (Software gives the value as 0.0068.) There is strong evidence ( $X^2 = 14.15$ ,  $df = 4$ ,  $P < 0.01$ ) that there is a relationship between fruit consumption and physical activity.

$$df = 4$$

$p$	0.01	0.005
$\chi^2$	13.28	14.86

We can check our work by adding the expected counts to obtain the row and column totals, as in the table. These are the same as those in the table of observed counts except for small roundoff errors.

## USE YOUR KNOWLEDGE

### 9.7 Find the expected counts.



HEALTH

Refer to Example 9.10. Compute the expected counts and display them in a  $3 \times 3$  table. Check your work by adding the expected counts to obtain row and column totals. These should be the same as those in the table of observed counts except for small roundoff errors.

### 9.8 Find the $X^2$ statistic.



HEALTH

Refer to the previous exercise. Use the formula on page 538 to compute the contributions to the chi-square statistic for each cell in the table. Verify that their sum is 14.15.

### 9.9 Find the $P$ -value.



HEALTH

For each of the following give the degrees of freedom and an appropriate bound on the  $P$ -value for the  $X^2$  statistic.

- (a)  $X^2 = 19.00$  for a  $5 \times 4$  table
- (b)  $X^2 = 19.00$  for a  $4 \times 5$  table
- (c)  $X^2 = 7.50$  for a  $2 \times 2$  table
- (d)  $X^2 = 1.60$  for a  $2 \times 2$  table

### 9.10 Time spent on Facebook: the chi-square test.



FACE

Refer to Example 9.2 (page 531). Use the chi-square test to assess the relationship between gender and increased amount of time spent on Facebook in the last year. State your conclusion.

## The chi-square test and the $z$ test

A comparison of the proportions of “successes” in two populations leads to a  $2 \times 2$  table. We can compare two population proportions either by the chi-square test or by the two-sample  $z$  test from Section 8.2. In fact, *these tests always give exactly*

the same result, because the  $X^2$  statistic is equal to the square of the  $z$  statistic, and  $z$  critical values are equal to the squares of the corresponding  $\chi^2(1)$  critical values. The advantage of the  $z$  test is that we can test either one-sided or two-sided alternatives. The chi-square test always tests the two-sided alternative. Of course, the chi-square test can compare more than two populations, whereas the  $z$  test compares only two.

## USE YOUR KNOWLEDGE

### 9.11 Comparison of conditional distributions.



Consider the following  $2 \times 2$  table.

Response variable	Observed counts		Total
	Explanatory variable		
	1	2	
Yes	75	95	170
No	135	115	250
Total	210	210	420

- Compute the conditional distribution of the response variable for each of the two explanatory-variable categories.
- Display the distributions graphically.
- Write a short paragraph describing the two distributions and how they differ.

### 9.12 Expected cell counts and the chi-square test.



Refer to Exercise 9.11. You decide to use the chi-square test to compare these two conditional distributions.

- What is the expected count for the first cell (observed count is 75)?

- (b) Computer software gives you  $\chi^2 = 3.95$ . What are the degrees of freedom for this statistic?
- (c) Using Table F, give an appropriate bound on the  $P$ -value.

### 9.13 Compare the chi-square test with the $z$ test.



Refer to the previous two exercises and the significance test for comparing two proportions (page 517).

- (a) Set up the problem as a comparison between two proportions. Describe the population proportions, state the null and alternative hypotheses, and give the sample proportions.
- (b) Carry out the significance test to compare the two proportions. Report the  $z$  statistic, the  $P$ -value, and your conclusion.
- (c) Compare the  $P$ -value for this significance test with the one that you reported in the previous exercise.
- (d) Verify that the square of the  $z$  statistic is the  $\chi^2$  statistic given in the previous exercise.

## Models for two-way tables

The chi-square test for the presence of a relationship between the two variables in a two-way table is valid for data produced from several different study designs. The precise statement of the null hypothesis of “no relationship” in terms of population parameters is different for different designs. We now describe two of these settings in detail. *An essential requirement is that each experimental unit or subject is counted only once in the data table.*



### Comparing several populations: the first model

Let’s think about the setting of Example 9.8 from a slightly different perspective. Suppose that we are interested in the relationship between physical activity and year of study in college. We will assume that the design called for independent SRSs of students from each of the four years. Here we have an example of *separate and independent random samples* from each of  $c$  populations. The  $c$  columns of the two-way table represent the populations. There is a single

categorical response variable, physical activity. The  $r$  rows of the table correspond to the values of the response variable, physical activity.

We know that the  $z$  test for comparing the two proportions of successes and the chi-square test for the  $2 \times 2$  table are equivalent. The  $r \times c$  table allows us to compare more than two populations or more than two categories of response, or both. In this setting, the null hypothesis “no relationship between column variable and row variable” becomes

$H_0$ : The distribution of the response variable is the same in all  $c$  populations.

Because the response variable is categorical, its distribution just consists of the probabilities of its  $r$  possible values. The null hypothesis says that these probabilities (or population proportions) are the same in all  $c$  populations.

## Example

### 9.11 Physical activity: comparing subpopulations based on year of study.

In our scenario based on Example 9.8, we compare four populations:

Population 1: first-year students

Population 2: second-year students

Population 3: third-year students

Population 4: fourth-year students

The null hypothesis for the chi-square test is

$H_0$ : The distribution of physical activity is the same in all four populations.

The alternative hypothesis for the chi-square test is

$H_a$ : The distribution of physical activity is not the same in all four populations.

The parameters of the model are the proportions of low, moderate, and vigorous physical activity in each of the four years of study.

More generally, if we take an independent SRS from each of  $c$  populations and classify each outcome into one of  $r$  categories, we have an  $r \times c$  table of population proportions. There are  $c$  different sets of proportions to be compared. There are  $c$  groups of subjects, and a single categorical variable with  $r$  possible values is measured for each individual.

## MODEL FOR COMPARING SEVERAL POPULATIONS USING TWO-WAY TABLES

Select independent SRSs from each of  $c$  populations, of sizes  $n_1, n_2, \dots, n_c$ . Classify each individual in a sample according to a categorical response variable with  $r$  possible values. There are  $c$  different probability distributions, one for each population.

The null hypothesis is that the distributions of the response variable are the same in all  $c$  populations. The alternative hypothesis says that these  $c$  distributions are not all the same.

### Testing independence: the second model

A second model for which our analysis of  $r \times c$  tables is valid is illustrated by the ongoing fright symptoms study, Example 9.3. There, a *single* sample from a *single* population was classified according to two categorical variables.

### Example

#### 9.12 Ongoing fright symptoms and gender: testing independence.

The single population studied is college students. Each college student was classified according to the following categorical variables: “Ongoing fright symptoms,” with possible responses “Yes” and “No,” and “Gender,” with possible responses “Men” and “Women.” The null hypothesis for the chi-square test is

$H_0$  : “Ongoing fright symptoms” and “Gender” are independent.

The parameters of the model are the probabilities for each of the four possible combinations of values of the row and column variables. If the null hypothesis is true, the multiplication rule for independent events says that these can be found as the products of outcome probabilities for each variable alone.



multiplication rule, p. 283

More generally, take an SRS from a single population and record the values of two categorical variables, one with  $r$  possible values and the other with  $c$  possible

values. The data are summarized by recording the number of individuals for each possible combination of outcomes for the two random variables. This gives  $r \times c$  an table of counts. Each of these  $r \times c$  possible outcomes has its own probability. The probabilities give the joint distribution of the two categorical variables.



joint distribution, p. 141



marginal distributions, p. 142

Each of the two categorical random variables has a distribution. These are the marginal distributions because they are the sums of the population proportions in the rows and columns.

The null hypothesis “no relationship” now states that the row and column variables are independent. The multiplication rule for independent events tells us that the joint probabilities are the products of the marginal probabilities.

## Example

### 9.13 The joint distribution and the two marginal distributions.

The joint probability distribution gives a probability for each of the four cells in our  $2 \times 2$  table of “Ongoing fright symptoms” and “Gender.” The marginal distribution for “Ongoing fright symptoms” gives probabilities for each of the two possible categories; the marginal distribution for “Gender” gives probabilities for each of the two possible gender categories.

Independence between “Ongoing fright symptoms” and “Gender” implies that the joint distribution can be obtained by multiplying the appropriate terms from the two marginal distributions. For example, the probability that a randomly chosen college student has ongoing fright symptoms *and* is male is equal to the probability that the student has ongoing symptoms *times* the probability that the student is male. The hypothesis that “Ongoing fright symptoms” and “Gender” are independent says that the multiplication rule applies to *all* outcomes.

## MODEL FOR EXAMINING INDEPENDENCE IN TWO-WAY

## TABLES

Select an SRS of size  $n$  from a population. Measure two categorical variables for each individual.

The null hypothesis is that the row and column variables are independent. The alternative hypothesis is that the row and column variables are dependent.

You can distinguish between the two models by examining the design of the study. In the independence model, there is a single sample. The column totals and row totals are random variables. The total sample size  $n$  is set by the researcher; the column and row sums are known only after the data are collected.

For the comparison-of-populations model, on the other hand, there is a sample from each of two or more populations. The column sums are the sample sizes selected at the design phase of the research.

The null hypothesis in both models says that there is no relationship between the column variable and the row variable. The precise statement of the hypothesis differs, depending on the sampling design. Fortunately, *the test of the hypothesis of “no relationship” is the same for both models*; it is the chi-square test. There are yet other statistical models for two-way tables that justify the chi-square test of the null hypothesis “no relation,” made precise in ways suitable for these models.

Statistical methods related to the chi-square test also allow the analysis of three-way and higher-way tables of count data. You can find a discussion of these topics in advanced texts on categorical data.<sup>5</sup>

### meta-analysis

## BEYOND THE BASICS

### Meta-analysis

Policymakers wanting to make decisions based on research are sometimes faced with the problem of summarizing the results of many studies. These studies may show effects of different magnitudes, some highly significant and some not significant. What *overall conclusion* can we draw?

**Meta-analysis** is a collection of statistical techniques designed to combine information from different but similar studies. Each individual study must be examined with care to ensure that its design and data quality are adequate. The basic idea is to compute a measure of the effect of interest for each study. These are then combined, usually by taking some sort of weighted average, to produce a summary measure for all of the studies. Of course, a confidence

interval for the summary is included in the results. Here is an example.

## Example

### 9.14 Do we eat too much salt?



← LOOK BACK

relative risk, p. 520

Evidence from a variety of sources suggests that diets high in salt are associated with risks to human health. To investigate the relationship between salt intake and stroke, information from 14 studies was combined in a meta-analysis.<sup>6</sup> Subjects were classified based on the amount of salt in their normal diet. They were followed for several years and then classified according to whether or not they had developed cardiovascular disease (CVD). A total of 104, 933 subjects were studied, and 5161 of them developed CVD. Here are the data from one of the studies:<sup>7</sup>

	Low salt	High salt
CVD	88	112
No CVD	1081	1134
Total	1169	1246

Let's look at the relative risk for this study. We first find the proportion of subjects who developed CVD in each group. For the subjects with a low salt intake the proportion who developed CVD is

$$881169=0.0753$$

or 75 per thousand; for the high-salt group, the proportion is

$$1121246=0.0899$$

or 90 per thousand. We can now compute the relative risk as the ratio of these two proportions. We choose to put the high-salt group in the numerator. The relative risk is

$$0.0899/0.0753=1.19$$

Relative risk greater than 1 means that the high-salt group developed more CVD than the low-salt group.

When the data from all 14 studies were combined, the relative risk was reported as 1.17 with a 95% confidence interval of (1.02, 1.32). Since this interval does not include the value 1, corresponding to equal proportions in the two groups, we conclude that the higher CVD rates are not the same for the two diets ( $P < 0.05$ ). The high-salt diet is associated with a 17% higher rate of CVD than the low-salt diet.

## USE YOUR KNOWLEDGE

### 9.14 A different view of the relative risk.

In the previous example, we computed the relative risk for the high-salt group relative to the low-salt group. Now, compute the relative risk for the low-salt group relative to the high-salt group by inverting the relative risk reported for the meta analysis in Example 9.14, that is, compute  $1/1.17$ . Then restate the last paragraph of the exercise with this change. (*Hint:* For the lower confidence limit, use 1 divided by the upper limit for the original ratio and do a similar calculation for the upper limit.)

## Section 9.1 Summary

The **null hypothesis** for  $r \times c$  tables of count data is that there is no relationship between the row variable and the column variable.

**Expected cell counts** under the null hypothesis are computed using the formula

$$\text{expected count} = \text{row total} \times \text{column total} / n$$

The null hypothesis is tested by the **chi-square statistic**, which compares the observed counts with the expected counts:

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Under the null hypothesis,  $X^2$  has approximately the  $\chi^2$  distribution with  $(r - 1)(c - 1)$  degrees of freedom. The  $P$ -value for the test is

$$P(\chi^2 \geq X^2)$$

where  $\chi^2$  is a random variable having the  $\chi^2(\text{df})$  distribution with  $\text{df} = (r - 1)(c - 1)$ .

The chi-square approximation is adequate for practical use when the average expected cell count is 5 or greater and all individual expected counts are 1 or greater, except in the case of  $2 \times 2$  tables. All four expected counts in a  $2 \times 2$  table should be 5 or greater.

For two-way tables we first compute percents or proportions that describe the relationship of interest. Then, we compute expected counts, the  $\chi^2$  statistic, and the  $P$ -value.

Two different models for generating  $r \times c$  tables lead to the chi-square test. In the first model, independent SRSs are drawn from each of  $c$  populations, and each observation is classified according to a categorical variable with  $r$  possible values. The null hypothesis is that the distributions of the row categorical variable are the same for all  $c$  populations. In the second model, a single SRS is drawn from a population, and observations are classified according to two categorical variables having  $r$  and  $c$  possible values. In this model,  $H^0$  states that the row and column variables are independent.

## 9.2 Goodness of Fit

When you complete this section, you will be able to

- Compute expected counts given a sample size and the probabilities specified by a null hypothesis for a chi-square goodness-of-fit test.
- Find the chi-square test statistic and its  $P$ -value.
- Interpret the results of a chi-square goodness-of-fit significance test.

In the last section, we discussed the use of the chi-square test to compare categorical-variable distributions of  $c$  populations. We now consider a slight variation on this scenario where we compare a sample from one population with a hypothesized distribution. Here is an example that illustrates the basic ideas.

### Example

#### 9.15 Sampling in the Adequate Calcium Today (ACT) study.

The ACT study was designed to examine relationships among bone growth patterns, bone development, and calcium intake. Participants were over 14,000 adolescents from six states: Arizona (AZ), California (CA), Hawaii (HI), Indiana (IN), Nevada (NV), and Ohio (OH). After the major goals of the study were completed, the investigators decided to do an additional analysis of the written comments made by the participants during the study. Because the number of participants was so large, a sampling plan was devised to select sheets containing the written comments of approximately 10% of the participants. A systematic sample (see page 204) of every tenth comment sheet was retrieved from each storage container for analysis.<sup>8</sup> Here are the counts for each of the six states:

Number of study participants in the sample						
AZ	CA	HI	IN	NV	OH	Total
167	257	257	297	107	482	1567

There were 1567 study participants in the sample. We will use the proportions of students from each of the states in the original sample of over 15,000

participants as the population values.<sup>9</sup> Here are the proportions:

Population proportions						
AZ	CA	HI	IN	NV	OH	Total
0.105	0.172	0.164	0.188	0.070	0.301	100.000

Let's see how well our sample reflects the state population proportions. We start by computing expected counts. Since 10.5% of the population is from Arizona, we expect the sample to have about 10.5% from Arizona. Therefore, since the sample has 1567 subjects, our expected count for Arizona is

$$\text{expected count for Arizona} = 0.105(1567) = 164.535$$

Here are the expected counts for all six states:

Expected counts						
AZ	CA	HI	IN	NV	OH	Total
164.54	269.52	256.99	294.60	109.69	471.67	1567.01

## USE YOUR KNOWLEDGE

### 9.15 Why is the sum 1567.01?



ACT

Refer to the table of expected counts in Example 9.15. Explain why the sum of the expected counts is 1567.01 and not 1567.

### 9.16 Calculate the expected counts.



ACT

Refer to Example 9.15. Find the expected counts for the other five states. Report your results with three places after the decimal as we did for Arizona.

As we saw with the expected counts in the analysis of two-way tables in Section 9.1, we do not really expect the observed counts to be *exactly* equal to the expected counts. Different samples under the same conditions would give different counts. We expect the average of these counts to be equal to the expected counts when the null hypothesis is true. How close do we think the counts and the expected counts should be?

We can think of our table of observed counts in Example 9.15 as a one-way table with six cells, each with a count of the number of subjects sampled from a particular state. Our question of interest is translated into a null hypothesis that says that the observed proportions of students in the six states can be viewed as random samples from the subjects in the ACT study. The alternative hypothesis is that the process generating the observed counts, a form of systematic sampling in this case, does not provide samples that are compatible with this hypothesis. In other words, the alternative hypothesis says that there is some bias in the way that we selected the subjects whose comments we will examine.

Our analysis of these data is very similar to the analyses of two-way tables that we studied in Section 9.1. We have already computed the expected counts. We now construct a chi-square statistic that measures how far the observed counts are from the expected counts. Here is a summary of the procedure:

## THE CHI-SQUARE GOODNESS-OF-FIT TEST

Data for  $n$  observations of a categorical variable with  $k$  possible outcomes are summarized as observed counts,  $n_1, n_2, \dots, n_k$ , in  $k$  cells. The null hypothesis specifies probabilities  $p_1, p_2, \dots, p_k$  for the possible outcomes. The alternative hypothesis says that the true probabilities of the possible outcomes are not the probabilities specified in the null hypothesis.

For each cell, multiply the total number of observations  $n$  by the specified probability to determine the expected counts:

$$\text{expected counts} = np_i$$

The **chi-square statistic** measures how much the observed cell counts differ from the expected cell counts. The formula for the statistic is

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

The degrees of freedom are  $k - 1$  and  $P$ -values are computed from the chi-square distribution.

Use this procedure when the expected counts are all 5 or more.

## Example

### 9.16 The goodness-of-fit test for the ACT study.

For Arizona, the observed count is 167. In Example 9.15, we calculated the expected count, 164.535. The contribution to the chi-square statistic for Arizona is

$$\frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} = \frac{(167 - 164.535)^2}{164.535} = 0.0369$$

We use the same approach to find the contributions to the chi-square statistic for the other five states. The expected counts are all at least 5, so we can proceed with the significance test.

The sum of these six values is the chi-square statistic,

$$\chi^2 = 0.93$$

The degrees of freedom are the number of cells minus 1:  $df = 6 - 1 = 5$ . We calculate the  $P$ -value using Table F or software. From Table F, we can determine  $P > 0.25$ . We conclude that the observed counts are compatible with the hypothesized proportions. The data do not provide any evidence that our systematic sample was biased with respect to selection of subjects from different states.

## USE YOUR KNOWLEDGE

### 9.17 Compute the chi-square statistic.



ACT

For each of the other five states, compute the contribution to the chi-square statistic using the method illustrated for Arizona in Example 9.16. Use the expected counts that you calculated in Exercise 9.16 for these calculations. Show that the sum of these values is the chi-square statistic.

## Example

### 9.17 The goodness-of-fit test from software.



ACT

Software output from Minitab and SPSS for this problem is given in Figure 9.6. Both report the  $P$ -value as 0.968. Note that the SPSS output includes a column titled “Residual.” For tables of counts, a residual for a cell is defined as

$$\text{residual} = \text{observed count} - \text{expected count} / \text{expected count}$$

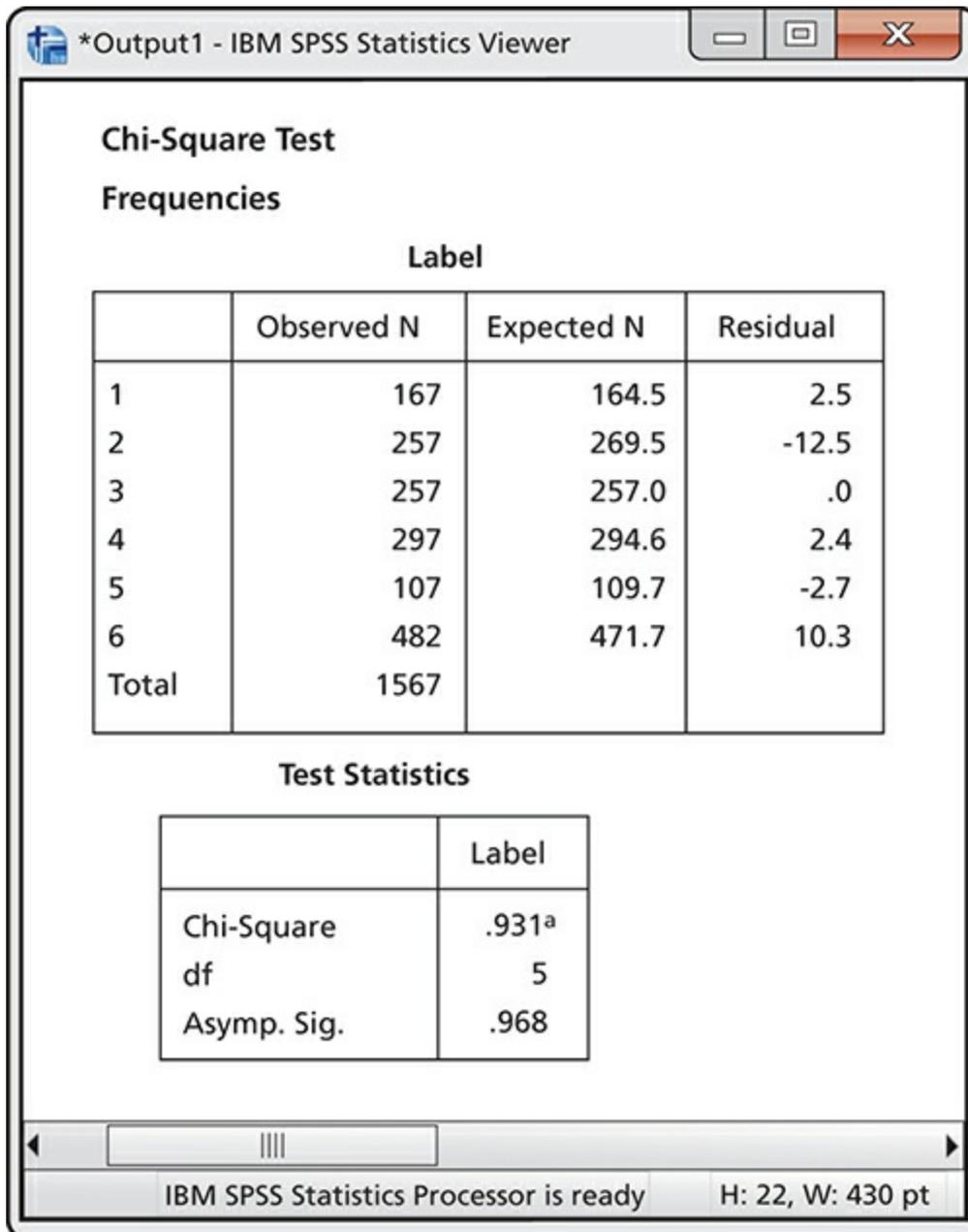
Note that the chi-square statistic is the sum of the squares of these residuals.

The image shows a screenshot of the Minitab software interface. The window title is "Minitab". The main content area displays the results of a "Chi-Square Goodness-of-Fit Test for Observed Counts in Variable: Count". The test uses category names in the state. The results are presented in a table with columns for Category, Observed, Test Proportion, Expected, and Contribution to Chi-Sq. Below this table, summary statistics are provided: N = 1567, DF = 5, Chi-Sq = 0.930840, and P-Value = 0.968. At the bottom of the window, there is a prompt to "Open a Minitab project file".

Category	Observed	Test Proportion	Expected	Contribution to Chi-Sq
AZ	167	0.105	164.535	0.036930
CA	257	0.172	269.524	0.581954
HI	257	0.164	256.988	0.000001
IN	297	0.188	294.596	0.019617
NV	107	0.070	109.690	0.065969
OH	482	0.301	471.667	0.226369

N	DF	Chi-Sq	P-Value
1567	5	0.930840	0.968



**FIGURE 9.6**

(a) Minitab and (b) SPSS output for Example 9.17.

Some software packages do not provide routines for computing the chi-square goodness-of-fit test. However, there is a very simple trick that can be used to produce the results from software that can analyze two-way tables. Make a two-way table in which the first column contains  $k$  cells with the observed counts. Add a second column with counts that correspond *exactly* to the probabilities specified by the null hypothesis, with a very large number of observations. Then perform the chi-square significance test for two-way tables.

## USE YOUR KNOWLEDGE

### 9.18 Distribution of M&M colors.



M&M Mars Company has varied the mix of colors for M&M'S Plain Chocolate Candies over the years. These changes in color blends are the result of consumer preference tests. Most recently, the color distribution is reported to be 13% brown, 14% yellow, 13% red, 20% orange, 24% blue, and 16% green.<sup>10</sup> You open up a 14-ounce bag of M&M'S and find 61 brown, 59 yellow, 49 red, 77 orange, 141 blue, and 88 green. Use a goodness-of-fit test to examine how well this bag fits the percents stated by the M&M Mars Company.

## Example

### 9.18 The sign test as a goodness-of-fit test.

In Example 7.12 (page 439) we used a sign test to examine the effect of the full moon on aggressive behaviors of dementia patients. The study included 15 patients, 14 of whom exhibited a greater number of aggressive behaviors on moon days than on other days. The sign test tests the null hypothesis that patients are equally likely to exhibit more aggressive behaviors on moon days than on other days. Since  $n = 15$ , the sample proportion is  $\hat{p} = 14/15$  and the null hypothesis is  $H^0: p = 0.5$ .

To look at these data from the viewpoint of goodness of fit, we think of the data as two counts: patients who had a greater number of aggressive behaviors on moon days and patients who had a greater number of aggressive behaviors on other days.

Counts		
Moon	Other	Total
14	1	15

If the two outcomes are equally likely, the expected counts are both 7.5 ( $15 \times 0.5$ ). The expected counts are both greater than 5, so we can proceed with the

significance test.

The test statistic is

$$\begin{aligned}X^2 &= (14 - 7.5)27.5 + (1 - 7.5)27.5 \\ &= 5.633 + 5.633 \\ &= 11.27\end{aligned}$$

We have  $k = 2$ , so the degrees of freedom are 1. From Table F we conclude that  $P < 0.001$ .

In Example 7.12, we tested the null hypothesis versus the one-sided alternative that there was a “moon effect.” Within the framework of the goodness-of-fit test, we test only the general alternative hypothesis that the distribution of the counts does not follow the specified probabilities. Note that the  $P$ -value in Example 7.12 was calculated using the binomial distribution. The value was 0.000488, approximately one-half of the value that we reported from Table F in Example 9.18.

## USE YOUR KNOWLEDGE

### 9.19 Is the coin fair?

In Example 4.3 (page 234) we learned that the South African statistician John Kerrich tossed a coin 10, 000 times while imprisoned by the Germans during World War II. The coin came up heads 5067 times.

- Formulate the question about whether or not the coin was fair as a goodness-of-fit hypothesis.
- Perform the chi-square significance test and write a short summary of the results.

## Section 9.2 Summary

The **chi-square goodness-of-fit test** is used to compare the sample distribution of a categorical variable from a population with a hypothesized distribution. The data for  $n$  observations with  $k$  possible outcomes are summarized as observed counts,  $n_1, n_2, \dots, n_k$ , in  $k$  cells. The **null hypothesis** specifies probabilities for the possible outcomes.

The analysis of these data is similar to the analyses of two-way tables discussed in Section 9.1. For each cell, the **expected count** is determined by multiplying the total number of observations  $n$  by the specified probability  $p_j$ . The null hypothesis

is tested by the usual **chi-square statistic**, which compares the observed counts,  $n_i$ , with the expected counts. Under the null hypothesis,  $X^2$  has approximately the  $\chi^2$  distribution with  $df = k - 1$ .

## CHAPTER 9 Exercises

*For Exercise 9.1 to 9.3, see page 535–536; for Exercise 9.4 to 9.6, see page 542–543; for Exercise 9.7 to 9.10, see page 544; for Exercise 9.11 to 9.13, see page 545; for Exercise 9.14, see page 550; for Exercise 9.15 and 9.16, see page 552; for Exercise 9.17, see page 553; for Exercise 9.18, see page 555; and for Exercise 9.19, see page 556.*

### 9.20 Translate each problem into a $2 \times 2$ table

In each of the following scenarios, translate the problem into one that can be analyzed using a  $2 \times 2$  table.

(a) Two website designs are being compared. Fifty students have agreed to be subjects for the study, and they are randomly assigned to watch one of the designs for as long as they like. For each student the study directors record whether or not the website is watched for more than a minute. For the first design, 12 students watched for more than a minute; for the second, 5 watched for more than a minute.

(b) Samples of first-year students and fourth-year students were asked if they were in favor of a new proposed core curriculum. Among the first-year students, 85 said “Yes” and 276 said “No.” For the fourth-year students, 117 said “Yes” and 104 said “No.”

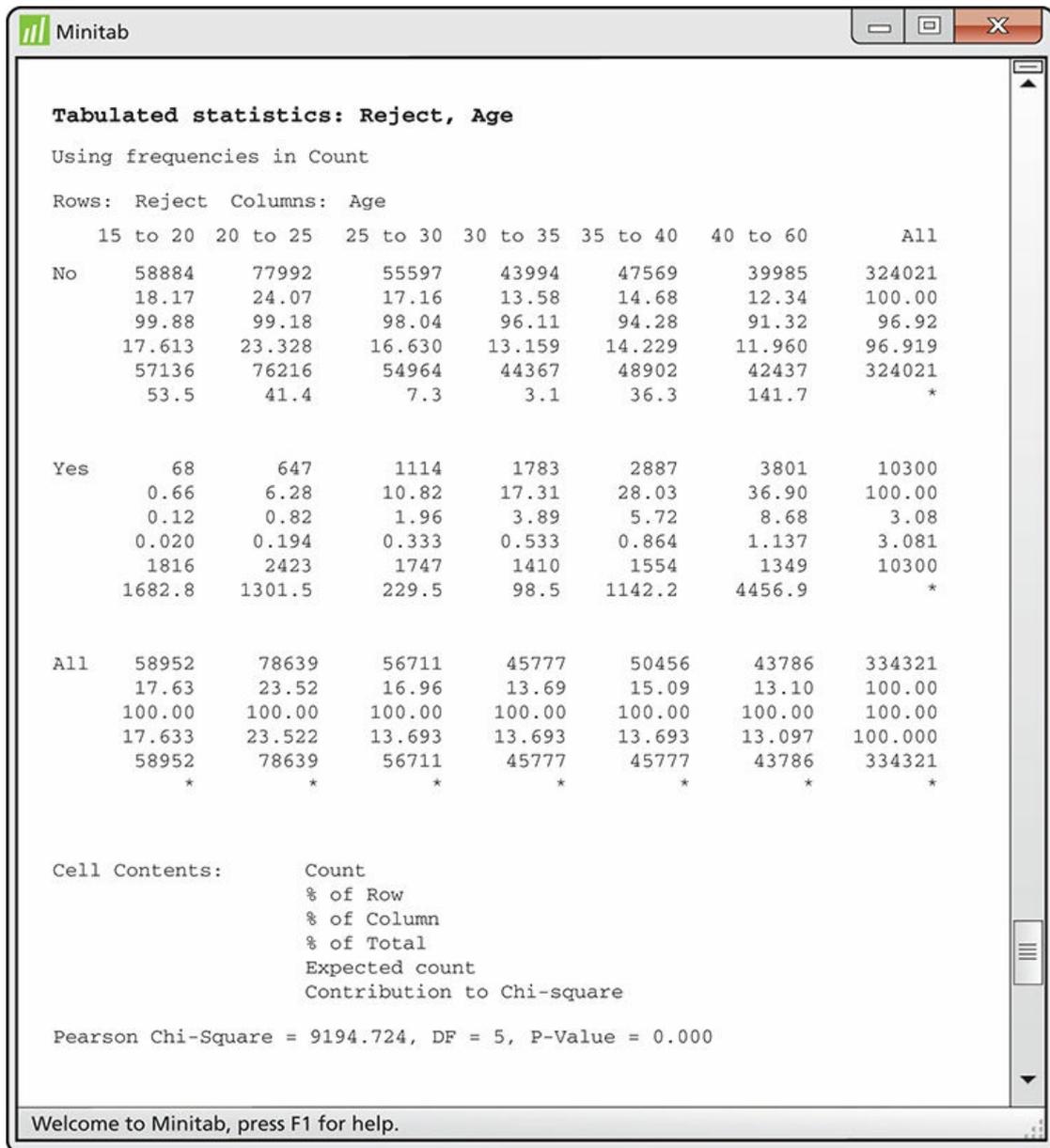
### 9.21 Find the joint distribution, the marginal distributions, and the conditional distributions

Refer to the previous exercise. For each scenario, identify the joint distribution, the marginal distributions, and the conditional distributions.

### 9.22 Read the output

Exercise 8.58 (page 523) gives data on individuals rejected for military service in the Cuban War of Independence in 1898 because they did not have enough teeth. In that exercise you compared the rejection rate for those under the age of 20 with the rejection rate for those over 40. Figure 9.7 gives software output for the table that classifies the recruits into six age categories. Use the output to find the joint distribution, the marginal distributions, and the conditional distributions for these data. 

**TEETH**



**FIGURE 9.7**  
Computer output for Exercise 9.22.

### 9.23 Relationship or explanatory and response variables?

In each of the following scenarios, determine whether the goal is to describe the relationship between an explanatory variable and a response variable or to simply describe the relationship between two categorical variables. There may not always be a clear correct answer, but you need to give reasons for the answer you choose. If there are explanatory and response variables, identify them.

- (a) A large sample of undergraduates is classified by major and year of study.
- (b) Equal-sized samples of first-year, second-year, third-year, and fourth-year undergraduates are selected. Each student is asked “Do you eat five or more servings of fruits or vegetables per day?”
- (c) Television programs are classified as low, medium, or high for violence content and by morning, afternoon, prime time, or late night for the time of day that they are broadcast.

(d) The setting of Exercise 9.22, which examines age and rejection rate for military recruits.

## 9.24 Choose the appropriate conditional distributions

Refer to the previous exercise. For each scenario, choose which conditional distribution you would use to describe the data. Give reasons for your answers.

## 9.25 Sexual harassment in middle and high schools

A nationally representative survey of students in grades 7 to 12 asked about the experience of these students with respect to sexual harassment.<sup>11</sup> One question asked how many times the student had witnessed sexual harassment in school. Here are the data categorized by gender:  HARAS1

Gender	Times witnessed		
	Never	Once	More than once
Girls	140	192	671
Boys	106	125	732

Find the expected counts for this  $2 \times 3$  table.

## 9.26 Do the significance test

Refer to the previous exercise. Compute the chi-square statistic and the  $P$ -value. Write a short summary of your conclusions from the analysis of these data.  HARAS1

## 9.27 Sexual harassment online or in person

In the study described in Exercise 9.25, the students were also asked whether or not they were harassed in person and whether or not they were harassed online. Here are the data for the girls:  HARASG

Harassed in person	Harassed online	
	Yes	No
Yes	321	200
No	40	441

(a) Analyze these data using the method presented in Chapter 8 for comparing two proportions (page 508).

(b) Analyze these data using the method presented in this chapter for examining a relationship between two categorical variables in a  $2 \times 2$  table.

(c) Use this example to explain the relationship between the chi-square test and the  $z$  test for comparing two proportions.

(d) The number of girls reported in this exercise is not the same as the number reported for Exercise 9.25. Suggest a possible reason for this difference.

## 9.28 Data for the boys

Refer to the previous exercise. Here are the corresponding data for boys:  HARASB

Harassed online		
Harassed in person	Yes	No
Yes	183	154
No	48	578

Using these data, repeat the analyses that you performed for the girls in Exercise 9.27. How do the results for the boys differ from those that you found for girls?

## 9.29 Repeat your analysis

In part (a) of Exercise 9.27, you had to decide which variable was explanatory and which variable was response when you computed the proportions to be compared.

- Did you use harassed online or harassed in person as the explanatory variable? Explain the reasons for your choice.
- Repeat the analysis that you performed in Exercise 9.27 with the other choice for the explanatory variable.
- Summarize what you have learned from comparing the results of using the different choices for analyzing these data.

## 9.30 Which model?

Refer to the four scenarios in Exercise 9.23. For each, determine whether the model corresponds to the comparison of several populations or to the test of independence. Give reasons for your answers.

## 9.31 Is the die fair?

You suspect that a die has been altered so that the outcomes of a roll, the numbers 1 to 6, are not equally likely. You toss the die 600 times and obtain the following results:  DIE

Outcome	1	2	3	4	5	6
Count	89	82	123	115	100	91

Compute the expected counts that you would need to use in a goodness-of-fit test for these data.

## 9.32 Perform the significance test

Refer to the previous exercise. Find the chi-square test statistic and its  $P$ -value and write a short summary of your conclusions.

## 9.33 The value of online courses

A Pew Internet survey asked college presidents whether or not they believed that online courses

offer an equal educational value when compared with courses taken in the classroom. The presidents were classified by the type of educational institution. Here are the data:<sup>12</sup>  ONLINE

Response	Institution type			
	4-year private	4-year public	2-year private	For profit
Yes	36	50	66	54
No	62	48	34	45

(a) Discuss different ways to plot the data. Choose one way to make a plot and give reasons for your choice.

(b) Make the plot and describe what it shows.

### 9.34 Do the answers depend upon institution type?

Refer to the previous exercise. You want to examine whether or not the data provide evidence that the belief that online and classroom courses offer equal educational value varies with the type of institution of the president.  ONLINE

(a) Formulate this question in terms of appropriate null and alternative hypotheses.

(b) Perform the significance test. Report the test statistic, the degrees of freedom, and the  $P$ -value.

(c) Write a short summary explaining the results.

### 9.35 Compare the college presidents with the general public

Refer to Exercise 9.33. Another Pew Internet survey asked the general public about their opinions on the value of online courses. Of the 2142 people who participated in the survey, 621 responded “Yes” to the question “Do you believe that online courses offer an equal educational value when compared with courses taken in the classroom?”  ONLINE

(a) Use the data given in Exercise 9.33 to find the number of college presidents who responded “Yes” to the question.

(b) Construct a two-way table that you can use to compare the responses of the general public with the responses of the college presidents.

(c) Is it meaningful to interpret the marginal totals or percents for this table? Explain your answer.

(d) Analyze the data in your two-way table and summarize the results.

### 9.36 Remote deposit capture

The Federal Reserve has called remote deposit capture (RDC) “the most important development the [U.S.] banking industry has seen in years.” This service allows users to scan checks and to transmit the scanned images to a bank for posting.<sup>13</sup> In its annual survey of community banks, the American Bankers Association asked banks whether or not they offered this service.<sup>14</sup> Here are the results classified by the asset size (in millions of dollars) of the bank:  RDCA

Offer RDC

Asset size	Yes	No
Under \$100	63	309
\$101-\$200	59	132
\$201 or more	112	85

- (a) Summarize the results of this survey question numerically and graphically.
- (b) Test the null hypothesis that there is no association between the size of a bank, measured by assets, and whether or not they offer RDC. Report the test statistic, the  $P$ -value, and your conclusion.

### 9.37 Health care fraud

Most errors in billing insurance providers for health care services involve honest mistakes by patients, physicians, or others involved in the health care system. However, fraud is a serious problem. The National Health Care Anti-fraud Association estimates that approximately \$68 billion is lost to health care fraud each year.<sup>15</sup> When fraud is suspected, an audit of randomly selected billings is often conducted. The selected claims are then reviewed by experts, and each claim is classified as allowed or not allowed. The distributions of the amounts of claims are frequently highly skewed, with a large number of small claims and a small number of large claims. Since simple random sampling would likely be overwhelmed by small claims and would tend to miss the large claims, stratification is often used. See the section on stratified sampling in Chapter 3 (page 196). Here are data from an audit that used three strata based on the sizes of the claims (small, medium, and large):<sup>16</sup>  **BILLER**

Stratum	Sampled claims	Number not allowed
Small	57	6
Medium	17	5
Large	5	1

- (a) Construct the  $3 \times 2$  table of counts for these data that includes the marginal totals.
- (b) Find the percent of claims that were not allowed in each of the three strata.
- (c) To perform a significance test, combine the medium and large strata. Explain why we do this.
- (d) State an appropriate null hypothesis to be tested for these data.
- (e) Perform the significance test and report your test statistic with degrees of freedom and the  $P$ -value. State your conclusion.

### 9.38 Population estimates

Refer to the previous exercise. One reason to do an audit such as this is to estimate the number of claims that would not be allowed if all claims in a population were examined by experts. We have an estimate of the proportion of unallowed claims from each stratum based on our sample. We know the corresponding population proportion for each stratum. Therefore, if we take the sample proportions of unallowed claims and multiply by the population sizes, we would have the estimates that we need. Here are the population sizes for the three strata:

### Stratum    Claims in strata

Small	3342
Medium	246
Large	58

(a) For each stratum, estimate the total number of claims that would not be allowed if all claims in the strata had been audited.

(b) Give margins of error for your estimates. (*Hint:* You first need to find standard errors for your sample estimates using material presented in Chapter 8 (page 490). Then you need to use the rules for variances from Chapter 4 (page 275) to find the standard errors for the population estimates. Finally, you need to multiply by  $z^*$  to determine the margins of error.)

## 9.39 DFW rates

One measure of student success for colleges and universities is the percent of admitted students who graduate. Studies indicate that a key issue in retaining students is their performance in so-called gateway courses. These are courses that serve as prerequisites for other key courses that are essential for student success. One measure of student performance in these courses is the DFW rate, the percent of students who receive grades of D, F, or W (withdraw). A major project was undertaken to improve the DFW rate in a gateway course at a large midwestern university. The course curriculum was revised to make it more relevant to the majors of the students taking the course, a small group of excellent teachers taught the course, technology (including clickers and online homework) was introduced, and student support outside the classroom was increased. The following table gives data on the DFW rates for the course over three years.<sup>17</sup> In Year 1, the traditional course was given; in Year 2, a few changes were introduced; and in Year 3, the course was substantially revised.

Year	DFW rate	Number of students taking course
Year 1	42.3%	2408
Year 2	24.9%	2325
Year 3	19.9%	2126

Do you think that the changes in this gateway course had an impact on the DFW rate? Write a report giving your answer to this question. Support your answer by an analysis of the data.  LIE

## 9.40 Lying to a teacher

One of the questions in a survey of high school students asked about lying to teachers.<sup>18</sup> The following table gives the numbers of students who said that they lied to a teacher at least once during the past year, classified by gender.

Lied at least once	Gender	
	Male	Female
Yes	3,228	10,295
No	9,659	4,620

(a) Add the marginal totals to the table.

(b) Calculate appropriate percents to describe the results of this question.

(c) Summarize your findings in a short paragraph.

(d) Test the null hypothesis that there is no association between gender and lying to teachers. Give the test statistic and the  $P$ -value (with a sketch similar to the one on page 539) and summarize your conclusion. Be sure to include numerical and graphical summaries.

### 9.41 When do Canadian students enter private career colleges?

A survey of 13,364 Canadian students who enrolled in private career colleges was conducted to understand student participation in the private postsecondary educational system.<sup>19</sup> In one part of the survey, students were asked about their field of study and about when they entered college. Here are the results:  CANF

Field of study	Number of students	Time of Entry	
		Right after high school	Later
Trades	942	34%	66%
Design	584	47%	53%
Health	5085	40%	60%
Media/IT	3148	31%	69%
Service	1350	36%	64%
Other	2255	52%	48%

In this table, the second column gives the number of students in each field of study. The next two columns give the marginal distribution of time of entry for each field of study.

- Use the data provided to make the  $6 \times 2$  table of counts for this problem.
- Analyze the data.
- Write a summary of your conclusions. Be sure to include the results of your significance testing as well as a graphical summary.

### 9.42 Government loans for Canadian students in private career colleges

Refer to the previous exercise. The survey also asked about how these college students paid for their education. A major source of funding was government loans. Here are the survey percents of Canadian private students who use government loans to finance their education by field of study:

 CANGOV

Field of study	Number of students	Percent using government loans
Trades	942	45%
Design	599	53%
Health	5234	55%
Media/IT	3238	55%
Service	1378	60%
Other	2300	47%

- Construct the  $6 \times 2$  table of counts for this exercise.
- Test the null hypothesis that the percent of students using government loans to finance their education does not vary with field of study. Be sure to provide all the details of your significance test.

(c) Summarize your analysis and conclusions. Be sure to include a graphical summary.

(d) The number of students reported in this exercise is not the same as the number reported in Exercise 9.41. Suggest a possible reason for this difference.

### 9.43 Other funding for Canadian students in private career colleges

Refer to the previous exercise. Another major source of funding was parents, family, or spouse. The following table gives the survey percents of Canadian private students who rely on these sources to finance their education by field of study. 

Field of study	Number of students	Percent using parents/family/spouse
Trades	942	20%
Design	599	37%
Health	5234	26%
Media/IT	3238	16%
Service	1378	18%
Other	2300	41%

Answer the questions in the previous exercise for these data.

### 9.44 Why not use a chi-square test?

As part of the study on ongoing fright symptoms due to exposure to horror movies at a young age, the following table was created based on the written responses from 119 students. Explain why a chi-square test is not appropriate for this table.

Movie or video	Percent of students who reported each problem			
	Type of Problem			
	Bedtime		Waking	
	Short term	Enduring	Short term	Enduring
<i>Poltergeist</i> (n 29)	68	7	64	32
<i>Jaws</i> (n 23)	39	4	83	43
<i>Nightmare on Elm Street</i> (n 16)	69	13	37	31
<i>Thriller</i> (music video) (n 16)	40	0	27	7
<i>It</i> (n 24)	64	0	64	50
<i>The Wizard of Oz</i> (n 12)	75	17	50	8
<i>E.T.</i> (n 11)	55	0	64	27

### 9.45 Waking versus bedtime symptoms

As part of the study on ongoing fright symptoms due to exposure to horror movies at a young age, the following table was presented to describe the lasting impact these movies have had during bedtime and waking life: 

Bedtime symptoms	Waking symptoms	
	Yes	No
Yes	36	33

- (a) What percent of the students have lasting waking-life symptoms?
- (b) What percent of the students have both waking-life and bedtime symptoms?
- (c) Test whether there is an association between waking-life and bedtime symptoms. State the null and alternative hypotheses, the  $\chi^2$  statistic, and the  $P$ -value.

### 9.46 Construct a table with no association

Construct a  $3 \times 3$  table of counts where there is no apparent association between the row and column variables.

### 9.47 Can you construct the joint distribution from the marginal distributions?

Here are the row and column totals for a two-way table with two rows and two columns:

$a$	$b$	150
$c$	$d$	150
100	200	300

Find *two different* sets of counts  $a$ ,  $b$ ,  $c$ , and  $d$  for the body of the table. This demonstrates that the relationship between two variables cannot be obtained solely from the two marginal distributions of the variables.

### 9.48 Which model?

Refer to Exercises 9.37, 9.39, 9.40, 9.42, and 9.45. For each, state whether you are comparing two or more populations (the first model for two-way tables) or testing independence between two categorical variables (the second model).

### 9.49 Are Mexican Americans less likely to be selected as jurors?

Refer to Exercise 8.99 (page 528) concerning *Castaneda v. Partida*, the case where the Supreme Court review used the phrase “two or three standard deviations” as a criterion for statistical significance. Recall that there were 181, 535 persons eligible for jury duty, of whom 143, 611 were Mexican Americans. Of the 870 people selected for jury duty, 339 were Mexican Americans. We are interested in finding out if there is an association between being a Mexican American and being selected as a juror. Formulate this problem using a two-way table of counts. Construct the  $2 \times 2$  table using the variables Mexican American or not and juror or not. Find the  $\chi^2$  statistic and its  $P$ -value. Square the  $z$  statistic that you obtained in Exercise 8.99 and verify that the result is equal to the  $\chi^2$  statistic.

### 9.50 Goodness of fit to a standard Normal distribution

Computer software generated 500 random numbers that should look as if they are from the standard Normal distribution. They are categorized into five groups: (1) less than or equal to  $-0.6$ ; (2) greater

than  $-0.6$  and less than or equal to  $-0.1$ ; (3) greater than  $-0.1$  and less than or equal to  $0.1$ ; (4) greater than  $0.1$  and less than or equal to  $0.6$ ; and (5) greater than  $0.6$ . The counts in the five groups are 139, 102, 41, 78, and 140, respectively. Find the probabilities for these five intervals using Table A. Then compute the expected number for each interval for a sample of 500. Finally, perform the goodness-of-fit test and summarize your results.

### 9.51 More on the goodness of fit to a standard Normal distribution

Refer to the previous exercise. Use software to generate your own sample of 500 standard Normal random variables, and perform the goodness-of-fit test. Choose a different set of intervals than the ones used in the previous exercise.

### 9.52 Goodness of fit to the uniform distribution

Computer software generated 500 random numbers that should look as if they are from the uniform distribution on the interval 0 to 1 (see page 74). They are categorized into five groups: (1) less than or equal to 0.2; (2) greater than 0.2 and less than or equal to 0.4; (3) greater than 0.4 and less than or equal to 0.6; (4) greater than 0.6 and less than or equal to 0.8; and (5) greater than 0.8. The counts in the five groups are 114, 92, 108, 101, and 85, respectively. The probabilities for these five intervals are all the same. What is this probability? Compute the expected number for each interval for a sample of 500. Finally, perform the goodness-of-fit test and summarize your results.

### 9.53 More on goodness of fit to the uniform distribution

Refer to the previous exercise. Use software to generate your own sample of 800 uniform random variables on the interval from 0 to 1, and perform the goodness-of-fit test. Choose a different set of intervals than the ones used in the previous exercise.

### 9.54 Suspicious results?

An instructor who assigned an exercise similar to the one described in the previous exercise received homework from a student who reported a  $P$ -value of 0.999. The instructor suspected that the student did not use the computer for the assignment but just made up some numbers for the homework. Why was the instructor suspicious? How would this scenario change if there were 2000 students in the class?

### 9.55 Is there a random distribution of trees?

In Example 6.1 (page 352) we examined data concerning the longleaf pine trees in the Wade Tract and concluded that the distribution of trees in the tract was not random. Here is another way to examine the same question. First, we divide the tract into four equal parts, or quadrants, in the east–west direction. Call the four parts  $Q_1$  to  $Q_4$ . Then we take a random sample of 100 trees and count

the number of trees in each quadrant. Here are the data:  TREEQ

Quadrant	Q1	Q2	Q3	Q4
Count	18	22	39	21

(a) If the trees are randomly distributed, we expect to find 25 trees in each quadrant. Why? Explain your answer.

- (b) We do not really expect to get *exactly* 25 trees in each quadrant. Why? Explain your answer.
- (c) Perform the goodness-of-fit test for these data to determine if these trees are randomly scattered. Write a short report giving the details of your analysis and your conclusion.

# 10 Inference for Regression

CHAPTER



## 10.1 Simple Linear Regression

## 10.2 More Detail about Simple Linear Regression

## Introduction

In this chapter we continue our study of relationships between variables and describe methods for inference when there is a single quantitative response variable and a single quantitative explanatory variable. The descriptive tools we learned in Chapter 2—scatterplots, least-squares regression, and correlation—are essential preliminaries to inference and also provide a foundation for confidence intervals and significance tests.

We first met the sample mean  $\bar{x}$  in Chapter 1 as a measure of the center of a collection of observations. Later we learned that when the data are a random sample from a population, the sample mean is an estimate of the population mean  $\mu$ . In Chapters 6 and 7, we used  $\bar{x}$  as the basis for confidence intervals and significance tests for inference about  $\mu$ .

Now we will follow the same approach for the problem of fitting straight lines to data. In Chapter 2 we met the least-squares regression line  $\hat{y} = b_0 + b_1x$  as a description of a straight-line relationship between a response variable  $y$  and an explanatory variable  $x$ . At that point we did not distinguish between sample and population. Now we will think of the least-squares line computed from a sample as an estimate of a *true* regression line for the population.

Following the common practice of using Greek letters for population parameters, we will write the population line as  $\beta_0 + \beta_1x$ . This notation reminds us that the intercept of the fitted line  $b_0$  estimates the intercept of the population line  $\beta_0$ , and the fitted slope  $b_1$  estimates the slope of the population line  $\beta_1$ .

The methods detailed in this chapter will help us answer questions such as

- Is the trend in the annual number of tornadoes reported in the United States approximately linear? If so, what is the average yearly increase in the number of tornadoes? How many are predicted for next year?
- What is the relationship between a female college student's body mass index and physical activity level measured by a pedometer?
- Among North American universities, is there a strong negative correlation between the binge-drinking rate and the average price for a bottle of beer at establishments within a two-mile radius of campus?

## 10.1 Simple Linear Regression

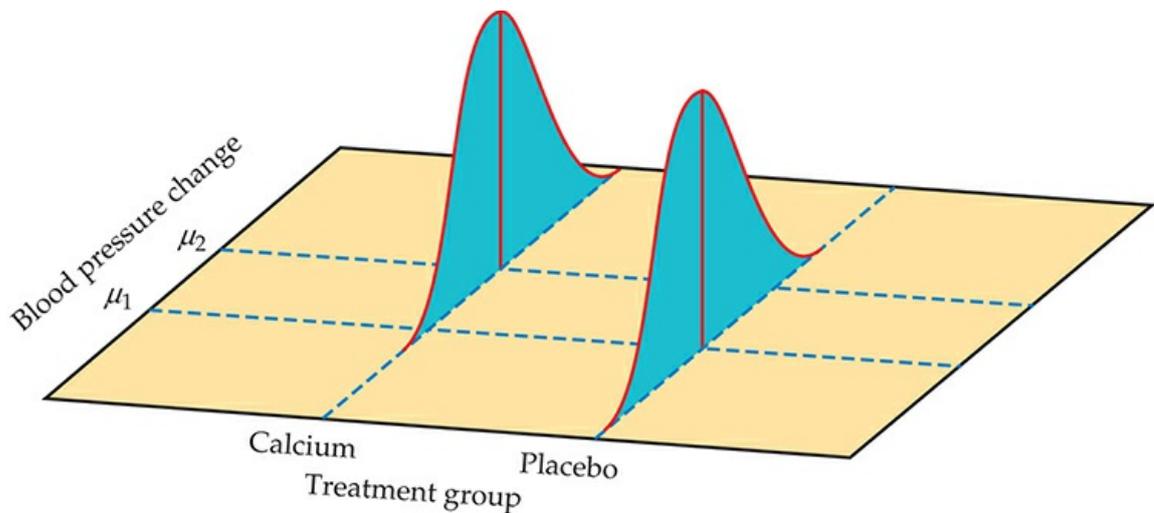
When you complete this section, you will be able to

- Describe the simple linear regression model in terms of a population regression line and the deviations of the response variable  $y$  from this line.
- Interpret linear regression output from statistical software to obtain the least-squares regression line and model standard deviation.
- Distinguish the model deviations  $\varepsilon_i$  from the residuals  $e_i$  that are obtained from a least-squares fit to a data set.
- Use diagnostic plots to check the assumptions of the simple linear regression model.
- Construct and interpret a level  $C$  confidence interval for the population intercept and for the population slope.
- Perform a level  $\alpha$  significance test for the population intercept and for the population slope.
- Construct and interpret a level  $C$  confidence interval for a mean response and a level  $C$  prediction interval for a future observation when  $x = x^*$ .

### Statistical model for linear regression

Simple linear regression studies the relationship between a response variable  $y$  and a single explanatory variable  $x$ . We expect that different values of  $x$  will produce different mean responses for  $y$ . We encountered a similar but simpler situation in Chapter 7 when we discussed methods for comparing two population means. Figure 10.1 illustrates the statistical model for a comparison of blood pressure change in two groups of experimental subjects, one group taking a calcium supplement and the other a placebo. We can think of the treatment (placebo or calcium) as the explanatory variable in this example. This model has two important parts:

- The mean change in blood pressure may be different in the two populations. These means are labeled  $\mu_1$  and  $\mu_2$  in Figure 10.1.
- Individual changes vary within each population according to a Normal distribution. The two Normal curves in Figure 10.1 describe these responses. These Normal distributions have the same spread, indicating that the population standard deviations are assumed to be equal.



**FIGURE 10.1**

The statistical model for comparing responses to two treatments; the mean response varies with the treatment.

In linear regression the explanatory variable  $x$  is quantitative and can have many different values. Imagine, for example, giving different amounts of calcium  $x$  to different groups of subjects. We can think of the values of  $x$  as defining different **subpopulations**, one for each possible value of  $x$ . Each subpopulation consists of all individuals in the population having the same value of  $x$ . If we conducted an experiment with five different amounts of calcium, we could view these values as defining five different subpopulations.

### subpopulations

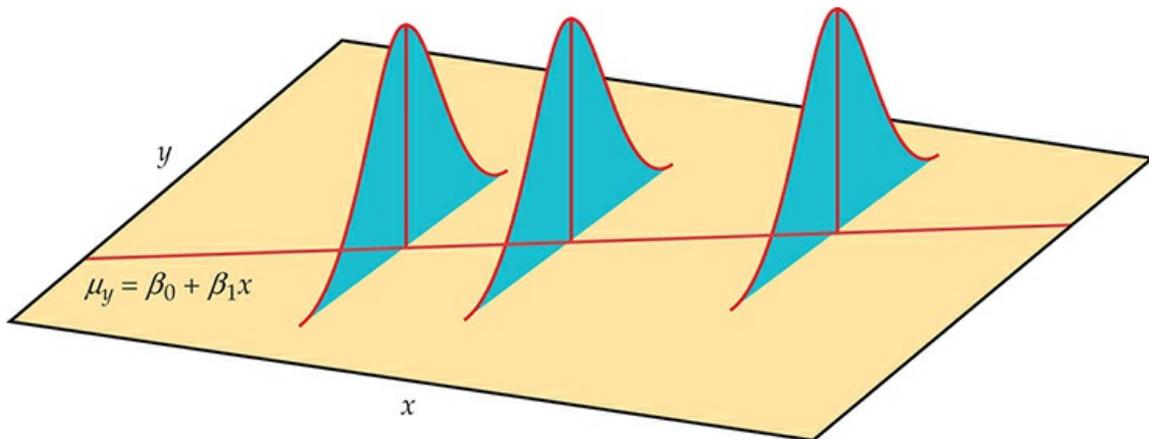
The statistical model for simple linear regression also assumes that for each value of  $x$ , the observed values of the response variable  $y$  are Normally distributed with a mean that depends on  $x$ . We use  $\mu_y$  to represent these means. In general, the means  $\mu_y$  can change as  $x$  changes according to any sort of pattern. In **simple linear regression** we assume that the means all lie on a line when plotted against  $x$ . To summarize, this model also has two important parts:

### simple linear regression

- The mean of the response variable  $y$  changes as  $x$  changes. The means all lie on a straight line. That is,  $\mu_y = \beta_0 + \beta_1 x$ .
- Individual responses  $y$  with the same  $x$  vary according to a Normal distribution. This variation, measured by the standard deviation  $\sigma$  is the same for all values of  $x$ .

This statistical model is pictured in Figure 10.2. The line describes how the mean response  $\mu_y$  changes with  $x$ . This is the **population regression line**. The three Normal curves show how the response  $y$  will vary for three different values of the explanatory variable  $x$ .

population regression line



**FIGURE 10.2**

The statistical model for linear regression; the mean response is a straight-line function of the explanatory variable.

## Data for simple linear regression



The data for a linear regression are observed values of  $y$  and  $x$ . The model takes each  $x$  to be a known quantity. In practice,  $x$  may not be exactly known. *If the error in measuring  $x$  is large, more advanced inference methods are needed.* The response  $y$  for a given  $x$  is a random variable. The linear regression model describes the mean and standard deviation of this random variable  $y$ . These unknown parameters must be estimated from the data.

We will use the following example to explain the fundamentals of simple linear regression. Because regression calculations in practice are always done by statistical software, we will rely on computer output for the arithmetic. In Section 10.2, we give an example that illustrates how to do the work with a calculator if software is unavailable.

### Example

#### 10.1 Relationship between BMI and physical activity.



Decrease in physical activity is considered to be a major contributor to the increase in prevalence of overweight and obesity in the general adult population. Because the prevalence of physical inactivity among college students is similar to that of the adult population, many researchers feel that a clearer understanding of college students' physical activity behaviors is needed to develop early interventions. As part of one study, researchers looked at the relationship between physical activity (PA) measured with a pedometer and body mass index (BMI).<sup>1</sup> Each participant wore a pedometer for a week, and the average number of steps taken per day (in thousands) was recorded. Various body composition variables, including BMI (in kilograms per square meter,  $\text{kg}/\text{m}^2$ ), were also measured. We will consider a sample of 100 female undergraduates.

Before starting our analysis, it is appropriate to consider the extent to which the results can reasonably be generalized. In the original study, undergraduate volunteers were obtained at a large southeastern public university through classroom announcements and campus flyers.



*The potential for bias should always be considered when obtaining volunteers.* In this case, the participants were screened, and those with severe health issues, as well as varsity athletes, were excluded. As a result, the researchers considered these volunteers as an SRS from the population of undergraduates at this university. However, they acknowledged the limitations of their study, stating that similar investigations at universities of different sizes and in other climates of the United States are needed.

In the statistical model for predicting BMI from physical activity, subpopulations are defined by the explanatory variable, physical activity. We could think about sampling women from this university, each averaging the same number of steps per day—say, 9000. Variation in genetic makeup, lifestyle, and diet would be sources of variation that would result in different values of BMI for this subpopulation.

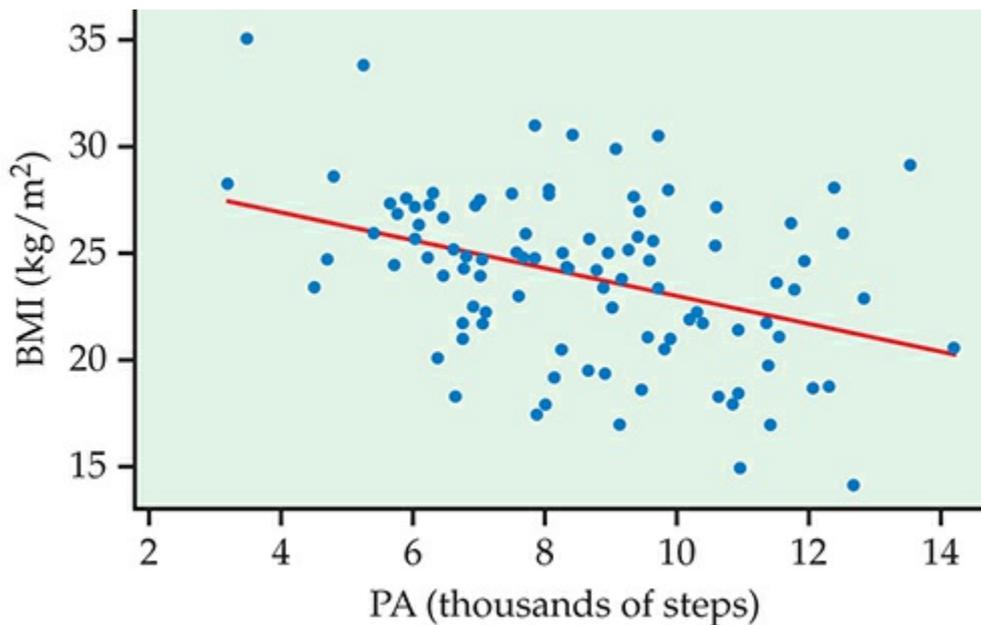
## Example

### 10.2 Graphical display of BMI and physical activity.



scatterplot, p. 88

We start our analysis with a scatterplot of the data. Figure 10.3 is a plot of BMI versus physical activity for our sample of 100 participants. We use the variable names BMI and PA. The least-squares regression line is also shown in the plot. There is a negative association between BMI and PA that appears approximately linear. There is also a considerable amount of scatter about this least-squares regression line.



**FIGURE 10.3**

Scatterplot of BMI versus physical activity (PA) with the least-squares line, for Example 10.2.



*Always start with a graphical display of the data.* There is no point in fitting a linear model if the relationship does not, at least approximately, appear linear. Now that we have confirmed an approximate linear relationship, we return to predicting BMI for different subpopulations, defined by the explanatory variable physical activity.

Our statistical model assumes that the BMI values are Normally distributed with a mean  $\mu_y$  that depends upon  $x$  in a linear way. Specifically,

$$\mu_y = \beta_0 + \beta_1 x$$

This population regression line gives the average BMI for all values of  $x$ . We cannot observe this line because the observed responses  $y$  vary about their means.

The statistical model for linear regression consists of the population regression line and a description of the variation of  $y$  about the line. This was displayed in Figure 10.2 with the line and the three Normal curves. The following equation expresses this idea:

$$\text{DATA} = \text{FIT} + \text{RESIDUAL}$$

The FIT part of the model consists of the subpopulation means, given by the expression  $\beta_0 + \beta_1 x$ . The RESIDUAL part represents deviations of the data from the line of population means. We assume that these deviations are Normally distributed with standard deviation  $\sigma$ .

We use  $\varepsilon$  (the lowercase Greek letter epsilon) to stand for the RESIDUAL part

of the statistical model. A response  $y$  is the sum of its mean and a chance deviation  $\varepsilon$  from the mean. These model deviations  $\varepsilon$  represent “noise,” that is, variation in  $y$  due to other causes that prevent the observed  $(x, y)$ -values from forming a perfectly straight line on the scatterplot.

## SIMPLE LINEAR REGRESSION MODEL

Given  $n$  observations of the explanatory variable  $x$  and the response variable  $y$ ,

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

the **statistical model for simple linear regression** states that the observed response  $y_i$  when the explanatory variable takes the value  $x_i$  is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Here  $\beta_0 + \beta_1 x_i$  is the mean response when  $x = x_i$ . The deviations  $\varepsilon_i$  are assumed to be independent and Normally distributed with mean 0 and standard deviation  $\sigma$ .

The **parameters of the model** are  $\beta_0$ ,  $\beta_1$  and  $\sigma$ .



Because the means  $\mu_y$  lie on the line  $\mu_y = \beta_0 + \beta_1 x$ , they are all determined by  $\beta_0$  and  $\beta_1$ . Thus, once we have estimates of  $\beta_0$  and  $\beta_1$  the linear relationship determines the estimates of  $\mu_y$  for all values of  $x$ . Linear regression allows us to do inference not only for subpopulations for which we have data but also for those corresponding to  $x$ 's not present in the data. These  $x$ -values can be both within and outside the range of observed  $x$ 's. *However, extreme caution must be taken when performing inference for an  $x$ -value outside the range of the observed  $x$ 's because there is no assurance that the same linear relationship between  $\mu_y$  and  $x$  holds.*

Given the simple linear regression model, we will now learn how to do inference about

- the slope  $\beta_1$  and the intercept  $\beta_0$  of the population regression line,
- the mean response  $\mu_y$  for a given value of  $x$  and
- an individual future response  $y$  for a given value of  $x$ .

## Estimating the regression parameters



least-squares regression, p. 113

The method of least squares presented in Chapter 2 fits a line to summarize a relationship between the observed values of an explanatory variable and a response variable. Now we want to use the least-squares line as a basis for inference about a population from which our observations are a sample. *We can do this only when the statistical model just presented holds.* In that setting, the slope  $b_1$  and intercept  $b_0$  of the least-squares line



$$\hat{y} = b_0 + b_1x$$

estimate the slope  $\beta_1$  and the intercept  $\beta_0$  of the population regression line.

Using the formulas from Chapter 2 (page 115), the slope of the least-squares line is

$$b_1 = r s_y / s_x$$

and the intercept is

$$b_0 = \bar{y} - b_1 \bar{x}$$

Here,  $r$  is the correlation between  $y$  and  $x$ ,  $s_y$  is the standard deviation of  $y$ , and  $s_x$  is the standard deviation of  $x$ . Notice that if the slope is 0, so is the correlation, and vice versa. We will discuss this relationship more later in the chapter.



correlation, p. 103

The predicted value of  $y$  for a given value  $x^*$  of  $x$  is the point on the least-squares line  $\hat{y} = b_0 + b_1x^*$ . This is an unbiased estimator of the mean response  $\mu_y$  when  $x = x^*$ . The **residual** is

**residual**

$$e_i = \text{observed response} - \text{predicted response}$$

$$= y_i - \hat{y}_i$$

$$= y_i - b_0 - b_1x_i$$

The residuals  $e_i$  correspond to the model deviations  $\varepsilon_i$ . The  $e_i$  sum to 0, and the  $\varepsilon_i$

come from a population with mean 0. Because we do not observe the  $\varepsilon_i$ , we use the residuals to check the model assumptions of the  $\varepsilon_i$ .

The remaining parameter to be estimated is  $\sigma$ , which measures the variation of  $y$  about the population regression line. Because this parameter is the standard deviation of the model deviations, it should come as no surprise that we use the residuals to estimate it. As usual, we work first with the variance and take the square root to obtain the standard deviation.

For simple linear regression, the estimate of  $\sigma^2$  is the average squared residual

$$s^2 = \frac{\sum e_i^2}{n-2}$$

$$= \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$



sample variance, p. 42

We average by dividing the sum by  $n - 2$  in order to make  $s^2$  an unbiased estimate of  $\sigma^2$ . The sample variance of  $n$  observations uses the divisor  $n - 1$  for this same reason. The quantity  $n - 2$  is called the degrees of freedom for  $s^2$ . The estimate of **the model standard deviation  $\sigma$**  is given by

model standard deviation  $\sigma$



$$s = \sqrt{s^2}$$

We will use statistical software to calculate the regression for predicting BMI from physical activity for Example 10.1. In entering the data, we chose the names PA for the explanatory variable and BMI for the response. *It is good practice to use names, rather than just  $x$  and  $y$ , to remind yourself which variables the output describes.*

## Example

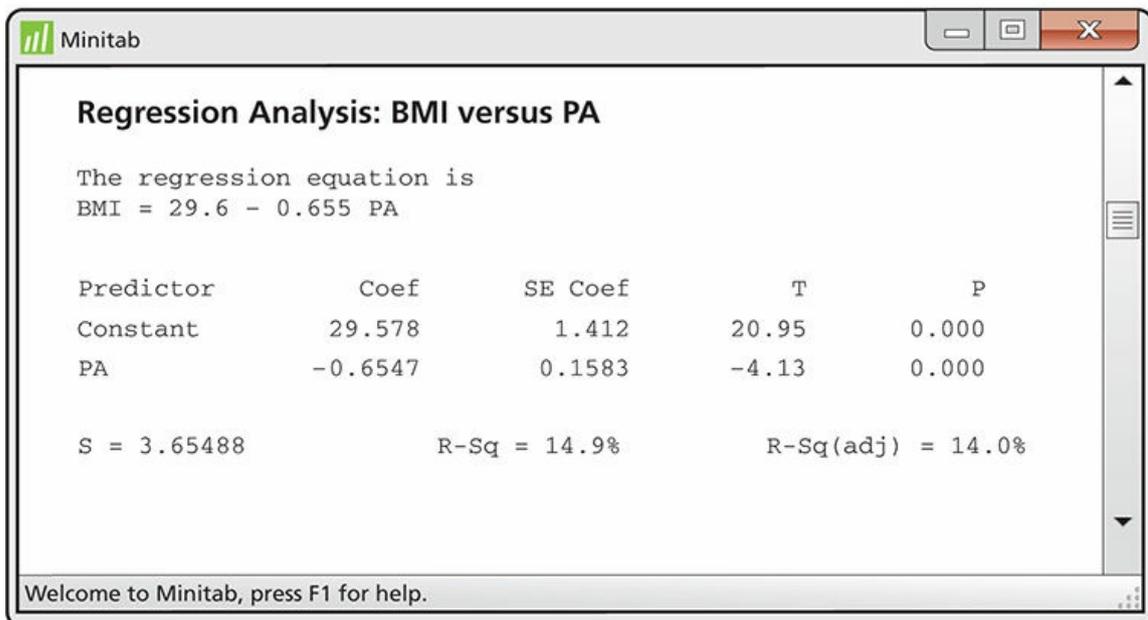
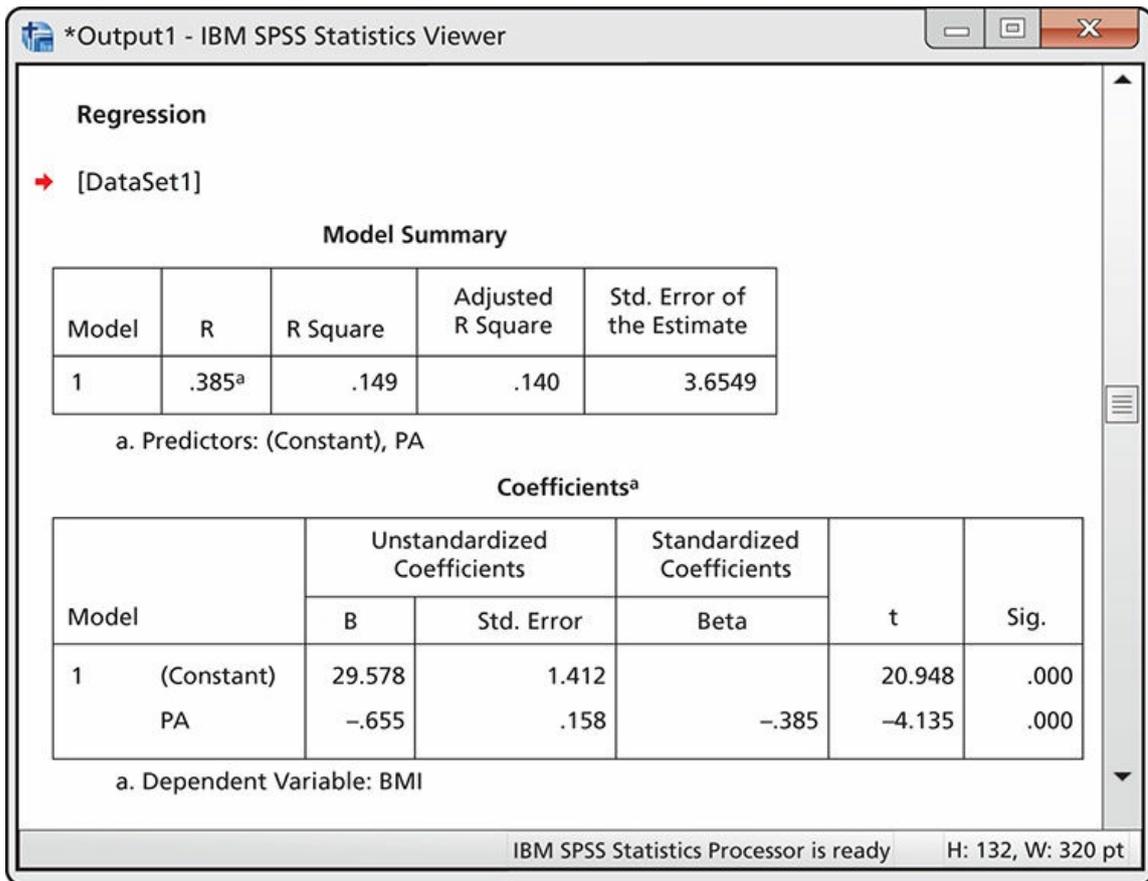
### 10.3 Statistical software output for BMI and physical activity.

Figure 10.4 gives the outputs from three commonly used statistical software

packages and Excel. Other software will give similar information. The SPSS output reports estimates of our three parameters as  $b_0 = 29.578$ ,  $b_1 = -0.655$ , and  $s = 3.6549$ . Be sure that you can find these entries in this output and the corresponding values in the other outputs.

The least-squares regression line is the straight line that is plotted in Figure 10.3. We would report it as

$$\text{BMI} = 29.578 - 0.655\text{PA}$$



Excel

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	<i>Regression Statistics</i>						
4	Multiple R	0.38540906					
5	R Square	0.14854014					
6	Adjusted R Square	0.13985178					
7	Standard Error	3.65488311					
8	Observations	100					
9							
10	ANOVA						
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	1	228.3771867	228.3772	17.09644	7.50303E-05	
13	Residual	98	1309.100713	13.35817			
14	Total	99	1537.4779				
15							
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-Value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	29.5782471	1.411978287	20.94809	5.71E-38	26.77622218	32.3802721
18	PA	-0.65468577	0.158336132	-4.13478	7.5E-05	-0.968898666	-0.340472865

SAS

Root MSE	3.65488	R-Square	0.1485
Dependent Mean	23.93900	Adj R-Sq	0.1399
Coeff Var	15.26748		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	29.57825	1.41198	20.95	<.0001
PA	PA	1	-0.65469	0.15834	-4.13	<.0001

**FIGURE 10.4**

Regression output from SPSS, Minitab, Excel, and SAS for the physical activity example.



with a model standard deviation of  $s = 3.655$ . Note that the number of digits provided varies with the software used and we have rounded the values to three decimal places. *It is important to avoid cluttering up your report of the results of a statistical analysis with many digits that are not relevant.* Software often reports many more digits than are meaningful or useful.

The outputs contain other information that we will ignore for now. Computer outputs often give more information than we want or need. This is done to reduce user frustration when a software package does not print out the particular statistics wanted for an analysis. *The experienced user of statistical software learns to ignore the parts of the output that are not needed for the current problem.*

## Example

### 10.4 Predicted values and residuals for BMI.

We can now use the least-squares regression equation to find the predicted BMI corresponding to any value of PA. Suppose that a female college student averages 8000 steps per day. We predict that this person will have a BMI of

$$29.578 - 0.655(8) = 24.338$$

If her actual BMI is 25.655, then the residual would be

$$y - \hat{y} = 25.655 - 24.338 = 1.317$$



Now that we have fitted a line, we should check the conditions that the simple linear regression model imposes on this fit. *There is no point in trying to do statistical inference if the data do not, at least approximately, meet the conditions that are the foundation for the inference.*

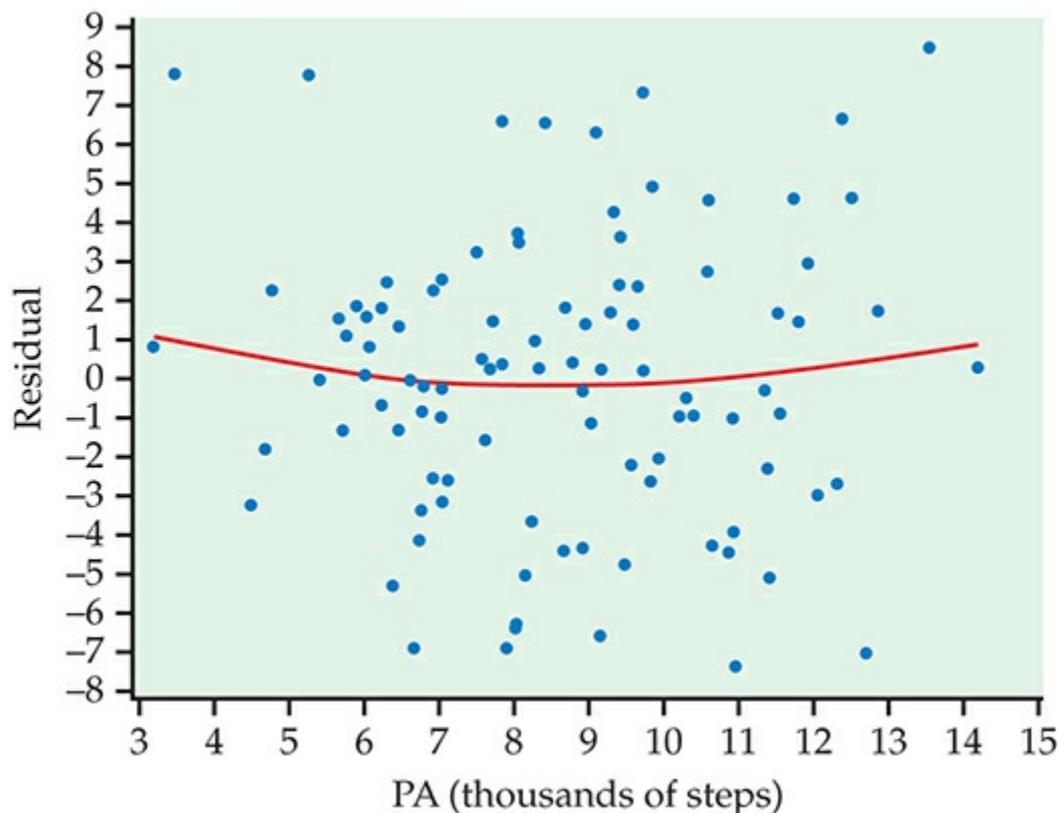
This check is done through a visual examination of the residuals for Normality, constant variance, and any remaining patterns in the data. We usually plot the residuals both against the case number (especially if this reflects the order in which the observations were collected) and against the explanatory variable. For this

example, we will just look at the residuals against the explanatory variable.



scatterplot smoothers, p. 96

Figure 10.5 gives a plot of the residuals versus physical activity with a smooth-function fit. The smooth function suggests that the residuals increase slightly at both low and high physical activity levels. This could mean that a curved relationship between BMI and physical activity would better fit the data. It also could just be chance variation.



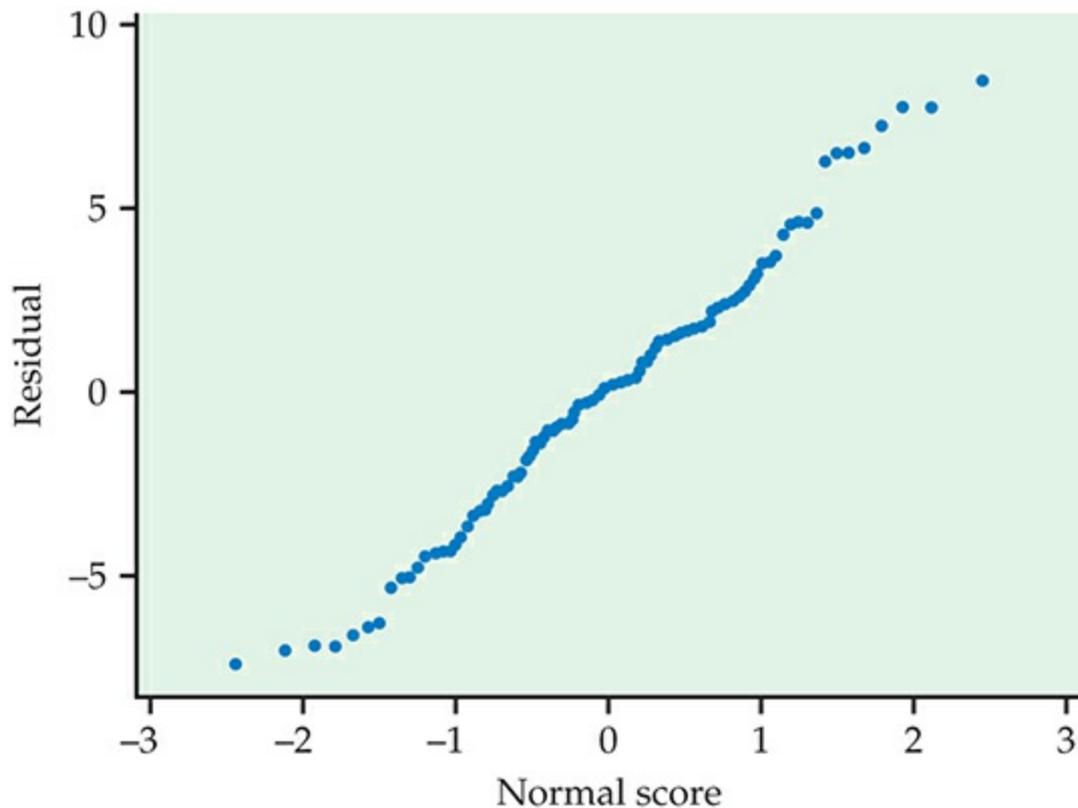
**FIGURE 10.5**

Plot of residuals versus physical activity (PA) with a smooth function, for the physical activity example.

Notice that there is a large positive residual near each end of the physical activity range. Since the effect does not appear to be particularly large, we will ignore this for the present analysis and investigate this further in Exercise 11.21 (page 636).

In Figure 10.5 the spread of the residuals is roughly uniform across the range of PA, suggesting that the assumption of a common standard deviation is reasonable. There also do not appear to be any outliers or influential observations. Finally, Figure 10.6 is a Normal quantile plot of the residuals. Because the plot looks fairly straight, we are confident that we do not have a serious violation of our assumption

that the residuals are Normally distributed.



**FIGURE 10.6**

Normal quantile plot of the residuals for the physical activity example.

## USE YOUR KNOWLEDGE

### 10.1 Understanding a linear regression model.

Consider a linear regression model with  $\mu_y = 51.6 + 3.1x$  and standard deviation  $\sigma = 5.2$ .

- What is the slope of the population regression line?
- Explain clearly what this slope says about the change in the mean of  $y$  for a change in  $x$ .
- What is the subpopulation mean when  $x = 10$ ?
- Using the 68–95–99.7 rule, between what two values would approximately 95% of the observed responses,  $y$  fall when  $x = 10$ ? (*Hint*: Refer to the two important parts of a linear regression on page 565)

### 10.2 More on BMI and physical activity.

Refer to Examples 10.3 (page 569) and 10.4 (page 572).

- (a) What is the predicted BMI for a woman who averages 9400 steps per day?
- (b) If an observed BMI at  $x = 9.4$  were 24.3, what would be the residual?
- (c) Suppose that you wanted to use the estimated population regression line to examine the predicted BMI for a woman who averages 4000, 10,000, or 16,000 steps per day. Discuss the appropriateness of using the equation to predict BMI for each of these activity levels.

## Confidence intervals and significance tests

Chapter 7 presented confidence intervals and significance tests for means and differences in means. In each case, inference rested on the standard errors of estimates and on  $t$  distributions. Inference in simple linear regression is similar in principle. For example, the confidence intervals have the form

$$\text{estimate} \pm t^* SE_{\text{estimate}}$$

where  $t^*$  is a critical point of a  $t$  distribution. The formulas for the estimate and standard error, however, are more complicated.



central limit theorem, p. 307

As a consequence of the model assumptions about the deviations  $\varepsilon$  the sampling distributions of  $b_0$  and  $b_1$  are Normally distributed with means  $\beta_0$  and  $\beta_1$  and standard deviations that are multiples of  $\sigma$ , the model parameter that describes the variability about the true regression line. In fact, even if the  $\varepsilon_i$  are not Normally distributed, a general form of the central limit theorem tells us that the distributions of  $b_0$  and  $b_1$  will be approximately Normal.

Because we do not know  $\sigma$  we estimate it by  $s$  the variability of the data about the least-squares line. When we do this, we move from the Normal distribution to  $t$  distributions with degrees of freedom  $n - 2$ , the degrees of freedom of  $s$ . We give formulas for the standard errors  $SE_{b_1}$  and  $SE_{b_0}$  in Section 10.2. For now, we will concentrate on the basic ideas and let the computer do the computations.

### CONFIDENCE INTERVAL AND SIGNIFICANCE TEST FOR THE REGRESSION SLOPE

A level  $C$  confidence interval for the slope  $\beta_1$  is

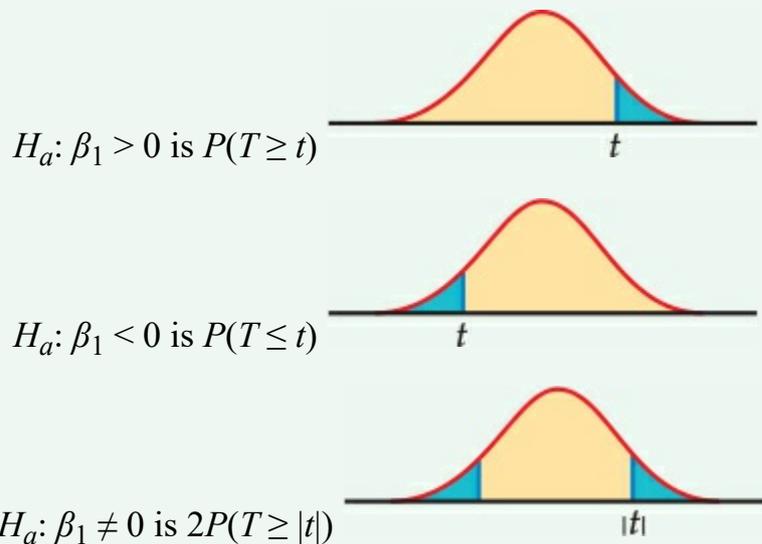
$$b_1 \pm t^* SE_{b_1}$$

In this expression  $t^*$  is the value for the  $t(n-2)$  density curve with area  $C$  between  $-t^*$  and  $t^*$ .

To test the hypothesis  $H_0: \beta_1 = 0$  compute the **test statistic**

$$t = b_1 / SE_{b_1}$$

The **degrees of freedom** are  $n - 2$ . In terms of a random variable  $T$  having the  $t(n - 2)$  distribution, the  $P$ -value for a test of  $H_0$  against



Formulas for confidence intervals and significance tests for the intercept  $\beta_0$  are exactly the same, replacing  $b_1$  and  $SE_{b_1}$  by  $b_0$  and its standard error  $SE_{b_0}$ . Although computer outputs often include a test of  $H_0: \beta_0 = 0$  this information usually has little practical value. From the equation for the population regression line,  $\mu_y = \beta_0 + \beta_1 x$ , we see that  $\beta_0$  is the mean response corresponding to  $x = 0$ . In many practical situations, this subpopulation does not exist or is not interesting.

On the other hand, the test of  $H_0: \beta_1 = 0$  is quite useful. When we substitute  $\beta_1 = 0$  in the model, the  $x$  term drops out and we are left with

$$\mu_y = \beta_0$$

This model says that the mean of  $y$  does *not* vary with  $x$ . In other words, all the  $y$ 's come from a single population with mean  $\beta_0$  which we would estimate by  $\bar{y}$ . The hypothesis  $H_0: \beta_1 = 0$  therefore says that there is *no straight-line relationship between  $y$  and  $x$*  and that linear regression of  $y$  on  $x$  is of no value for predicting  $y$ .

## Example

## 10.5 Statistical software output, continued.

The computer outputs in Figure 10.4 (pages 570 and 571) for the BMI problem contain the information needed for inference about the regression slope and intercept. Let's look at the SPSS output. The column labeled Std. Error gives the standard errors of the estimates. The value of  $SE_{b_1}$  appears on the line labeled with the variable name for the explanatory variable, PA. It is given as 0.158. In a summary we would report that the regression coefficient for the average number of steps per day is  $-0.655$  with a standard error of 0.158.

The  $t$  statistic and  $P$ -value for the test of  $H_0: \beta_1 = 0$  against the two-sided alternative  $H_a: \beta_1 \neq 0$  appear in the columns labeled  $t$  and Sig. We can verify the  $t$  calculation from the formula for the standardized estimate:

$$t = b_1 / SE_{b_1} = -0.655 / 0.158 = -4.14$$

The  $P$ -value is given as 0.000. This is a rounded number, and from that information we can conclude that  $P < 0.0005$ . The other outputs in Figure 10.4 also indicate that the  $P$ -value is very small. We will report the result as  $P < 0.001$  because 1 chance in 1000 is sufficiently small for us to decisively reject the null hypothesis.

We have found a statistically significant linear relationship between physical activity and BMI. The estimated slope is more than 4 standard deviations away from zero. Because this is highly unlikely to happen if the true slope is zero, we have strong evidence for our claim.



Note, however, that this is not the same as concluding that we have found a strong linear relationship between the response and explanatory variables in this example. We saw in Figure 10.3 that there was a lot of scatter about the regression line. *A very small  $P$ -value for the significance test for a zero slope does not necessarily imply that we have found a strong relationship.*

A confidence interval will provide additional information about the relationship.

### Example

## 10.6 Confidence interval for the slope.

A confidence interval for  $\beta_1$  requires a critical value  $t^*$  from the  $t(n - 2) = t(98)$  distribution. In Table D there are entries for 80 and 100 degrees of freedom. The values for these rows are very similar. To be conservative, we will use the larger critical value, for 80 degrees of freedom. Find the confidence level values at the bottom of the table. In the 95% confidence column, the entry for 80 degrees of freedom is  $t^* = 1.990$ .

To compute the 95% confidence interval for  $\beta_1$  we combine the estimate of the slope with the margin of error:

$$\begin{aligned} b_1 \pm t^* SE_{b_1} &= -0.655 \pm (1.990)(0.158) \\ &= -0.655 \pm 0.314 \end{aligned}$$

The interval is  $(-0.969 \pm -0.341)$ . This agrees with the intervals given by the software outputs that provide this information in Figure 10.4. We estimate that an increase of 1000 steps per day is associated with a decrease in BMI of between 0.341 and 0.969  $\text{kg/m}^2$ .

Note that the intercept in this example is not of practical interest. It estimates average BMI when the activity level is 0, a value that isn't realistic. For this reason, we do not compute a confidence interval for  $\beta_0$  or discuss the significance test available in the software.

## USE YOUR KNOWLEDGE

### 10.3 Significance test for the slope.

Test the null hypothesis that the slope is zero versus the two-sided alternative in each of the following settings using the  $\alpha = 0.05$  significance level:

- (a)  $n = 25$ ,  $\hat{y} = 30.5 + 1.8x$ , and  $SE_{b_1} = 0.95$
- (b)  $n = 25$ ,  $\hat{y} = 32.8 + 2.0x$ , and  $SE_{b_1} = 0.95$
- (c)  $n = 100$ ,  $\hat{y} = 28.3 + 1.7x$ , and  $SE_{b_1} = 0.55$

### 10.4 95% confidence interval for the slope.

For each of the settings in the previous exercise, find the 95% confidence interval for the slope.

## Confidence intervals for mean response

Besides performing inference about the slope (and sometimes the intercept) in a linear regression, we may want to use the estimated regression line to make predictions about the response  $y$  at certain values of  $x$ . We may be interested in the mean response for different subpopulations or in the response of future observations at different values of  $x$ . In either case, we would want an estimate and associated margin of error.

For any specific value of  $x$ , say  $x^*$ , the mean of the response  $y$  in this subpopulation is given by

$$\mu_y = \beta_0 + \beta_1 x^*$$

To estimate this mean from the sample, we substitute the estimates  $b_0$  and  $b_1$  for  $\beta_0$  and  $\beta_1$ :

$$\hat{\mu}_y = b_0 + b_1 x^*$$

A confidence interval for  $\mu_y$  adds to this estimate a margin of error based on the standard error  $SE_{\hat{\mu}_y}$ . (The formula for the standard error is given in Section 10.2.)

### CONFIDENCE INTERVAL FOR A MEAN RESPONSE

A level  $C$  confidence interval for the mean response  $\mu_y$  when  $x$  takes the value  $x^*$  is

$$\hat{\mu}_y \pm t^* SE_{\hat{\mu}_y}$$

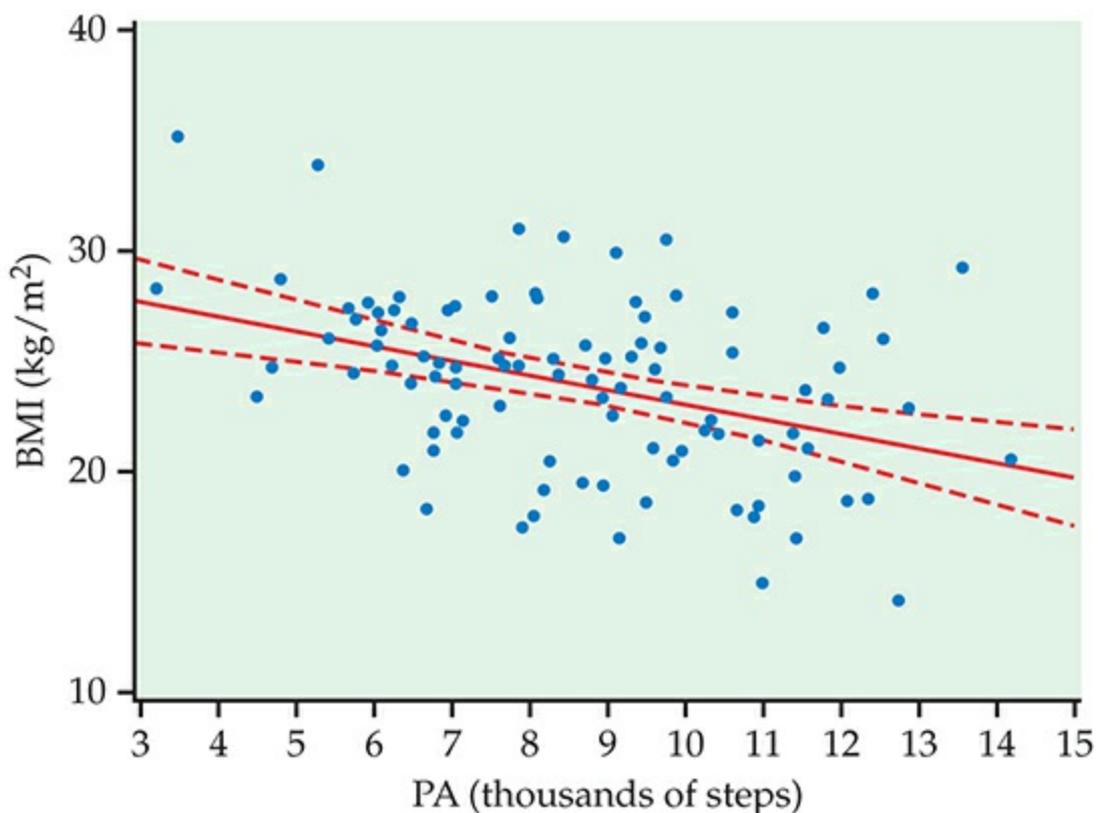
where  $t^*$  is the value for the  $t(n-2)$  density curve with area  $C$  between  $-t^*$  and  $t^*$

Many computer programs calculate confidence intervals for the mean response corresponding to each of the  $x$ -values in the data. Some can calculate an interval for any value  $x^*$  of the explanatory variable. We will use a plot to illustrate these intervals.

## Example

### 10.7 Confidence intervals for the mean response.

Figure 10.7 shows the upper and lower confidence limits on a graph with the data and the least-squares line. The 95% confidence limits appear as dashed curves. For any  $x^*$  the confidence interval for the mean response extends from the lower dashed curve to the upper dashed curve. The intervals are narrowest for values of  $x^*$  near the mean of the observed  $x$ 's and widen as  $x^*$  moves away from  $\bar{x}$ .



**FIGURE 10.7**

The 95% confidence limits (dashed curves) for the mean response for the physical activity example.

Some software will do these calculations directly if you input a value for the explanatory variable. Other software will calculate the intervals for each value of  $x$  in the data set. Creating a new data set with an additional observation with  $x$  equal to the value of interest and  $y$  missing will often work.

## Example

### 10.8 Confidence interval for an average of 9000 steps per day.

Let's find the confidence interval for the average BMI at  $x = 9.0$ . Our predicted BMI is

$$\begin{aligned}\widehat{\text{BMI}} &= 29.578 - 0.655PA \\ &= 29.578 - 0.655(9.0) \\ &= 23.7\end{aligned}$$

Software tells us that the 95% confidence interval for the mean response is 23.0 to 24.4 kg/m<sup>2</sup>.



If we sampled many women who averaged 9000 steps per day, we would expect their average BMI to be between 23.0 and 24.4 kg/m<sup>2</sup>. Note that many of the observations in Figure 10.7 lie outside the confidence bands. *These confidence intervals do not tell us what BMI to expect for a single observation at a particular average steps per day.* We need a different kind of interval, a prediction interval, for this purpose.

## Prediction intervals

In the last example, we predicted the average BMI for an average of 9000 steps per day. Suppose that we now want to predict an observation of BMI for a woman averaging 9000 steps per day. Our best guess for the BMI is what we obtained using the regression equation, that is, 23.7 kg/m<sup>2</sup>. The margin of error, on the other hand, is larger because it is harder to predict an individual value than to predict the mean.

The predicted response  $\hat{y}$  for an individual case with a specific value  $x^*$  of the explanatory variable  $x$  is

$$\hat{y} = b_0 + b_1x^*$$

This is the same as the expression for  $\hat{\mu}_y$ . That is, the fitted line is used both to estimate the mean response when  $x = x^*$  and to predict a single future response. We use the two notations  $\hat{\mu}_y$  and  $\hat{y}$  to remind ourselves of these two distinct uses.

A useful prediction should include a margin of error to indicate its accuracy.

The interval used to predict a future observation is called a **prediction interval**. Although the response  $y$  that is being predicted is a random variable, the interpretation of a prediction interval is similar to that for a confidence interval.

### prediction interval

Consider doing the following many times:

- Draw a sample of  $n$  observations  $(x_i, y_i)$  and then one additional observation  $(x^*, y)$ .
- Calculate the 95% prediction interval for  $y$  when  $x = x^*$  using the sample of size  $n$ .

Then 95% of the prediction intervals will contain the value of  $y$  for the additional observation. In other words, the probability that this method produces an interval that contains the value of a future observation is 0.95.

The form of the prediction interval is very similar to that of the confidence interval for the mean response. The difference is that the standard error  $SE_{\hat{y}}$  used in the prediction interval includes both the variability due to the fact that the least-squares line is not exactly equal to the true regression line *and* the variability of the future response variable  $y$  around the subpopulation mean. (The formula for  $SE_{\hat{y}}$  appears in Section 10.2.)

## PREDICTION INTERVAL FOR A FUTURE OBSERVATION

A level  $C$  prediction interval for a future observation on the response variable  $y$  from the subpopulation corresponding to  $x^*$  is

$$\hat{y} \pm t^* SE_{\hat{y}}$$

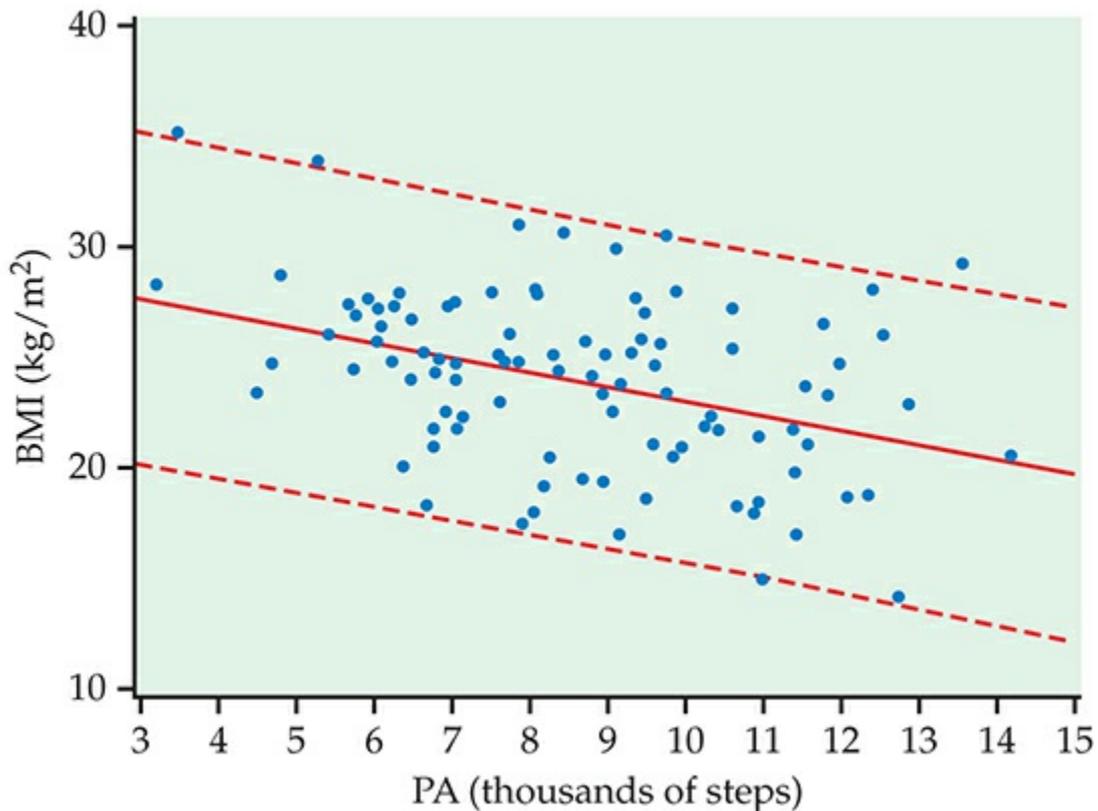
where  $t^*$  is the value for the  $t(n-2)$  density curve with area  $C$  between  $-t^*$  and  $t^*$ .

Again, we use a graph to illustrate the results.

### Example

#### 10.9 Prediction intervals for BMI.

Figure 10.8 shows the upper and lower prediction limits, along with the data and the least-squares line. The 95% prediction limits are indicated by the dashed curves. Compare this figure with Figure 10.7, which shows the 95% confidence limits drawn to the same scale. The upper and lower limits of the prediction intervals are farther from the least-squares line than are the confidence limits. This results in most, but not all, of the observations in Figure 10.8 lying within the prediction bands.



**FIGURE 10.8**

The 95% prediction limits (dashed curves) for individual responses for the physical activity example. Compare with Figure 10.7. The limits are wider because the margins of error incorporate the variability about the subpopulation means.

The comparison of Figures 10.7 and 10.8 reminds us that the interval for a single future observation must be larger than an interval for the mean of its subpopulation.

## Example

### 10.10 Prediction interval for an average of 9000 steps per day.

Let's find the prediction interval for a future observation of BMI for a college-aged woman who averages 9000 steps per day. The predicted value is the same as the estimate of the average BMI that we calculated in Example 10.8, that is,  $23.7 \text{ kg/m}^2$ . Software tells us that the 95% prediction interval is 16.4 to  $31.0 \text{ kg/m}^2$ . This interval is extremely wide, covering BMI values that are classified as underweight and obese. Because of the large amount of scatter about the regression line, prediction intervals here are relatively useless.



*Although a larger sample would better estimate the population regression line, it would not reduce the degree of scatter about the line.* This means that prediction intervals for BMI, given activity level, will *always* be wide. This example clearly demonstrates that a very small  $P$ -value for the significance test for a zero slope does not necessarily imply that we have found a strong predictive relationship.

## USE YOUR KNOWLEDGE

### 10.5 Margin of error for the predicted mean.

Refer to Example 10.8 (page 578). What is the 95% margin of error of  $\mu^y$  when  $x = 9.0$ ? Would you expect the margin of error to be larger, smaller, or the same for  $x = 5.0$ ? Explain your answer.

### 10.6 Margin of error for the predicted response.

Refer to Example 10.10. What is the 95% margin of error of  $\hat{y}$  when  $x = 9.0$ ? If you increased the sample size from  $n = 100$  to  $n = 400$ , would you expect the 95% margin of error to be roughly twice as large, half as small, or the same for  $x = 9.0$ ? Explain your answer.

## Transforming variables



We started our analysis of Example 10.1 with a scatterplot to check whether the relationship between BMI and physical activity could be summarized with a straight line. We followed that with a residual plot (Figure 10.5) and a Normal quantile plot (Figure 10.6) to check Normality and any remaining patterns in the data. *A check of model assumptions should always be done prior to inference.*

When there is a violation, it is best to consult an expert, as a more sophisticated regression model is likely needed. However, when the relationship between  $y$  and  $x$  is not linear, sometimes we can make it linear by a transformation of one or both of the variables. Here is an example.

## Example

### 10.11 Relationship between speed driven and fuel efficiency.



MPHMPG

Computers in some vehicles calculate various quantities related to the vehicle's performance. One of these is the fuel efficiency, or gas mileage, expressed as miles per gallon (mpg). Another is the average speed in miles per hour (mph). For one vehicle equipped in this way, mpg and mph were recorded each time the gas tank was filled, and the computer was then reset.<sup>2</sup> How does the speed at which the vehicle is driven affect the fuel efficiency? We will work with a simple random sample of 60 observations.

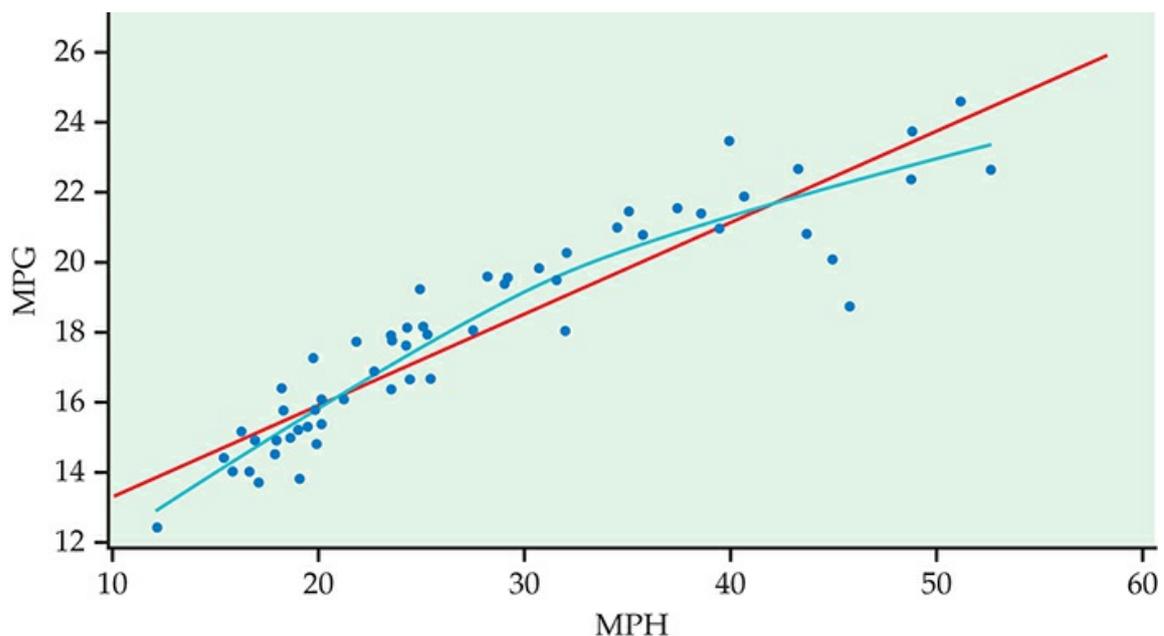
Our statistical modeling for this data set concerns the process by which speed affects the fuel efficiency. Except possibly for the owner, no one cares about the particular vehicle. The results are interesting only if they can be applied to other, similar vehicles that are driven under similar conditions. Although we would not

expect the parameters that describe the relationship between speed and fuel efficiency to be *exactly* the same for similar vehicles, we would expect to find qualitatively similar results.

## Example

### 10.12 Graphical display of the fuel efficiency and speed relationship.

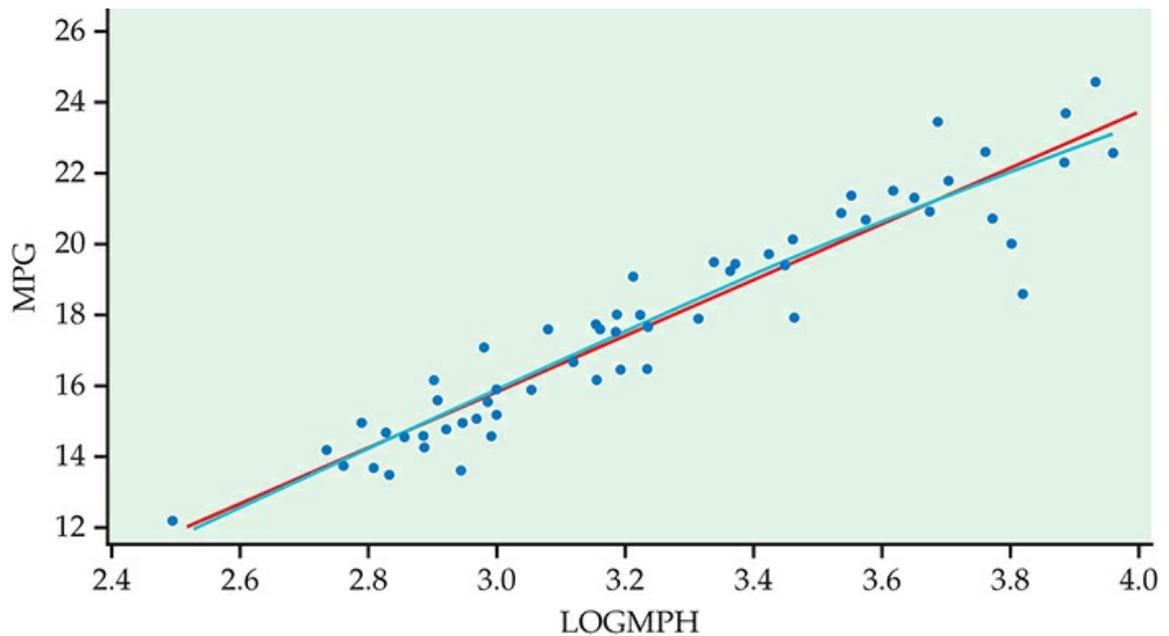
Figure 10.9 is a plot of fuel efficiency versus speed for our sample of 60 observations. We use the variable names MPG and MPH. The least-squares regression line and a smooth function are also shown in the plot. Although there is a positive association between MPG and MPH, the fit is not linear. The smooth function shows us that the relationship levels off somewhat with increasing speed.



**FIGURE 10.9**

Scatterplot of MPG versus MPH with a smooth function and the least-squares line, for Example 10.12. The relationship between MPG and MPH does not appear to be linear.

Given this nonlinearity, we need to make a choice about how to proceed. One approach would be to confine our interest to speeds that are 30 mph or less, a region where it appears that a line would be a good fit to the data. Another possibility is to consider a transformation that will make the relationship approximately linear for the entire set of data.



**FIGURE 10.10**

Scatterplot of MPG versus the logarithm of MPH with a smooth function and the least-squares line, for Example 10.13. Here, the line and smooth function are very close.

## Example

### 10.13 Is this relationship linear?

One type of function that looks similar to the smooth-function fit in Figure 10.9 is a logarithm. Therefore, we will examine the effect of transforming speed by taking the natural logarithm. The result is shown in Figure 10.10. In this plot the smooth function and the line are quite close. We are satisfied that the relationship between the log of speed and fuel efficiency is approximately linear for this set of data. We also see that there is only a small amount of scatter about the regression line, suggesting a strong predictive relationship.



log transformation, p. 93



Although this transformation has resulted in an approximately linear

relationship, there are still other assumptions of the simple linear model that must be met. For this example, one can show that these assumptions are also satisfied, so statistical inference can be performed. *In other cases, transforming a variable may help linearity but harm the Normality and constant-variance assumptions.* In those cases a more sophisticated model is needed.

---

## BEYOND THE BASICS

### Nonlinear regression

When the relationship is not linear, we often use models that allow for various types of curved relationships. These models are called **nonlinear models**.

#### nonlinear models

The technical details are much more complicated for nonlinear models. In general, we cannot write down simple formulas for the parameter estimates; we use a computer to solve systems of equations to find the estimates. However, the basic principles are those that we have already learned. For example,

$$\text{DATA} = \text{FIT} + \text{RESIDUAL}$$

still applies. The FIT is a nonlinear (curved) function, and the residuals are assumed to be an SRS from the  $N(0, \sigma)$  distribution. The nonlinear function contains parameters that must be estimated from the data. Approximate standard errors for these estimates are part of the standard output provided by software. Here is an example.

---

### Example

#### 10.14 Investing in one's bone health.



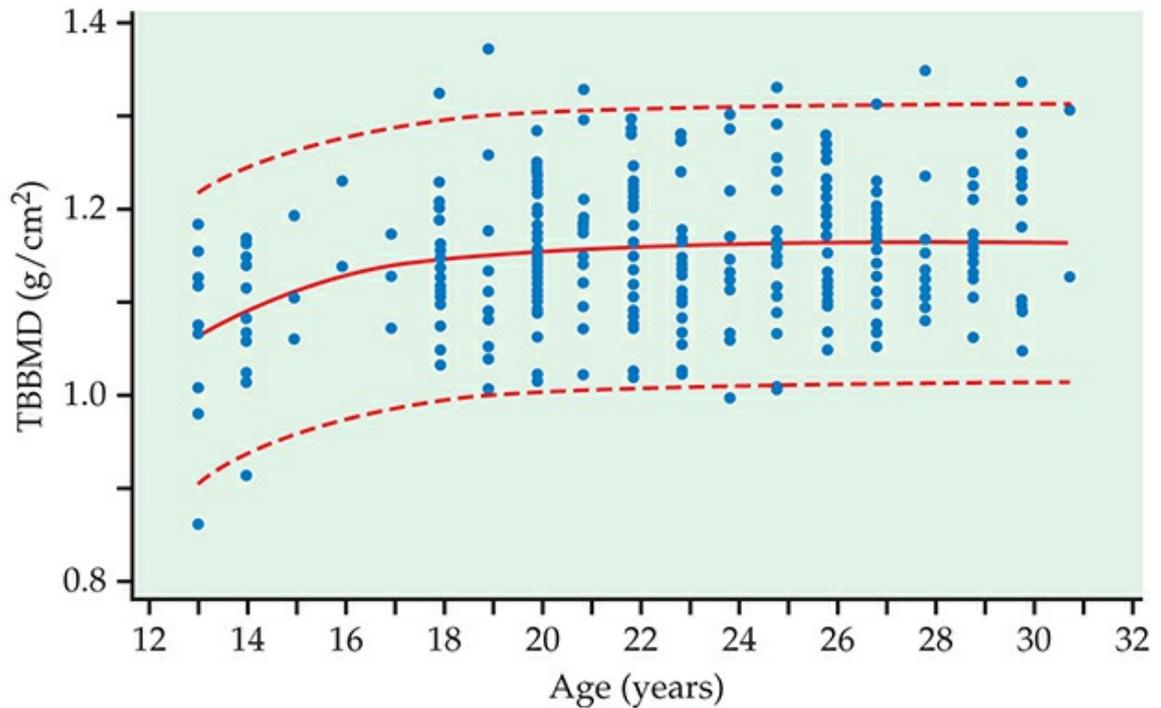
As we age, our bones become weaker and are more likely to break. Osteoporosis (or weak bones) is the major cause of bone fractures in older women. Various researchers have studied this problem by looking at how and when bone mass is accumulated by young women. They've determined that up to 90% of a person's peak bone mass is acquired by age 18 in girls.<sup>3</sup> This makes youth the best time to invest in stronger bones.

Figure 10.11 displays data for a measure of bone strength, called "total body bone mineral density" (TBBMD), and age for a sample of 256 young women.<sup>4</sup> TBBMD is measured in grams per square centimeter ( $\text{g}/\text{cm}^2$ ), and age is recorded in years. The solid curve is the nonlinear fit, and the dashed curves are 95% prediction limits. Similar to our example of BMI and activity level, there is a large amount of scatter about the fitted curve. Although prediction intervals may be useless in this case, the researchers can draw some conclusions regarding the relationship.

The fitted nonlinear equation is

$$\hat{y} = 1.162e^{-1.162+0.28x} + e^{-1.162+0.28x}$$

In this equation,  $\hat{y}$  is the predicted value of TBBMD, the response variable; and  $x$  is age, the explanatory variable. A straight line would not do a very good job of summarizing the relationship between TBBMD and age. At first, TBBMD increases with age, but then it levels off as age increases. The value of the function where it is level is called "peak bone mass"; it is a parameter in the nonlinear model. The estimate is 1.162 and the standard error is 0.008. Software gives the 95% confidence interval as (1.146, 1.178). Other calculations could be done to determine the age by which up to 90% of this peak bone mass is acquired.



**FIGURE 10.11**

Plot of total body bone mineral density versus age.

The long-range goals of the researchers who conducted this study include developing intervention programs (exercise and increasing calcium intake have been shown to be effective) for young women that will increase their TBBMD.

### SECTION 10.1 Summary

The statistical model for **simple linear regression** assumes that the means of the response variable  $y$  fall on a line when plotted against  $x$  with the observed  $y$ 's varying Normally about these means. For  $n$  observations, this model can be written

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where  $i = 1, 2, \dots, n$ , and the  $\varepsilon_i$  are assumed to be independent and Normally distributed with mean 0 and standard deviation  $\sigma$ . Here  $\beta_0 + \beta_1 x_i$  is the mean response when  $x = x_i$ . The **parameters** of the model are  $\beta_0$ ,  $\beta_1$  and  $\sigma$ .

The **population regression line** intercept and slope,  $\beta_0$  and  $\beta_1$  are estimated by the intercept and slope of the **least-squares regression line**,  $b_0$  and  $b_1$ . The parameter  $\sigma$  is estimated by

$$s = \sqrt{\frac{\sum e_i^2}{n-2}}$$

where the  $e_i$  are the **residuals**

$$e_i = y_i - \hat{y}_i$$

Prior to inference, always examine the residuals for Normality, constant variance,

and any other remaining patterns in the data. **Plots of the residuals** both against the case number and against the explanatory variable are usually done as part of this examination.

A **level  $C$  confidence interval for  $\beta_1$**  is

$$b_1 \pm t^* SE_{b_1}$$

where  $t^*$  is the value for the  $t(n - 2)$  density curve with area  $C$  between  $-t^*$  and  $t^*$

The **test of the hypothesis  $H_0: \beta_1 = 0$**  is based on the  **$t$  statistic**

$$t = b_1 / SE_{b_1}$$

and the  $t(n - 2)$  distribution. This tests whether there is a straight-line relationship between  $y$  and  $x$ . There are similar formulas for confidence intervals and tests for  $\beta_0$  but these are meaningful only in special cases.

The **estimated mean response** for the subpopulation corresponding to the value  $x^*$  of the explanatory variable is

$$\mu^{\wedge}y = b_0 + b_1 x^*$$

A **level  $C$  confidence interval for the mean response** is

$$\mu^{\wedge}y \pm t^* SE_{\mu^{\wedge}}$$

where  $t^*$  is the value for the  $t(n - 2)$  density curve with area  $C$  between  $-t^*$  and  $t^*$

The **estimated value of the response variable  $y$**  for a future observation from the subpopulation corresponding to the value  $x^*$  of the explanatory variable is

$$y^{\wedge} = b_0 + b_1 x^*$$

A **level  $C$  prediction interval** for the estimated response is

$$y^{\wedge} \pm t^* SE_{y^{\wedge}}$$

where  $t^*$  is the value for the  $t(n - 2)$  density curve with area  $C$  between  $-t^*$  and  $t^*$ . The standard error for the prediction interval is larger than that for the confidence interval because it also includes the variability of the future observation around its subpopulation mean.

Sometimes a **transformation** of one or both of the variables can make their relationship linear. However, these transformations can harm the assumptions of Normality and constant variance, so it is important to examine the residuals.

## 10.2 More Detail about Simple Linear Regression

When you complete this section, you will be able to

- Construct a linear regression ANOVA table.
- Use an ANOVA table to perform the ANOVA  $F$  test and draw appropriate conclusions regarding  $H_0: \beta_1 = 0$ .
- Use an ANOVA table to compute the square of the sample correlation and provide an interpretation of it in terms of explained variation.
- Perform, using a calculator, inference in simple linear regression when a computer is not available.
- Differentiate the formulas for the standard error that we use for a confidence interval for the mean response and the standard error that we use for a prediction interval when  $x = x^*$ .
- Test the hypothesis that there is no linear association in the population and summarize the results.
- Explain the close connection between the tests  $H_0: \beta_1 = 0$  and  $H_0: \rho = 0$ .

In this section we study three topics. The first is analysis of variance for regression. If you plan to read Chapter 11 on multiple regression, you should study this material. The second topic concerns computations for regression inference. The section we just completed assumes that you have access to software or a statistical calculator. Here we present and illustrate the use of formulas for the inference procedures. Finally, we discuss inference for correlation.

### Analysis of variance for regression

The usual computer output for regression includes additional calculations called **analysis of variance**. Analysis of variance, often abbreviated ANOVA, is essential for multiple regression (Chapter 11) and for comparing several means (Chapters 12 and 13). Analysis of variance summarizes information about the sources of variation in the data. It is based on the

analysis of variance

$$\text{DATA} = \text{FIT} + \text{RESIDUAL}$$

framework.

The total variation in the response  $y$  is expressed by the deviations  $y_i - \bar{y}$ . If these deviations were all 0, all observations would be equal and there would be no variation in the response. There are two reasons why the individual observations  $y_i$  are not all equal to their mean  $\bar{y}$ .

1. The responses  $y_i$  correspond to different values of the explanatory variable  $x$  and will differ because of that. The fitted value  $\hat{y}_i$  estimates the mean response for  $x_i$ . The differences  $\hat{y}_i - \bar{y}$  reflect the variation in mean response due to differences in the  $x_i$ . This variation is accounted for by the regression line because the  $\hat{y}$ 's lie exactly on the line.
2. Individual observations will vary about their mean because of variation within the subpopulation of responses for a fixed  $x_i$ . This variation is represented by the residuals  $y_i - \hat{y}_i$  that record the scatter of the actual observations about the fitted line.

The overall deviation of any  $y$  observation from the mean of the  $y$ 's is the sum of these two deviations:

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

In terms of deviations, this equation expresses the idea that DATA = FIT + RESIDUAL.

Several times we have measured variation by an average of squared deviations. If we square each of the preceding three deviations and then sum over all  $n$  observations, it can be shown that the sums of squares add:

$$\Sigma(y_i - \bar{y})^2 = \Sigma(\hat{y}_i - \bar{y})^2 + \Sigma(y_i - \hat{y}_i)^2$$

We rewrite this equation as

$$\text{SST} = \text{SSM} + \text{SSE}$$

where

$$\text{SST} = \Sigma(y_i - \bar{y})^2$$

$$\text{SSM} = \Sigma(\hat{y}_i - \bar{y})^2$$

$$\text{SSE} = \Sigma(y_i - \hat{y}_i)^2$$

The SS in each abbreviation stands for **sum of squares**, and the T, M, and E stand for total, model, and error, respectively. ("Error" here stands for deviations from the line, which might better be called "residual" or "unexplained variation.") The total variation, as expressed by SST, is the sum of the variation due to the straight-line model (SSM) and the variation due to deviations from this model (SSE). This

partition of the variation in the data between two sources is the heart of analysis of variance.

### sum of squares

If  $H_0: \beta_1 = 0$  were true, there would be no subpopulations and all of the  $y$ 's should be viewed as coming from a single population with mean  $\mu_y$ . The variation of the  $y$ 's would then be described by the sample variance

$$s_y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$$



degrees of freedom, p. 44

The numerator in this expression is SST. The denominator is the total degrees of freedom, or simply DFT.

Just as the total sum of squares SST is the sum of SSM and SSE, the total degrees of freedom DFT is the sum of DFM and DFE, the degrees of freedom for the model and for the error:

$$DFT = DFM + DFE$$

The model has one explanatory variable  $x$ , so the degrees of freedom for this source are  $DFM = 1$ . Because  $DFE = n - 1$ , this leaves  $DFT = n - 2$  as the degrees of freedom for error.

For each source, the ratio of the sum of squares to the degrees of freedom is called the **mean square**, or simply MS. The general formula for a mean square is

### mean square

$$MS = \frac{\text{sum of squares}}{\text{degrees of freedom}}$$

Each mean square is an average squared deviation. MST is just  $s_y^2$ , the sample variance that we would calculate if all of the data came from a single population. MSE is also familiar to us:

$$MSE = s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

It is our estimate of  $\sigma^2$ , the variance about the population regression line.

## SUMS OF SQUARES, DEGREES OF FREEDOM, AND MEAN SQUARES

**Sums of squares** represent variation present in the responses. They are

calculated by summing squared deviations. **Analysis of variance** partitions the total variation between two sources.

The sums of squares are related by the formula

$$SST = SSM + SSE$$

That is, the total variation is partitioned into two parts, one due to the model and one due to deviations from the model.

**Degrees of freedom** are associated with each sum of squares. They are related in the same way:

$$DFT = DFM + DFE$$

To calculate **mean squares**, use the formula

$$MS = \frac{\text{sum of squares}}{\text{degrees of freedom}}$$

In Section 2.4 (page 120) we noted that  $r^2$  is the fraction of variation in the values of  $y$  that is explained by the least-squares regression of  $y$  on  $x$ . The sums of squares make this interpretation precise. Recall that  $SST = SSM + SSE$ . It is an algebraic fact that

interpretation of  $r^2$

$$r^2 = \frac{SSM}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Because  $SST$  is the total variation in  $y$  and  $SSM$  is the variation due to the regression of  $y$  on  $x$ , this equation is the precise statement of the fact that  $r^2$  is the fraction of variation in  $y$  explained by  $x$  in the linear regression.

## The ANOVA $F$ test

The null hypothesis  $H_0: \beta_1 = 0$  that  $y$  is not linearly related to  $x$  can be tested by comparing  $MSM$  with  $MSE$ . The ANOVA test statistic is an  **$F$  statistic**,

$F$  statistic

$$F = \frac{MSM}{MSE}$$

When  $H_0$  is true, this statistic has an  $F$  distribution with 1 degree of freedom in the numerator and  $n - 2$  degrees of freedom in the denominator. These degrees of freedom are those of  $MSM$  and  $MSE$ . Just as there are many  $t$  statistics, there are many  $F$  statistics. The ANOVA  $F$  statistic is not the same as the  $F$  statistic of equality of spread.



When  $\beta_1 \neq 0$ , MSM tends to be large relative to MSE. So large values of  $F$  are evidence against  $H_0$  in favor of the two-sided alternative.

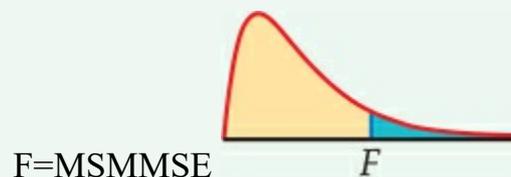
## ANALYSIS OF VARIANCE $F$ TEST

In the simple linear regression model, the hypotheses

$$H_0 \beta_1 = 0$$

$$H_a \beta_1 \neq 0$$

are tested by the  $F$  statistic



The  $P$ -value is the probability that a random variable having the  $F(1, n - 2)$  distribution is greater than or equal to the calculated value of the  $F$  statistic.

The  $F$  statistic tests the same null hypothesis as one of the  $t$  statistics that we encountered earlier in this chapter, so it is not surprising that the two are related. It is an algebraic fact that  $t^2 = F$  in this case. For linear regression with one explanatory variable, we prefer the  $t$  form of the test because it more easily allows us to test one-sided alternatives and is closely related to the confidence interval for  $\beta_1$ .

The ANOVA calculations are displayed in an *analysis of variance table*, often abbreviated **ANOVA table**. Here is the format of the table for simple linear regression:

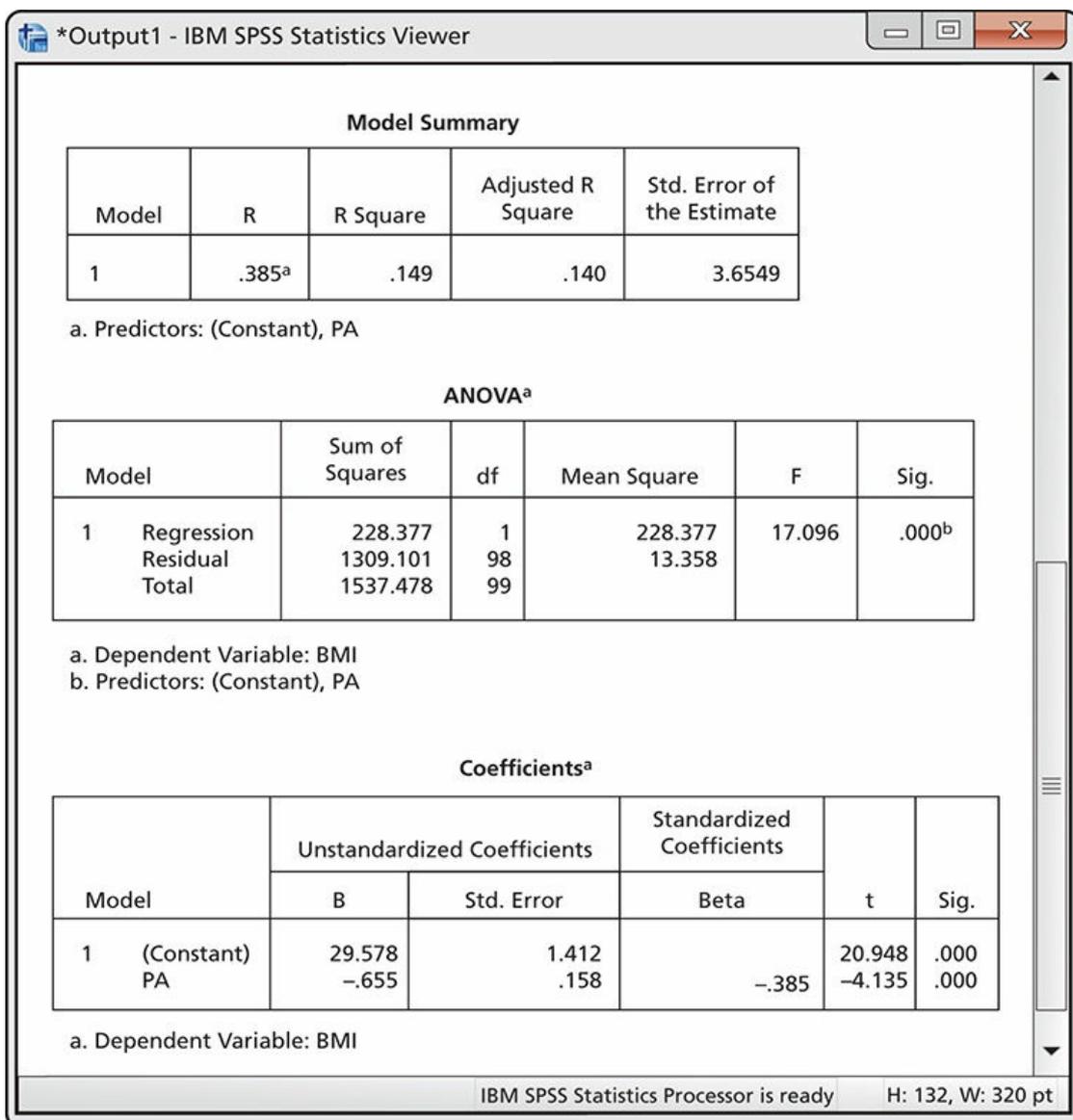
### ANOVA table

Degrees	Source of freedom	Sum of squares	Mean square	$F$
Model	1	$\Sigma(\hat{y}_i - \bar{y})^2$	SSM/DFM	MSM/MSE
Error	$n - 2$	$\Sigma(y_i - \hat{y}_i)^2$	SSE/DFE	
Total	$n - 1$	$\Sigma(y_i - \bar{y})^2$	SST/DFT	

## Example

### 10.15 Interpreting SPSS output for BMI and physical activity.

The entire output generated by SPSS for the physical activity study in Example 10.3 is given in Figure 10.12. Note that SPSS uses the labels Regression, Residual, and Total for the three sources of variation. We have called these Model, Error, and Total. Other statistical software packages may use slightly different labels. We round the calculated value of the  $F$  statistic to 17.10; the  $P$ -value is given as 0.000. This is a rounded value and we can conclude that  $P < 0.0005$ .



**FIGURE 10.12**

Regression output with ANOVA table for Example 10.15.



There is strong evidence against the null hypothesis that there is no relationship between BMI and average number of steps per day (PA). Now look at the output for the regression coefficients. The  $t$  statistic for PA is given as  $-4.135$ . If we square this number, we obtain the  $F$  statistic (accurate up to roundoff error). The value of  $r^2$  is also given in the output. Average number of steps per day explains only 14.9% of the variability in BMI. *Strong evidence against the null hypothesis that there is no relationship does not imply that a large percentage of the total variability is explained by the model.*

## Calculations for regression inference

We recommend using statistical software for regression calculations. With time and care, however, the work is feasible with a calculator. We will use the following example to illustrate how to perform inference for regression analysis using a calculator.

### Example

#### 10.16 Protein requirements via nitrogen balance.

Nitrogen balance studies are used to determine protein requirements for people. Each subject is fed three different controlled diets during three separate experimental periods. The three diets are similar with regard to all nutrients except protein.

Nitrogen balance is the difference between the amount of nitrogen consumed and the amount lost in feces and urine and by other means. Since virtually all the nitrogen in a diet comes from protein, nitrogen balance is an indicator of the amount of protein retained by the body. The protein requirement for an individual is the protein intake corresponding to a nitrogen balance of zero.

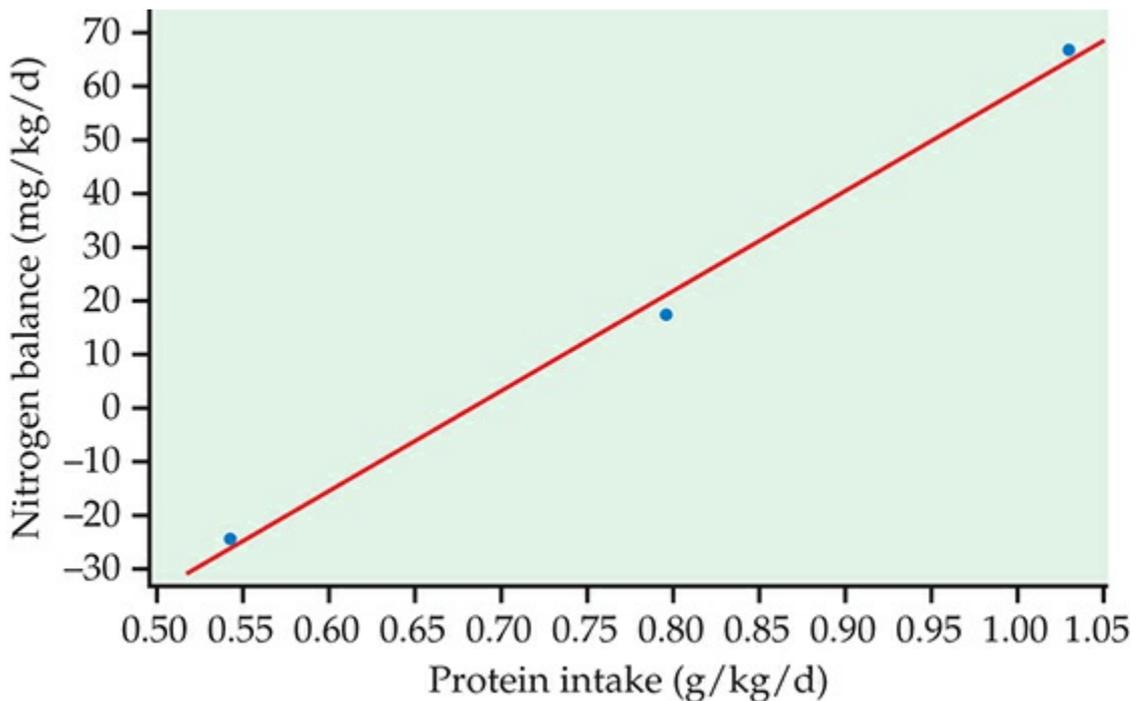
Linear regression is used to model the relationship between nitrogen balance, measured in milligrams of nitrogen per kilogram of body weight per day (mg/kg/d), and protein intake, measured in grams of protein per kilogram of body weight per day (g/kg/d). Here are the data for one subject:<sup>5</sup>

Protein intake ( $x$ )	0.543	0.797	1.030
Nitrogen balance ( $y$ )	-23.4	17.8	67.3

The data and the least-squares line are plotted in Figure 10.13. The strong straight-line pattern suggests that we can use linear regression to model the relationship between nitrogen balance and protein intake.



We begin our regression calculations by fitting the least-squares line. Fitting the line gives estimates  $b_1$  and  $b_0$  of the model parameters  $\beta_1$  and  $\beta_0$ . Next we examine the residuals from the fitted line and obtain an estimate  $s$  of the remaining parameter  $\sigma$ . These calculations are preliminary to inference. Finally, we use  $s$  to obtain the standard errors needed for the various interval estimates and significance tests. *Roundoff errors that accumulate during these calculations can ruin the final results. Be sure to carry many significant digits and check your work carefully.*



**FIGURE 10.13**

Scatterplot and regression line for Example 10.16.

### Preliminary calculations

After examining the scatterplot (Figure 10.13) to verify that the data show a straight-line pattern, we begin our calculations.

## Example

### 10.17 Summary statistics for nitrogen balance study.

We start by making a table with the mean and standard deviation for each of the variables, the correlation, and the sample size. These calculations should be familiar from Chapters 1 and 2. Here is the summary:

Variable	Mean	Standard deviation	Correlation	Sample size
Intake	$\bar{x}=0.79000$	$s_x = 0.24357545$	$r = 0.99698478$	$n = 3$
N balance	$\bar{y} = 20.56667$	$s_y = 45.4132506$		

These quantities are the building blocks for our calculations.

We will need one additional quantity for the calculations to follow. It is the expression  $\sum(x_i - \bar{x})^2$ . We obtain this quantity as an intermediate step when we calculate  $s_x$ . You could also find it using the fact that  $\sum(x_i - \bar{x})^2 = (n-1)s_x^2$ . You should verify that the value for our example is

$$\sum(x_i - \bar{x})^2 = 0.118658$$

Our first task is to find the least-squares line. This is easy with the building blocks that we have assembled.

## Example

### 10.18 Computing the least-squares regression line.

The slope of the least-squares line is

$$\begin{aligned} b_1 &= r s_y / s_x \\ &= 0.99698478 \cdot 45.4132506 / 0.24357545 \\ &= 185.882 \end{aligned}$$

The intercept is

$$\begin{aligned} b_0 &= \bar{y} - b_1 \bar{x} \\ &= 20.56667 - (185.882)(0.79000) \end{aligned}$$

$$= -126.280$$

The equation of the least-squares regression line is therefore

$$\hat{y} = -126.280 + 185.882x$$

This is the line shown in Figure 10.13.

We now have estimates of the first two parameters,  $\beta_0$  and  $\beta_1$ , of our linear regression model. Next, we find the estimate of the third parameter,  $\sigma$ : the standard deviation  $s$  about the fitted line. To do this we need to find the predicted values and then the residuals.

## Example

### 10.19 Computing the predicted values and residuals.

The first observation is an intake of  $x = 0.543$ . The corresponding predicted value of nitrogen balance is

$$\begin{aligned}\hat{y}_1 &= b_0 + b_1x_1 \\ &= -126.280 + (185.882)(0.543) \\ &= -25.346\end{aligned}$$

and the residual is

$$\begin{aligned}e_1 &= y_1 - \hat{y}_1 \\ &= -23.4 - (-25.346) \\ &= 1.946\end{aligned}$$

The residuals for the other intakes are calculated in the same way. You should verify that they are  $-4.068$  and  $2.122$ .

Notice that the sum of these three residuals is zero. When doing these calculations by hand, it is always helpful to check that the sum of the residuals is zero.

## Example

### 10.20 Computing $s^2$ .

The estimate of  $\sigma^2$  is  $s^2$ , the sum of the squares of the residuals divided by  $n - 2$ . The estimated standard deviation about the line is the square root of this quantity.

$$\begin{aligned}s^2 &= \frac{e_i^2}{n-2} \\ &= \frac{(1.946)^2 + (-4.068)^2 + (2.122)^2}{21} \\ &= 24.838\end{aligned}$$

So the estimate of the standard deviation about the line is

$$s = \sqrt{24.838} = 4.984$$

Now that we have estimates of the three parameters of our model, we can proceed to the more detailed calculations needed for regression inference.

#### Inference for slope and intercept

Confidence intervals and significance tests for the slope  $\beta_1$  and intercept  $\beta_0$  of the population regression line make use of the estimates  $b_1$  and  $b_0$  and their standard errors.

Some algebra and the rules for variances establishes that the standard deviation of  $b_1$  is



rules for variances, p. 275

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Similarly, the standard deviation of  $b_0$  is

$$\sigma_{b_0} = \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

To estimate these standard deviations, we need only replace  $\sigma$  by its estimate  $s$ .

## STANDARD ERRORS FOR ESTIMATED REGRESSION COEFFICIENTS

The **standard error of the slope**  $b_1$  of the least-squares regression line is

$$SE_{b1} = s \sqrt{\frac{1}{\sum (x_i - \bar{x})^2}}$$

The **standard error of the intercept**  $b_0$  is

$$SE_{b0} = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

The plot of the regression line with the data in Figure 10.13 shows a very strong relationship, but our sample size is very small. We assess the situation with a significance test for the slope.

## Example

**10.21 Testing the slope. First we need the standard error of the estimated slope:**

$$\begin{aligned} SE_{b1} &= s \sqrt{\frac{1}{\sum (x_i - \bar{x})^2}} \\ &= 4.9840118658 \\ &= 14.469 \end{aligned}$$

To test

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

calculate the  $t$  statistic:

$$\begin{aligned} t &= \frac{b_1}{SE_{b1}} \\ &= \frac{185.88214469}{14.469} = 12.847 \end{aligned}$$

Using Table D with  $n - 2 = 1$  degree of freedom, we conclude that  $P < 0.05$ . (The exact value obtained from software is 0.0494.) The data provide evidence in favor of a relationship between nitrogen balance and protein intake ( $t = 12.85$ ,  $df = 1$ ,  $P < 0.05$ ).

Three things are important to note about this example. First, the sample size is very small. Even though the estimated slope is more than 12 standard deviations away from zero, we have only barely attained the 0.05 standard for statistical

significance. *It is important to remember that we need to have a very large effect if we expect to detect it with a small sample size.* Second, we would, of course, prefer to have more than three observations for this analysis. However, for each protein intake, data are collected for about a month in order to calculate the nitrogen balance. Because of this expense of time and money, researchers typically use only three levels of intake. Third, because we expect balance to increase with increasing intake, a one-sided significance test is justified in this setting.



The significance test tells us that the data provide sufficient information to conclude that intake and balance are related. We use the estimate  $b_1$  and its confidence interval to further describe the relationship.

## Example

### 10.22 Computing a 95% confidence interval for the slope.

For the protein requirement problem, let's find a 95% confidence interval for the slope  $\beta_1$ . The degrees of freedom are  $n - 2 = 1$ , so  $t^*$  from Table D is 12.706. We compute

$$\begin{aligned} b_1 \pm t^* SE_{b_1} &= 185.882 \pm (12.706)(14.469) \\ &= 185.882 \pm 183.843 \end{aligned}$$

The interval is (2, 370).

Note the effect of the small sample size on the critical value  $t^*$ . With one additional observation, it would decrease to 4.303.

In this example, the intercept  $\beta_0$  does not have a meaningful interpretation. A protein intake of zero is theoretically possible, but we would not expect our linear model to be reasonable when extended to such an extreme value. For problems where inference for  $\beta_0$  is appropriate, the calculations are performed in the same way as those for  $\beta_1$ . Note that there is a different formula for the standard error, however.

**Confidence intervals for the mean response and prediction intervals for a future observation**

When we substitute a particular value  $x^*$  of the explanatory variable into the regression equation and obtain a value of  $\hat{y}$ , we can view the result in two ways:

1. We have estimated the mean response  $\mu_y$ .
2. We have predicted a future value of the response  $y$ .

The margins of error for these two uses are often quite different. Prediction intervals for an individual response are wider than confidence intervals for estimating a mean response. We now proceed with the details of these calculations. Once again, standard errors are the essential quantities. And once again, these standard errors are multiples of  $s$  our basic measure of the variability of the responses about the fitted line.

### STANDARD ERRORS FOR $\hat{\mu}$ AND $\hat{y}$

The standard error of  $\hat{\mu}$  is

$$SE_{\hat{\mu}} = s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

The standard error for predicting an individual response  $\hat{y}$  is

$$SE_{\hat{y}} = s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

Note that the only difference between the formulas for these two standard errors is the extra 1 under the square root sign in the standard error for prediction. This standard error is larger due to the additional variation of individual responses about the mean response. This additional variation remains regardless of the sample size  $n$  and is the reason that prediction intervals are wider than the confidence intervals for the mean response.

For the nitrogen balance example, we can think about the mean balance that would result if a particular protein intake was consumed many times. The confidence interval for the mean response would provide an interval estimate of this population value. On the other hand, we might want to predict a future observation under conditions similar to those used in the study, that is, for a one-month period, at a particular intake level. A prediction interval attempts to capture this future observation.

### Example

### 10.23 Computing a confidence interval for $\mu^{\wedge}$ .

Let's find a 95% confidence interval for the mean balance corresponding to an intake of 0.7 g/kg/d. The estimated mean balance is

$$\begin{aligned}\mu^{\wedge} &= b_0 + b_1 x_1 \\ &= -126.280 + (185.882)(0.7) \\ &= 3.837\end{aligned}$$

The standard error is

$$\begin{aligned}SE_{\mu^{\wedge}} &= s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \\ &= 4.98413 + (0.70 - 0.79) 20.118658\end{aligned}$$

To find the 95% confidence interval we compute

$$\begin{aligned}\mu^{\wedge} \pm t^* SE_{\mu^{\wedge}} &= 3.837 \pm (12.706)(3.158) \\ &= 3.837 \pm 40.126 \\ &= 4 \pm 40\end{aligned}$$

The interval is -36 to 44 mg/kg/d of nitrogen.

Calculations for the prediction intervals are similar. The only difference is the use of the formula for  $SE_{\hat{y}}$  in place of  $SE_{\mu^{\wedge}}$ . This results in a much wider interval.

Since the confidence interval for mean response includes the value 0, the corresponding intake 0.7 g/kg/d should be considered as a possible value for the intake requirement for this individual. Other intakes would also produce confidence intervals that include the value of 0 for mean balance. Here is one method that is commonly used to determine a single value of the requirement for an individual.

## Example

### 10.24 Estimating the protein requirement.

We define the estimated requirement for an individual to be the intake corresponding to zero balance using the fitted regression equation. To do this, we set the equation

$$\hat{\mu} = b_0 + b_1x$$

equal to 0 and solve for the intake  $x$ . So,

$$\begin{aligned}x &= -b_0/b_1 \\ &= -(-126.280)/185.882 \\ &= 0.68\end{aligned}$$

The estimated protein requirement for this individual is 0.68 g/kg/d.

If we repeat these calculations using data collected on a large number of individuals, we can estimate the requirement distribution for a population. There are many interesting statistical issues related to this problem, including estimating non-Normal population distributions.<sup>6</sup>

## Inference for correlation

The correlation coefficient is a measure of the strength and direction of the *linear* association between two variables. Correlation does not require an explanatory-response relationship between the variables. We can consider the sample correlation  $r$  as an estimate of the correlation in the population and base inference about the population correlation on  $r$ .



correlation, p. 103

The correlation between the variables  $x$  and  $y$  when they are measured for every member of a population is the **population correlation**. As usual, we use Greek letters to represent population parameters. In this case  $\rho$  (the Greek letter rho) is the population correlation.

### population correlation

When  $\rho = 0$ , there is no linear association in the population. In the important case where the two variables  $x$  and  $y$  are both Normally distributed, the condition  $\rho = 0$  is equivalent to the statement that  $x$  and  $y$  are independent. That is, there is no association of any kind between  $x$  and  $y$ . (Technically, the condition required is that  $x$  and  $y$  be **jointly Normal**. This means that the distribution of  $x$  is Normal and also that the conditional distribution of  $y$ , given any fixed value of  $x$  is Normal.) We therefore may wish to test the null hypothesis that a population correlation is 0.

### jointly Normal variables

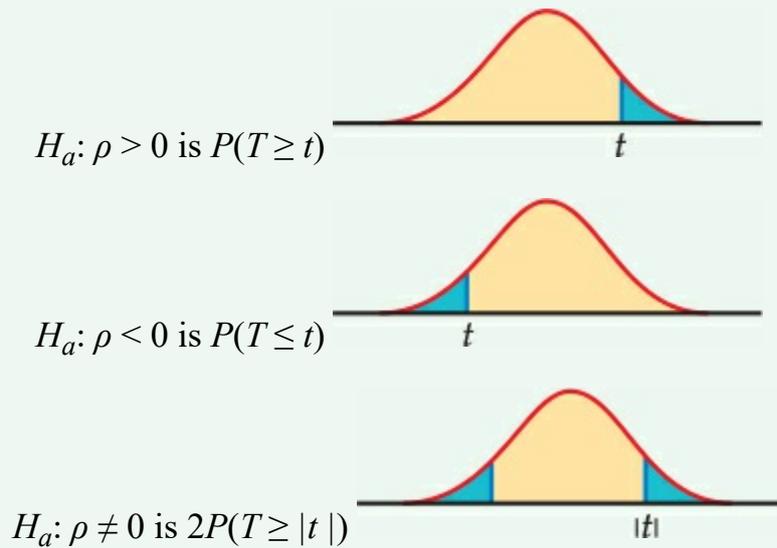
## TEST FOR A ZERO POPULATION CORRELATION

To test the hypothesis  $H_0: \rho = 0$  compute the ***t* statistic**

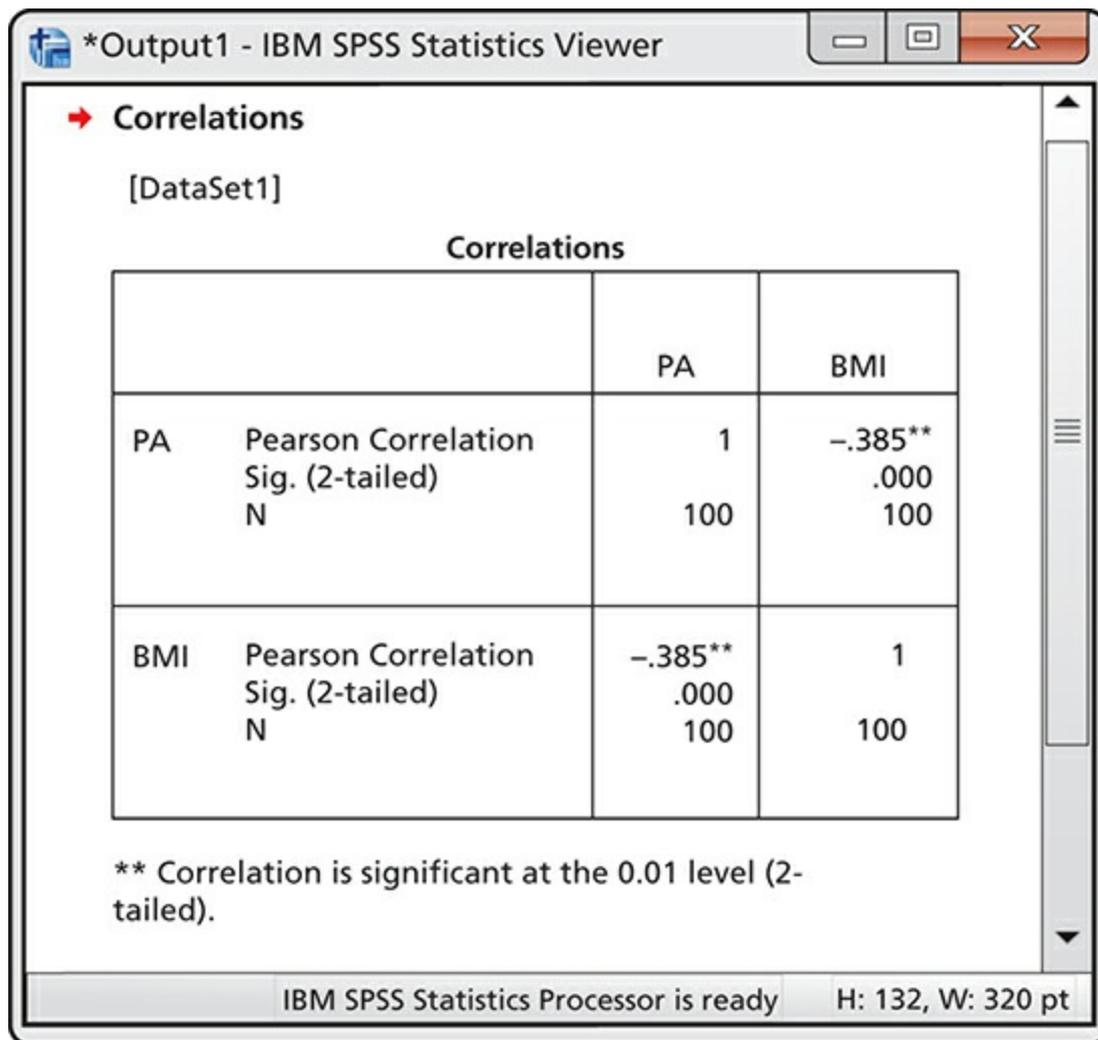
$$t = \frac{r\sqrt{n-2}}{1-r^2}$$

where  $n$  is the sample size and  $r$  is the sample correlation.

In terms of a random variable  $T$  having the  $t(n-2)$  distribution, the  $P$ -value for a test of  $H_1$  against



Most computer packages have routines for calculating correlations, and some will provide the significance test for the null hypothesis that  $\rho$  is zero.



**FIGURE 10.14**  
Correlation output for Example 10.25.

## Example

### 10.25 Correlation in the physical activity study.

The SPSS output for the physical activity example appears in Figure 10.14. The sample correlation between BMI and the average number of steps per day (PA) is  $r = -0.385$ . SPSS calls this a Pearson correlation to distinguish it from other kinds of correlations that it can calculate. The  $P$ -value for a two-sided test of  $H_0: \rho = 0$  is given as 0.000. This means that the actual  $P$ -value is less than 0.0005. We conclude that there is a nonzero correlation between BMI and PA.

If we wanted to test the one-sided alternative that the population correlation is negative, we divide the  $P$ -value in the output by 2, after checking that the sample coefficient is in fact negative.

If your software does not give the significance test, you can do the computations easily with a calculator.

## Example

### 10.26 Correlation test using a calculator.

The correlation between BMI and PA is  $r = -0.385$ . Recall that  $n = 100$ . The  $t$  statistic for testing the null hypothesis that the population correlation is zero is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \\ = \frac{-0.385}{\sqrt{\frac{1-(-0.385)^2}{100-2}}}$$

The degrees of freedom are  $n - 2 = 98$ . From Table D we conclude that  $P < 0.0001$ . This agrees with the SPSS output in Figure 10.14, where the  $P$ -value is given as 0.000. The data provide clear evidence that BMI and PA are related.

There is a close connection between the significance test for a correlation and the test for the slope in a linear regression. Recall that

$$b_1 = r \frac{S_y}{S_x}$$

From this fact we see that if the slope is 0, so is the correlation, and vice versa. It should come as no surprise to learn that the procedures for testing  $H_0: \beta_1 = 0$  and  $H_0: \rho = 0$  are also closely related. In fact, the  $t$  statistics for testing these hypotheses are numerically equal. That is,

$$t_{b_1} = t_r = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Check that this holds in both of our examples.

In our examples, the conclusion that there is a statistically significant correlation between the two variables would not come as a surprise to anyone familiar with the meaning of these variables. The significance test simply tells us whether or not there is evidence in the data to conclude that the population correlation is different from 0. The actual size of the correlation is of considerably more interest. We would therefore like to give a confidence interval for the population correlation. Unfortunately, most software packages do not perform this calculation. Because hand calculation of the confidence interval is very tedious, we

do not give the method here.<sup>7</sup>

## USE YOUR KNOWLEDGE

### 10.7 Research and development spending.



NSF

The National Science Foundation collects data on the research and development spending by universities and colleges in the United States.<sup>8</sup> Here are the data for the years 2003, 2006, and 2009:

Year	2003	2006	2009
Spending (billions of dollars)	40.1	47.8	54.9

Do the following by hand or with a calculator and verify your results with a software package.

- Make a scatterplot that shows the increase in research and development spending over time. Does the pattern suggest that the spending is increasing linearly over time?
- Find the equation of the least-squares regression line for predicting spending from year. Add this line to your scatterplot.
- For each of the three years, find the residual. Use these residuals to calculate the standard error  $s$ .
- Write the regression model for this setting. What are your estimates of the unknown parameters in this model?
- Compute a 95% confidence interval for the slope and summarize what this interval tells you about the increase in spending over time.

## SECTION 10.2 Summary

The **ANOVA table** for a linear regression gives the degrees of freedom, sum of squares, and mean squares for the model, error, and total sources of variation. The **ANOVA  $F$  statistic** is the ratio  $MSM/MSE$ . Under  $H_0: \beta_1 = 0$ , this statistic has an  $F(1, n - 2)$  distribution and is used to test  $H_0$  versus the two-sided alternative.

The **square of the sample correlation** can be expressed as

$$r^2 = \frac{SSM}{SST}$$

and is interpreted as the proportion of the variability in the response variable  $y$  that is explained by the explanatory variable  $x$  in the linear regression.

The **standard errors for  $b_0$  and  $b_1$**  are

$$SEb_0 = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}$$

$$SEb_1 = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}$$

The **standard error that we use for a confidence interval** for the estimated mean response for the subpopulation corresponding to the value  $x^*$  of the explanatory variable is

$$SE\hat{\mu} = s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

The **standard error that we use for a prediction interval** for a future observation from the subpopulation corresponding to the value  $x^*$  of the explanatory variable is

$$SE\hat{y} = s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

When the variables  $y$  and  $x$  are jointly Normal, the sample correlation is an estimate of the **population correlation  $\rho$** . The test of  $H_0: \rho = 0$  is based on the  **$t$  statistic**

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

which has a  $t(n - 2)$  distribution under  $H_0$ . This test statistic is numerically identical to the  $t$  statistic used to test  $H_0: \beta_1 = 0$ .

## CHAPTER 10 Exercises

For Exercises 10.1 and 10.2, see page 573; for Exercises 10.3 and 10.4, see page 576; for Exercises 10.5 and 10.6, see page 580; and for Exercise 10.7, see page 599.

### 10.8 What's wrong?

For each of the following, explain what is wrong and why.

- (a) The slope describes the change in  $x$  for a change in  $y$ .
- (b) The population regression line is  $y = b_0 + b_1x$ .
- (c) A 95% confidence interval for the mean response is the same width regardless of  $x$ .

### 10.9 What's wrong?

For each of the following, explain what is wrong and why.

- (a) The parameters of the simple linear regression model are  $b_0$ ,  $b_1$  and  $s$ .
- (b) To test  $H_0: b_1 = 0$  use a  $t$  test.
- (c) For a particular value of the explanatory variable  $x$ , the confidence interval for the mean response will be wider than the prediction interval for a future observation.

### 10.10 College debt versus the percent of students who borrow.

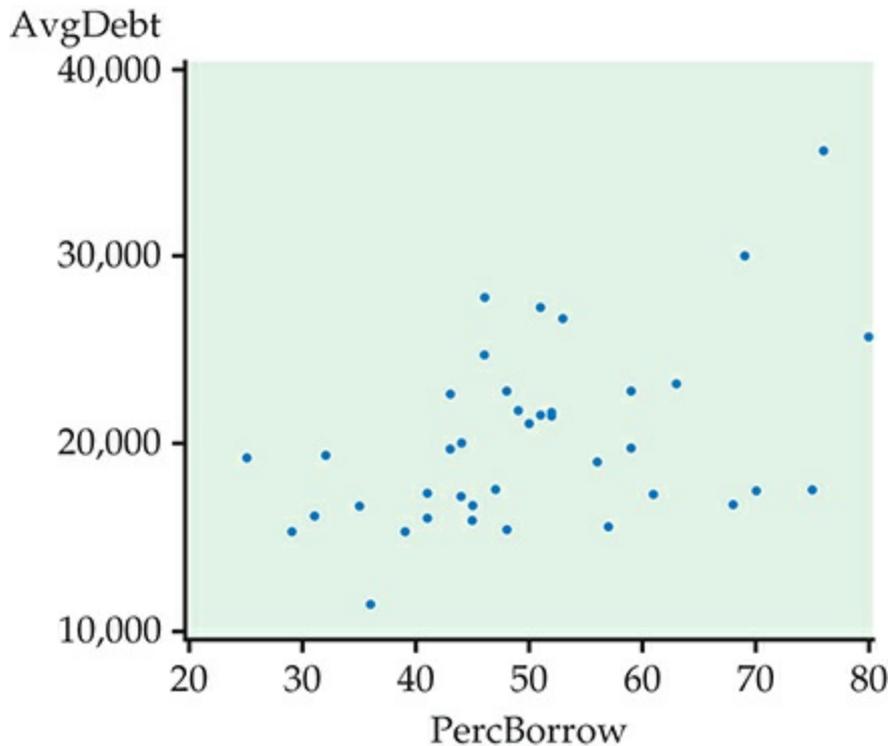
Kiplinger's "Best Values in Public Colleges" provides a ranking of U.S. public colleges based on a combination of various measures of academics and affordability.<sup>9</sup> We'll consider a random collection of 40 colleges from Kiplinger's 2011–2012 report and focus on the average debt in dollars at graduation (AvgDebt) and the percent of students who borrow (PercBorrow). 

- (a) A scatterplot of these two variables is shown in Figure 10.15. Describe the relationship. Are there any possible outliers or unusual values? Does a linear relationship between PercBorrow and AvgDebt seem reasonable?
- (b) Based on the scatterplot, approximately how much does the average debt change for a college with 10% more students who borrow?
- (c) The State University of New York–Fredonia is a school where 86% of the students borrow. Discuss the appropriateness of using this data set to predict the average debt for this school.

### 10.11 Can we consider this an SRS?

Refer to the previous exercise. The report states that Kiplinger's rankings focus on traditional four-year public colleges with broad-based curricula. Each year, they start with more than 500 schools

and then narrow the list down to roughly 120 based on academic quality before ranking them. The data set in the previous exercise is an SRS from their published list of 100 schools. As far as investigating the relationship between average debt and the percent of students who borrow, is it reasonable to consider this to be an SRS from the population of interest? Write a short paragraph explaining your answer. 



**FIGURE 10.15**

Scatterplot of average debt (in dollars) at graduation (AvgDebt) versus the percent of students who borrow (PercBorrow), for Exercise 10.10.

### 10.12 Predicting college debt.

Refer to Exercise 10.10. Figure 10.16 contains partial SAS output for the simple linear regression of AvgDebt on PercBorrow. 

- State the least-squares regression line.
- Construct a 95% confidence interval for the slope. What does this interval tell you about the change in average debt for a change in the percent who borrow?
- At Miami University, 51% of the students borrow, and the average debt is \$27,315. What is the residual?

The image shows a SAS window with two tables. The first table is titled 'Analysis of Variance' and the second is 'Parameter Estimates'. Both tables are presented in a standard SAS output format with columns for source/variable, degrees of freedom, sum of squares/parameter estimate, mean square/standard error, F/t value, and p-value.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	187482771	187482771	10.23	0.0028
Error	38	696437936	18327314		
Corrected Total	39	883920707			

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	11818	2739.85172	4.31	0.0001
PercBorrow	PercBorrow	1	168.98446	52.83425	3.20	0.0028

**FIGURE 10.16**

SAS output for Exercise 10.12.

### 10.13 More on predicting college debt.

Refer to the previous exercise. The University of Michigan–Ann Arbor is a school where 46% of the students borrow, and the average debt is \$27,828. The University of Wisconsin–La Crosse is a school where 69% of the students borrow, and the average debt is \$21,420. **BESTVAL**

- Using your answer to part (a) of the previous exercise, what is the predicted average debt for a student at the University of Michigan–Ann Arbor?
- What is the predicted average debt for the University of Wisconsin–La Crosse?
- Without doing any calculations, would the standard error for the estimated average debt be larger for the University of Michigan–Ann Arbor or the University of Wisconsin–La Crosse? Explain your answer.

### 10.14 Predicting college debt: other measures.

Refer to Exercise 10.10. Let’s now look at AvgDebt and its relationship with all seven measures available in the data set. In addition to the percent of students who borrow (PercBorrow), we have the admittance rate (Admit), the four-year graduation rate (Yr4Grad), in-state tuition after aid (InAfterAid), out-of-state tuition after aid (OutAfterAid), average aid per student (AvgAid), and the number of students per faculty member (StudPerFac). **BESTVAL**

- Generate scatterplots of each explanatory variable and AvgDebt. Do all these relationships look linear? Describe what you see.
- Fit each of the predictors separately and create a table that lists the explanatory variable, model standard deviation  $s$ , and the  $P$ -value for the test of a linear association.

(c) Which variable appears to be the best single predictor of average debt? Explain your answer.

### 10.15 Importance of Normal model deviations?

A general form of the central limit theorem tells us that the sampling distributions of  $b_0$  and  $b_1$  will be approximately Normal even if the model deviations are not Normally distributed. Using this fact, explain why the Normal distribution assumption is much more important for a prediction interval than for the confidence interval of the mean response at  $x = x^*$ .

### 10.16 Public university tuition: 2008 versus 2011.

Table 10.1 shows the in-state undergraduate tuition and required fees for 33 public universities in 2008 and 2011.  **TUITION**

(a) Plot the data with the 2008 in-state tuition (IN08) on the  $x$  axis and the 2011 tuition (IN11) on the  $y$  axis. Describe the relationship. Are there any outliers or unusual values? Does a linear relationship between the in-state tuition in 2008 and in 2011 seem reasonable?

(b) Run the simple linear regression and state the least-squares regression line.

(c) Obtain the residuals and plot them versus the 2008 in-state tuition amounts. Describe anything unusual in the plot.

(d) Do the residuals appear to be approximately Normal with constant variance? Explain your answer.

(e) The 5 California schools appear to follow the same linear trend as the other schools but have higher-than-predicted in-state tuition in 2011. Assume that this jump is particular to this state (financial troubles?), and remove these 5 observations and refit the model. How do the model parameters change?

(f) If you were to move forward with inference, which of these two model fits would you use? Write a short paragraph explaining your answer.

### 10.17 More on public university tuition.

Refer to the previous exercise. We'll now move forward with inference using the model fit you chose in part (f) of the previous exercise.  **TUITION**

**TABLE 10.1** In-State Tuition and Fees (in dollars) for 33 Public Universities

School	2008	2011	School	2008	2011	School	2008	2011
Penn State	13,706	15,984	Pittsburgh	13,642	16,132	Michigan	11,738	12,634
Rutgers	11,540	12,754	Michigan State	10,214	12,202	Maryland	8,005	8,655
Illinois	12,106	13,838	Minnesota	10,756	13,022	Missouri	8,467	8,989
Buffalo	6,285	7,482	Indiana	8,231	9,524	Ohio State	8,679	9,735
Virginia	9,300	11,786	Cal-Davis	8,635	13,860	Cal-Berkeley	7,656	12,834

Cal-Irvine	8,046	13,122	Purdue	7,750	9,478	Cal-San Diego	8,062	13,200
Oregon	6,435	8,789	Wisconsin	7,564	9,665	Washington	6,802	10,574
UCLA	7,551	12,686	Texas	8,532	9,794	Nebraska	6,584	7,563
Iowa	6,544	7,765	Colorado	7,278	9,152	Iowa State	6,360	7,486
North Carolina	5,397	7,009	Kansas	7,042	9,222	Arizona	5,542	9,286
Florida	3,778	5,657	Georgia Tech	6,040	9,652	Texas A&M	7,844	8,421

- (a) Give the null and alternative hypotheses for examining the linear relationship between 2008 and 2011 in-state tuition amounts.
- (b) Write down the test statistic and  $P$ -value for the hypotheses stated in part (a). State your conclusions.
- (c) Construct a 95% confidence interval for the slope. What does this interval tell you about the annual percent increase in tuition between 2008 and 2011?
- (d) What percent of the variability in 2011 tuition is explained by a linear regression model using the 2008 tuition?
- (e) Explain why inference on  $\beta_0$  is not of interest for this problem.

### 10.18 Even more on public university tuition.

Refer to the previous two exercises.  **TUITION**

- (a) The in-state tuition at State U was \$5100 in 2008. What is the predicted in-state tuition in 2011?
- (b) The in-state tuition at Moneypit U was \$15,700 in 2008. What is its predicted in-state tuition in 2011?
- (c) Discuss the appropriateness of using the fitted equation to predict tuition for each of these universities.

### 10.19 Out-of-state tuition.

Refer to Exercise 10.16. In addition to in-state tuition, out-of-state tuition for 2008 (OUT08) and 2011 (OUT11) was also obtained. Repeat parts (a) through (d) of Exercise 10.16 using these tuition rates. Does it appear we can use all the schools for this analysis or are there some unusual observations? Explain your answer.  **TUITION**

### 10.20 More on out-of-state tuition.

Refer to the previous exercise.  **TUITION**

- (a) Construct a 95% confidence interval for the slope. What does this interval tell you about the annual percent increase in out-of-state tuition between 2008 and 2011?
- (b) In Exercise 10.17(c) you constructed a similar 95% confidence interval for the annual percent increase in in-state tuition. Suppose that you want to test whether the increase is the same for both

tuition types. Given the two slope estimates  $b_1$  and standard errors, could we just do a variation of the two-sample  $t$  test from Chapter 7? Explain why or why not.

### 10.21 In-state versus out-of-state tuition.

Refer to the previous five exercises. We can also investigate whether there is a linear association between the in-state and out-of-state tuition. Perform a linear regression analysis using the 2011 data, complete with scatterplots and residual checks, and write a paragraph summarizing your findings.  **TUITION**

### 10.22 U.S. versus overseas stock returns.

Returns on common stocks in the United States and overseas appear to be growing more closely correlated as economies become more interdependent. Suppose that the following population regression line connects the total annual returns (in percent) on two indexes of stock prices:

$$\text{MEAN OVERSEAS RETURN} = -0.2 + 0.32 \times \text{U.S. RETURN}$$

- (a) What is  $\beta_0$  in this line? What does this number say about overseas returns when the U.S. market is flat (0% return)?
- (b) What is  $\beta_1$  in this line? What does this number say about the relationship between U.S. and overseas returns?
- (c) We know that overseas returns will vary in years when U.S. returns do not vary. Write the regression model based on the population regression line given above. What part of this model allows overseas returns to vary when U.S. returns remain the same?

### 10.23 Beer and blood alcohol.

How well does the number of beers a student drinks predict his or her blood alcohol content (BAC)? Sixteen student volunteers at Ohio State University drank a randomly assigned number of 12-ounce cans of beer. Thirty minutes later, a police officer measured their BAC. Here are the data:<sup>11</sup>

Student	1	2	3	4	5	6	7	8
Beers	5	2	9	8	3	7	3	5
BAC	0.10	0.03	0.19	0.12	0.04	0.095	0.07	0.06
Student	9	10	11	12	13	14	15	16
Beers	3	5	4	6	5	7	1	4
BAC	0.02	0.05	0.07	0.10	0.085	0.09	0.01	0.05

The students were equally divided between men and women and differed in weight and usual drinking habits. Because of this variation, many students don't believe that number of drinks predicts BAC well.  **BAC**

- (a) Make a scatterplot of the data. Find the equation of the least-squares regression line for predicting BAC from number of beers and add this line to your plot. What is  $r^2$  for these data? Briefly summarize what your data analysis shows.

(b) Is there significant evidence that drinking more beers increases BAC on the average in the population of all students? State hypotheses, give a test statistic and  $P$ -value, and state your conclusion.

(c) Steve thinks he can drive legally 30 minutes after he drinks 5 beers. The legal limit is  $BAC = 0.08$ . Give a 90% prediction interval for Steve's BAC. Can he be confident he won't be arrested if he drives and is stopped?

### 10.24 School budget and number of students.

Suppose that there is a linear relationship between the number of students  $x$  in a school system and the annual budget  $y$ . Write a population regression model to describe this relationship.

(a) Which parameter in your model is the fixed cost in the budget (for example, the salary of the principals and some administrative costs) that does not change as  $x$  increases?

(b) Which parameter in your model shows how total cost changes when there are more students in the system? Do you expect this number to be greater than 0 or less than 0?

(c) Actual data from various school systems will not fit a straight line exactly. What term in your model allows variation among schools of the same size  $x$ ?

### 10.25 Performance bonuses.

In the National Football League (NFL), performance bonuses now account for roughly 25% of player compensation.<sup>12</sup> Does tying a player's salary into performance bonuses result in better individual or team success on the field? Focusing on linebackers, let's look at the relationship between a player's end-of-year production rating and the percent of his salary devoted to incentive payments in that same year.  **PERFPAY**

(a) Use numerical and graphical methods to describe the two variables and summarize your results.

(b) Both variable distributions are non-Normal. Does this necessarily pose a problem for performing linear regression? Explain.

(c) Construct a scatterplot of the data and describe the relationship. Are there any outliers or unusual values? Does a linear relationship between the percent of salary and the player rating seem reasonable? Is it a very strong relationship? Explain.

(d) Run the simple linear regression and state the least-squares regression line.

(e) Obtain the residuals and assess whether the assumptions for the linear regression analysis are reasonable. Include all plots and numerical summaries used in doing this assessment.

### 10.26 Performance bonuses, continued.

Refer to the previous exercise.  **PERFPAY**

(a) Now run the simple linear regression for the variables  $\sqrt{\text{rating}}$  and percent of salary devoted to incentive payments.

(b) Obtain the residuals and assess whether the assumptions for the linear regression analysis are reasonable. Include all plots and numerical summaries used in doing this assessment.

(c) Construct a 95% confidence interval for the square root increase in rating given a 1% increase in the percent of salary devoted to incentive payments.

(d) Consider the values 0%, 20%, 40%, 60%, and 80% salary devoted to incentives. Compute the predicted rating for this model and for the one in the previous exercise. For the model in this problem, you will need to square the predicted value to get back to the original units.

(e) Plot the predicted values versus the percent and connect those values from the same model. For which regions of percent do the predicted values from the two models differ the most?

(f) Based on the comparison of regression models (both predicted values and residuals), which model do you prefer? Explain.

### 10.27 Sales price versus assessed value.

Real estate is typically reassessed annually for property tax purposes. This assessed value, however, is not necessarily the same as the fair market value of the property. Table 10.2 summarizes an SRS of 30 homes recently sold in a midwestern city.<sup>13</sup> Both variables are measured in thousands of dollars.  SALES

**TABLE 10.2** Sales Price and Assessed Value (in \$ thousands) of 30 Homes in a Midwestern City

Property	Sales price	Assessed value	Property	Sales price	Assessed value	Property	Sales price	Assessed value
1	179.9	188.7	2	240.0	220.4	3	113.5	118.1
4	281.5	232.4	5	186.0	188.1	6	275.0	240.1
7	281.5	232.4	8	210.0	211.8	9	210.0	168.0
10	184.0	180.3	11	186.5	294.7	12	239.0	209.2
13	185.0	162.3	14	251.0	236.8	15	180.0	123.7
16	160.0	191.7	17	255.0	245.6	18	220.0	19.3
19	160.0	181.6	20	200.0	177.4	21	265.0	307.2
22	190.0	229.7	23	150.5	168.9	24	189.0	194.4
25	157.0	143.9	26	171.5	201.4	27	157.0	143.9
28	175.0	181.0	29	159.0	125.1	30	229.0	195.3

(a) Inspect the data. How many homes have a sales price greater than the assessed value? Do you think this trend would be true for the larger population of all homes recently sold? Explain your answer.

(b) Make a scatterplot with assessed value on the horizontal axis. Briefly describe the relationship between assessed value and sales price.

(c) Report the least-squares regression line for predicting sales price from assessed value.

(d) Obtain the residuals and plot them versus assessed value. Property 11 was sold at a price substantially lower than the assessed value. Does this observation appear to be unusual in the residual plot? Approximately how many standard deviations is it away from its predicted value?

(e) Remove this observation and redo the least-squares fit. How have the least-squares regression line and model standard deviation changed?

(f) Check the residuals for this new fit. Do the assumptions for the linear regression analysis appear reasonable here? Explain your answer.

### 10.28 Sales price versus assessed value, continued.

Refer to the previous exercise. Let's consider the model fit with Property 11 excluded.  SALES

- (a) Calculate the predicted sales prices for homes currently assessed at \$155,000, \$220,000, and \$285,000.
- (b) Construct a 95% confidence interval for the slope and explain what this model tells you in terms of the relationship between assessed value and sales price.
- (c) Explain why inference on the intercept is not of interest.
- (d) Using the result from part (b), compare the estimated regression line with  $y = x$ , which says that, on average, the sales price is equal to the assessed value. Is there evidence that this model is not reasonable? In other words, is the sales price typically larger or smaller than the assessed value? Explain your answer.

### 10.29 Is the number of tornadoes increasing?

The Storm Prediction Center of the National Oceanic and Atmospheric Administration maintains a database of tornadoes, floods, and other weather phenomena. Table 10.3 summarizes the annual number of tornadoes in the United States between 1953 and 2012.<sup>14</sup>  TWISTER

- (a) Make a plot of the total number of tornadoes by year. Does a linear trend over years appear reasonable? Are there any outliers or unusual patterns? Explain your answer.
- (b) Run the simple linear regression and summarize the results, making sure to construct a 95% confidence interval for the average annual increase in the number of tornadoes.
- (c) Obtain the residuals and plot them versus year. Is there anything unusual in the plot?
- (d) Are the residuals Normal? Justify your answer.
- (e) The number of tornadoes in 2004 is much larger than expected under this linear model. Also, the number of tornadoes in 2012 is much smaller than predicted. Remove these observations and rerun the simple linear regression. Compare these results with the results in part (b). Do you think these two observations should be considered outliers and removed? Explain your answer.

### 10.30 Are the two fuel efficiency measurements similar?

**TABLE 10.3** Annual Number of Tornadoes in the United States Between 1953 and 2012

Year	Number of tornadoes						
1953	421	1968	660	1983	931	1998	1449
1954	550	1969	608	1984	907	1999	1340
1955	593	1970	653	1985	684	2000	1075

1956	504	1971	888	1986	764	2001	1215
1957	856	1972	741	1987	656	2002	934
1958	564	1973	1102	1988	702	2003	1374
1959	604	1974	947	1989	856	2004	1817
1960	616	1975	920	1990	1133	2005	1265
1961	697	1976	835	1991	1132	2006	1103
1962	657	1977	852	1992	1298	2007	1096
1963	464	1978	788	1993	1176	2008	1692
1964	704	1979	852	1994	1082	2009	1156
1965	906	1980	866	1995	1235	2010	1282
1966	585	1981	783	1996	1173	2011	1692
1967	926	1982	1046	1997	1148	2012	939

Refer to Exercise 7.30 (page 443). In addition to the computer calculating miles per gallon (mpg), the driver also recorded this measure by dividing the miles driven by the number of gallons at fill-up. The driver wants to determine if these calculations are different.  **MPGDIFF**

Fill-up	1	2	3	4	5	6	7	8	9	10
Computer	41.5	50.7	36.6	37.3	34.2	45.0	48.0	43.2	47.7	42.2
Driver	36.5	44.2	37.2	35.6	30.5	40.5	40.0	41.0	42.8	39.2
Fill-up	11	12	13	14	15	16	17	18	19	20
Computer	43.2	44.6	48.4	46.4	46.8	39.2	37.3	43.5	44.3	43.3
Driver	38.8	44.5	45.4	45.3	45.7	34.2	35.2	39.8	44.9	47.5

- Consider the driver's mpg calculations as the explanatory variable. Plot the data and describe the relationship. Are there any outliers or unusual values? Does a linear relationship seem reasonable?
- Run the simple linear regression and state the least-squares regression line.
- Summarize the results. Does it appear that the computer and driver calculations are the same? Explain.

### 10.31 Gambling and alcohol use by first-year college students.

Gambling and alcohol use are problematic behaviors for many college students. One study looked at 908 first-year students from a large northeastern university.<sup>15</sup> Each participant was asked to fill out the 10-item Alcohol Use Disorders Identification Test (AUDIT) and a 7-item inventory used in prior gambling research among college students. AUDIT assesses alcohol consumption and other alcohol-related risks and problems (a higher score means more risks). A correlation of 0.29 was reported between the frequency of gambling and the AUDIT score.

- What percent of the variability in AUDIT score is explained by frequency of gambling?
- Test the null hypothesis that the correlation between the gambling frequency and the AUDIT score is zero.
- The sample in this study represents 45% of the students contacted for the online study. To what extent do you think these results apply to all first-year students at this university? To what extent do you think these results apply to all first-year students? Give reasons for your answers.

### 10.32 Predicting water quality.

The index of biotic integrity (IBI) is a measure of the water quality in streams. IBI and land use measures for a collection of streams in the Ozark Highland ecoregion of Arkansas were collected as part of a study.<sup>16</sup> Table 10.4 gives the data for IBI, the percent of the watershed that was forest, and the area of the watershed in square kilometers for streams in the original sample with watershed area less than or equal to 70 km<sup>2</sup>  IBI

- Use numerical and graphical methods to describe the variable IBI. Do the same for area. Summarize your results.
- Plot the data and describe the relationship between IBI and area. Are there any outliers or unusual patterns?
- Give the statistical model for simple linear regression for this problem.
- State the null and alternative hypotheses for examining the relationship between IBI and area.
- Run the simple linear regression and summarize the results.
- Obtain the residuals and plot them versus area. Is there anything unusual in the plot?
- Do the residuals appear to be approximately Normal? Give reasons for your answer.
- Do the assumptions for the analysis of these data using the model you gave in part (c) appear to be reasonable? Explain your answer.

### 10.33 More on predicting water quality.

The researchers who conducted the study described in the previous exercise also recorded the percent of the watershed area that was forest for each of the streams.

Area	Forest	IBI	Area	Fc									
21	0	47	29	0	61	31	0	39	32	0	59	34	
34	0	76	49	3	85	52	3	89	2	7	74	70	
6	9	33	28	10	46	21	10	32	59	11	80	69	
47	17	78	8	17	53	8	18	43	58	21	88	54	
10	25	62	57	31	55	18	32	29	19	33	29	39	
49	33	78	9	39	71	5	41	55	14	43	58	9	
23	47	33	31	49	59	18	49	81	16	52	71	21	
32	59	64	10	63	41	26	68	82	9	75	60	54	
12	79	83	21	80	82	27	86	82	23	89	86	26	
16	95	67	26	95	56	26	100	85	28	100	91		

These data are also given in Table 10.4. Analyze these data using the questions in the previous exercise as a guide.  IBI

### 10.34 Comparing the analyses.

In Exercises 10.32 and 10.33, you used two different explanatory variables to predict IBI. Summarize the two analyses and compare the results. If you had to choose between the two explanatory variables for predicting IBI, which one would you prefer? Give reasons for your answer.



### 10.35 How an outlier can affect statistical significance.

Consider the data in Table 10.4 and the relationship between IBI and the percent of watershed area that was forest. The relationship between these two variables is almost significant at the 0.05 level. In this exercise you will demonstrate the potential effect of an outlier on statistical significance. Investigate what happens when you decrease the IBI to 0.0 for (1) an observation with 0% forest and (2) an observation with 100% forest. Write a short summary of what you learn from this exercise.



### 10.36 Predicting water quality for an area of 40 km<sup>2</sup>.

Refer to Exercise 10.32. IBI

- Find a 95% confidence interval for the mean response corresponding to an area of 40 km<sup>2</sup>.
- Find a 95% prediction interval for a future response corresponding to an area of 40 km<sup>2</sup>.
- Write a short paragraph interpreting the meaning of the intervals in terms of Ozark Highland streams.
- Do you think that these results can be applied to other streams in Arkansas or in other states? Explain why or why not.

### 10.37 Compare the predictions.

Consider Case 37 in Table 10.4 (8th row, 2nd column). For this case the area is 10 km<sup>2</sup> and the percent forest is 63%. A predicted index of biotic integrity based on area was computed in Exercise 10.32, while one based on percent forest was computed in Exercise 10.33. Compare these two estimates and explain why they differ. Use the idea of a prediction interval to interpret these results.



### 10.38 Reading test scores and IQ.

In Exercise 2.33 (page 100) you examined the relationship between reading test scores and IQ scores for a sample of 60 fifth-grade children. READIQ

- Run the regression and summarize the results of the significance tests.
- Rerun the analysis with the four possible outliers removed. Summarize your findings, paying particular attention to the effects of removing the outliers.

### 10.39 Leaning Tower of Pisa.

The Leaning Tower of Pisa is an architectural wonder. Engineers concerned about the tower's stability have done extensive studies of its increasing tilt. Measurements of the lean of the tower over time provide much useful information. The following table gives measurements for the years 1975 to 1987. The variable "lean" represents the difference between where a point on the tower would be if the tower were straight and where it actually is. The data are coded as tenths of a millimeter in excess of 2.9 meters, so that the 1975 lean, which was 2.9642 meters, appears in the table as 642. Only the last two digits of the year were entered into the computer.<sup>17</sup>  PISA

Year	75	76	77	78	79	80	81	82	83	84	85	86	87
Lean	642	644	656	667	673	688	696	698	713	717	725	742	757

- (a) Plot the data. Does the trend in lean over time appear to be linear?
- (b) What is the equation of the least-squares line? What percent of the variation in lean is explained by this line?
- (c) Give a 99% confidence interval for the average rate of change (tenths of a millimeter per year) of the lean.

### 10.40 More on the Leaning Tower of Pisa.

Refer to the previous exercise.  PISA

- (a) In 1918 the lean was 2.9071 meters. (The coded value is 71.) Using the least-squares equation for the years 1975 to 1987, calculate a predicted value for the lean in 1918. (Note that you must use the coded value 18 for year.)
- (b) Although the least-squares line gives an excellent fit to the data for 1975 to 1987, this pattern did not extend back to 1918. Write a short statement explaining why this conclusion follows from the information available. Use numerical and graphical summaries to support your explanation.

### 10.41 Predicting the lean in 2013.

Refer to the previous two exercises.  PISA

- (a) How would you code the explanatory variable for the year 2013?
- (b) The engineers working on the Leaning Tower of Pisa were most interested in how much the tower would lean if no corrective action was taken. Use the least-squares equation to predict the tower's lean in the year 2013. (*Note:* The tower was renovated in 2001 to make sure it does not fall down.)
- (c) To give a margin of error for the lean in 2013, would you use a confidence interval for a mean response or a prediction interval? Explain your choice.

### 10.42 Correlation between binge drinking and the average price of beer.

A recent study looked at 118 colleges to investigate the association between the binge-drinking rate and the average price for a bottle of beer at establishments within a two-mile radius of campus.<sup>18</sup> A correlation of  $-0.36$  was found. Explain this correlation.

### 10.43 Is this relationship significant?

Refer to the previous exercise. Test the null hypothesis that the correlation between the binge-drinking rate and the average price for a bottle of beer within a two-mile radius of campus is zero.

### 10.44 Does a math pretest predict success?

Can a pretest on mathematics skills predict success in a statistics course? The 62 students in an introductory statistics class took a pretest at the beginning of the semester. The least-squares regression line for predicting the score  $y$  on the final exam from the pretest score  $x$  was  $\hat{y} = 13.8 + 0.81x$ . The standard error of  $b_1$  was 0.43.

(a) Test the null hypothesis that there is no linear relationship between the pretest score and the score on the final exam against the two-sided alternative.

(b) Would you reject this null hypothesis versus the one-sided alternative that the slope is positive? Explain your answer.

### 10.45 Completing an ANOVA table.

How are returns on common stocks in overseas markets related to returns in U.S. markets? Consider measuring U.S. returns by the annual rate of return on the Standard & Poor's 500 stock index and overseas returns by the annual rate of return on the Morgan Stanley Europe, Australasia, Far East (EAFE) index.<sup>19</sup> Both are recorded in percents. We will regress the EAFE returns on the S&P 500 returns for the years 1993 to 2012. Here is part of the Minitab output for this regression:

```
The regression equation is
EAFE = - 0.168 + 0.845 S&P
Analysis of Variance
Source          DF      SS      MS      F
Regression          1    4947.2
Residual Error     18
Total              19   8251.5
```

Using the ANOVA table format on page 589 as a guide, complete the analysis of variance table.

### 10.46 Interpreting statistical software output.

Refer to the previous exercise. What are the values of the regression standard error  $s$  and the squared correlation  $r^2$ ?

### 10.47 Standard error and confidence interval for the slope.

Refer to the previous two exercises. The standard deviation of the S&P 500 returns for these years is 19.09%. From this and your work in the previous exercise, find the standard error for the least-squares slope  $b_1$ . Give a 95% confidence interval for the slope  $\beta_1$  of the population regression line.

### 10.48 Grade inflation.

The average undergraduate GPA for American colleges and universities was estimated based on a sample of institutions that published this information.<sup>20</sup> Here are the data for public schools in that report:

Year	1992	1996	2002	2007
GPA	2.85	2.90	2.97	3.01

Do the following by hand or with a calculator and verify your results with a software package.  **GRADEUP**

- Make a scatterplot that shows the increase in GPA over time. Does a linear increase appear reasonable?
- Find the equation of the least-squares regression line for predicting GPA from year. Add this line to your scatterplot.
- Compute a 95% confidence interval for the slope and summarize what this interval tells you about the increase in GPA over time.

### 10.49 Significance test of the correlation.

A study reported a correlation  $r = 0.5$  based on a sample size of  $n = 15$ ; another reported the same correlation based on a sample size of  $n = 25$ . For each, perform the test of the null hypothesis that  $\rho = 0$ . Describe the results and explain why the conclusions are different.

### 10.50 State and college binge drinking.

Excessive consumption of alcohol is associated with numerous adverse consequences. In one study, researchers analyzed binge-drinking rates from two national surveys, the Harvard School of Public Health College Alcohol Study (CAS) and the Centers for Disease Control and Prevention's Behavioral Risk Factor Surveillance System (BRFSS).<sup>21</sup> The CAS survey was used to provide an estimate of the college binge-drinking rate in each state, and the BRFSS was used to determine the adult binge-drinking rate in each state. A correlation of 0.43 was reported between these two rates for their sample of  $n = 40$  states. The college binge-drinking rate had a mean of 46.5% and standard deviation 13.5%. The adult binge-drinking rate had a mean of 14.88% and standard deviation 3.8%.

- Find the equation of the least-squares line for predicting the college binge-drinking rate from the adult binge-drinking rate.
- Give the results of the significance test for the null hypothesis that the slope is 0. (*Hint*: What is the relation between this test and the test for a zero correlation?)

### 10.51 SAT versus ACT.

The SAT and the ACT are the two major standardized tests that colleges use to evaluate candidates. Most students take just one of these tests. However, some students take both. Consider the scores of 60 students who did this. How can we relate the two tests? 

- Plot the data with SAT on the  $x$  axis and ACT on the  $y$  axis. Describe the overall pattern and any unusual observations.
- Find the least-squares regression line and draw it on your plot. Give the results of the significance test for the slope.

(c) What is the correlation between the two tests?

### 10.52 SAT versus ACT, continued.

Refer to the previous exercise. Find the predicted value of ACT for each observation in the data set.

#### SATACT

(a) What is the mean of these predicted values? Compare it with the mean of the ACT scores.

(b) Compare the standard deviation of the predicted values with the standard deviation of the actual ACT scores. If least-squares regression is used to predict ACT scores for a large number of students such as these, the average predicted value will be accurate but the variability of the predicted scores will be too small.

(c) Find the SAT score for a student who is one standard deviation above the mean ( $z=(x-\bar{x})/s=1$ ). Find the predicted ACT score and standardize this score. (Use the means and standard deviations from this set of data for these calculations.)

(d) Repeat part (c) for a student whose SAT score is one standard deviation below the mean ( $Z=-1$ ).

(e) What do you conclude from parts (c) and (d)? Perform additional calculations for different  $z$ 's if needed.

### 10.53 Matching standardized scores.

Refer to the previous two exercises. An alternative to the least-squares method is based on matching standardized scores. Specifically, we set

$$(y-\bar{y})/s_y=(x-\bar{x})/s_x$$

and solve for  $y$ . Let's use the notation  $y = a_0 + a_1x$  for this line. The slope is  $a_1 = s_y/s_x$  and the intercept is  $a_0 = \bar{y} - a_1\bar{x}$ . Compare these expressions with the formulas for the least-squares slope and intercept (page 592).  SATACT

(a) Using the data in the previous exercise, find the values of  $a_0$  and  $a_1$ .

(b) Plot the data with the least-squares line and the new prediction line.

(c) Use the new line to find predicted ACT scores. Find the mean and the standard deviation of these scores. How do they compare with the mean and standard deviation of the ACT scores?

### 10.54 Weight, length, and width of perch.

Here are data for 12 perch caught in a lake in Finland.<sup>22</sup>  PERCH

Weight (grams)	Length (cm)	Width (cm)	Weight (grams)	Length (cm)	Width (cm)
5.9	8.8	1.4	300.0	28.7	5.1
100.0	19.2	3.3	300.0	30.1	4.6
110.0	22.5	3.6	685.0	39.0	6.9
120.0	23.5	3.5	650.0	41.4	6.0

150.0	24.0	3.6	820.0	42.5	6.6
145.0	25.5	3.8	1000.0	46.6	7.6

In this exercise we will examine different models for predicting weight.

(a) Plot weight versus length and weight versus width. Do these relationships appear to be linear? Explain your answer.

(b) Run the regression using length to predict weight. Do the same using width as the explanatory variable. Summarize the results. Be sure to include the value of  $r^2$ .

### 10.55 Transforming the perch data.

Refer to the previous exercise.  PERCH

(a) Try to find a better model using a transformation of length. One possibility is to use the square. Make a plot and perform the regression analysis. Summarize the results.

(b) Do the same for width.

### 10.56 Creating a new explanatory variable.

Refer to the previous two exercises.  PERCH

(a) Create a new variable that is the product of length and width. Make a plot and run the regression using this new variable. Summarize the results.

(b) Write a short report summarizing and comparing the different regression analyses that you performed in this exercise and the previous two exercises.

### 10.57 Index of biotic integrity.

Refer to the data on the index of biotic integrity and area in Exercise 10.32 (page 606) and the additional data on percent watershed area that was forest in Exercise 10.33. Find the correlations among these three variables, perform the test of statistical significance, and summarize the results. Which of these test results could have been obtained from the analyses that you performed in

Exercises 10.32 and 10.33?  IBI

### 10.58 Food neophobia.

Food neophobia is a personality trait associated with avoiding unfamiliar foods. In one study of 564 children who were 2 to 6 years of age, food neophobia and the frequency of consumption of different types of food were measured.<sup>23</sup> Here is a summary of the correlations:

Type of food	Correlation
Vegetables	-0.27
Fruit	-0.16
Meat	-0.15
Eggs	-0.08
Sweet/fatty snacks	0.04

Starchy staples	-0.02
-----------------	-------

Perform the significance test for each correlation and write a summary about food neophobia and the consumption of different types of food.

### 10.59 A mechanistic explanation of popularity.

Previous experimental work has suggested that the serotonin system plays an important and causal role in social status. In other words, genes may predispose individuals to be popular/likable. As part of a recent study on adolescents, an experimenter looked at the relationship between the expression of a particular serotonin receptor gene, a person’s “popularity,” and the person’s rule-breaking (RB) behaviors.<sup>24</sup> RB was measured by both a questionnaire and video observation. The composite score is an equal combination of these two assessments. Here is a table of the correlations:

Rule-breaking measure	Popularity	Gene expression
Sample 1 ( $n = 123$ )		
RB.composite	0.28	0.26
RB.questionnaire	0.22	0.23
RB.video	0.24	0.20
Sample 1 Caucasians only ( $n = 96$ )		
RB.composite	0.22	0.23
RB.questionnaire	0.16	0.24
RB.video	0.19	0.16

For each correlation, test the null hypothesis that the corresponding true correlation is zero. Reproduce the table and mark the correlations that have  $P < 0.001$  with \*\*\*, those that have  $P < 0.01$  with \*\*, and those that have  $P < 0.05$  with \*. Write a summary of the results of your significance tests.

### 10.60 Resting metabolic rate and exercise.

Metabolic rate, the rate at which the body consumes energy, is important in studies of weight gain, dieting, and exercise. The following table gives data on the lean body mass and resting metabolic rate for 12 women and 7 men who are subjects in a study of dieting. Lean body mass, given in kilograms, is a person’s weight leaving out all fat. Metabolic rate is measured in calories burned per 24 hours, the same calories used to describe the energy content of foods. The researchers believe that lean body mass is an important influence on metabolic rate. 

Subject	Sex	Mass	Rate	Subject	Sex	Mass	Rate
1	M	62.0	1792	11	F	40.3	1189
2	M	62.9	1666	12	F	33.1	913
3	F	36.1	995	13	M	51.9	1460
4	F	54.6	1425	14	F	42.4	1124
5	F	48.5	1396	15	F	34.5	1052
6	F	42.0	1418	16	F	51.1	1347
7	M	47.4	1362	17	F	41.2	1204
8	F	50.6	1502	18	M	51.9	1867
9	F	42.0	1256	19	M	46.9	1439
10	M	48.7	1614				

---

(a) Make a scatterplot of the data, using different symbols or colors for men and women. Summarize what you see in the plot.

(b) Run the regression to predict metabolic rate from lean body mass for the women in the sample and summarize the results. Do the same for the men.

### **10.61 Resting metabolic rate and exercise, continued.**

Refer to the previous exercise. It is tempting to conclude that there is a strong linear relationship for the women but no relationship for the men. Let's look at this issue a little more carefully. 

#### **METRATE**

(a) Find the confidence interval for the slope in the regression equation that you ran for the females. Do the same for the males. What do these suggest about the possibility that these two slopes are the same? (The formal method for making this comparison is a bit complicated and is beyond the scope of this chapter.)

(b) Examine the formula for the standard error of the regression slope given on page 593. The term in the denominator is  $\sum(x_i - \bar{x})^2$ . Find this quantity for the females; do the same for the males. How do these calculations help to explain the results of the significance tests?

(c) Suppose that you were able to collect additional data for males. How would you use lean body mass in deciding which subjects to choose?

### **10.62 Inference over different ranges of X.**

Think about what would happen if you analyzed a subset of a set of data by analyzing only data for a restricted range of values of the explanatory variable. What results would you expect to change? Examine your ideas by analyzing the fuel efficiency data described in Example 10.11 (page 581). First, run a regression of MPG versus MPH using all cases. This least-squares regression line is shown in Figure 10.9. Next run a regression of MPG versus MPH for only those cases with speed less than or equal to 30 mph. Note that this corresponds to 3.4 in the log scale. Finally, do the same analysis with a restriction on the response variable. Run the analysis with only those cases with fuel efficiency less than or equal to 20 mpg. Write a summary comparing the effects of these two restrictions with each other and with the complete data set results.  **MPHMPG**

# 11 Multiple Regression

CHAPTER



**11.1 Inference for Multiple Regression**

**11.2 A Case Study**

## Introduction

In Chapter 10 we presented methods for inference in the setting of a linear relationship between a response variable  $y$  and a *single* explanatory variable  $x$ . In this chapter, we use *more than one* explanatory variable to explain or predict a single response variable.

Many of the ideas that we encountered in our study of simple linear regression carry over to the multiple linear regression setting. For example, the descriptive tools we learned in Chapter 2—scatterplots, least-squares regression, and correlation—are still essential preliminaries to inference and also provide a foundation for confidence intervals and significance tests.

The introduction of several explanatory variables leads to many additional considerations. In this short chapter we cannot explore all these issues. Rather, we will outline some basic facts about inference in the multiple regression setting and then illustrate the analysis with a case study whose purpose was to predict success in college based on several high school achievement scores.

## 11.1 Inference for Multiple Regression

When you complete this section, you will be able to

- Describe the multiple linear regression model in terms of a population regression line and the deviations of the response variable  $y$  from this line.
- Interpret regression output from statistical software to obtain the least-squares regression equation and model standard deviation, multiple correlation coefficient, ANOVA  $F$  test, and individual regression coefficient  $t$  tests.
- Explain the difference between the ANOVA  $F$  test and the  $t$  tests for individual coefficients.
- Interpret a level  $C$  confidence interval or significance test for a regression coefficient.
- Use diagnostic plots to check the assumptions of the multiple linear regression model.

### Population multiple regression equation

The simple linear regression model assumes that the mean of the response variable  $y$  depends on the explanatory variable  $x$  according to a linear equation

$$\mu_y = \beta_0 + \beta_1 x$$

For any fixed value of  $x$  the response  $y$  varies Normally around this mean and has a standard deviation  $\sigma$  that is the same for all values of  $x$ .

In the multiple regression setting, the response variable  $y$  depends on  $p$  explanatory variables, which we will denote by  $x_1, x_2, \dots, x_p$ . The mean response depends on these explanatory variables according to a linear function

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

Similar to simple linear regression, this expression is the **population regression equation**, and the observed values  $y$  vary about their means given by this equation.

#### population regression equation

Just as we did in simple linear regression, we can also think of this model in terms of subpopulations of responses. Here, each subpopulation corresponds to a particular set of values for *all* the explanatory variables  $x_1, x_2, \dots, x_p$ . In each subpopulation,  $y$  varies Normally with a mean given by the population regression

equation. The regression model assumes that the standard deviation  $\sigma$  of the responses is the same in all subpopulations.

## EXAMPLE

### 11.1 Predicting early success in college.

Our case study is based on data collected on science majors at a large university.<sup>1</sup> The purpose of the study was to attempt to predict success in the early university years. One measure of success was the cumulative grade point average (GPA) after three semesters. Among the explanatory variables recorded at the time the students enrolled in the university were average high school grades in mathematics (HSM), science (HSS), and English (HSE).

We will use high school grades to predict the response variable GPA. There are  $p = 3$  explanatory variables:  $x_1 = \text{HSM}$ ,  $x_2 = \text{HSS}$ , and  $x_3 = \text{HSE}$ . The high school grades are coded on a scale from 1 to 10, with 10 corresponding to A, 9 to A-, 8 to B+, and so on. These grades define the subpopulations. For example, the straight-C students are the subpopulation defined by  $\text{HSM} = 4$ ,  $\text{HSS} = 4$ , and  $\text{HSE} = 4$ .

One possible multiple regression model for the subpopulation mean GPAs is

$$\mu_{\text{GPA}} = \beta_0 + \beta_1 \text{HSM} + \beta_2 \text{HSS} + \beta_3 \text{HSE}$$

For the straight-C subpopulation of students, the model gives the subpopulation mean as

$$\mu_{\text{GPA}} = \beta_0 + \beta_1 4 + \beta_2 4 + \beta_3 4$$

## Data for multiple regression

The data for a simple linear regression problem consist of observations  $(x_i, y_i)$  of the two variables. Because there are several explanatory variables in multiple regression, the notation needed to describe the data is more elaborate. Each observation or case consists of a value for the response variable and for each of the explanatory variables. Call  $x_{ij}$  the value of the  $j$ th explanatory variable for the  $i$ th case. The data are then

$$\text{Case 1: } (x_{11}, x_{12}, \dots, x_{1p}, y_1)$$

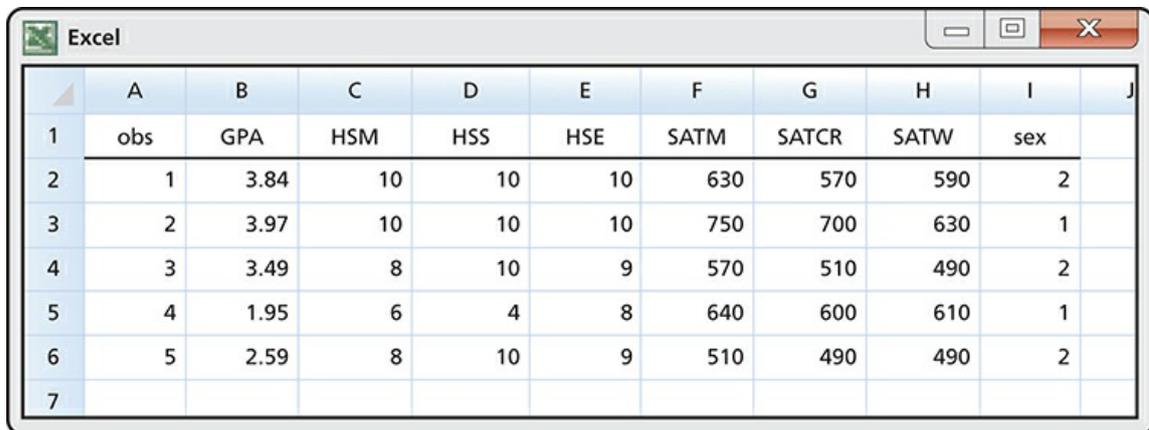
Case 2:  $(x_{21}, x_{22}, \dots, x_{2p}, y_2)$

⋮

Case  $n$ :  $(x_{n1}, x_{n2}, \dots, x_{np}, y_n)$

Here,  $n$  is the number of cases and  $p$  is the number of explanatory variables. Data are often entered into computer regression programs in this format. Each row is a case and each column corresponds to a different variable.

The data for Example 11.1, with several additional explanatory variables, appear in this format in the GPA data file. Figure 11.1 shows the first 5 rows entered into an Excel spreadsheet. Grade point average (GPA) is the response variable, followed by  $p = 7$  explanatory variables. There are a total of  $n = 150$  students in this data set.



	A	B	C	D	E	F	G	H	I	J
1	obs	GPA	HSM	HSS	HSE	SATM	SATCR	SATW	sex	
2	1	3.84	10	10	10	630	570	590	2	
3	2	3.97	10	10	10	750	700	630	1	
4	3	3.49	8	10	9	570	510	490	2	
5	4	1.95	6	4	8	640	600	610	1	
6	5	2.59	8	10	9	510	490	490	2	
7										

**FIGURE 11.1**

Format of data set for Example 11.1 in an Excel spreadsheet.

## USE YOUR KNOWLEDGE

### 11.1 Describing a multiple regression.

Traditionally, demographic and high school academic variables have been used to predict college academic success. One study investigated the influence of emotional health on GPA.<sup>2</sup> Data from 242 students who had completed their first two semesters of college were obtained. The researchers were interested in describing how students' second-semester grade point averages are explained by gender, a standardized test score, perfectionism, self-esteem, fatigue, optimism, and depressive symptomatology.

- (a) What is the response variable?
- (b) What is  $n$  the number of cases?
- (c) What is  $p$  the number of explanatory variables?
- (d) What are the explanatory variables?

## Multiple linear regression model

We combine the population regression equation and assumptions about variation to construct the multiple linear regression model. The subpopulation means describe the FIT part of our statistical model. The RESIDUAL part represents the variation of observations about the means.



**DATA = FIT + RESIDUAL, p. 567**

We will use the same notation for the residual that we used in the simple linear regression model. The symbol  $\varepsilon$  represents the deviation of an individual observation from its subpopulation mean.

We assume that these deviations are Normally distributed with mean 0 and an unknown model standard deviation  $\sigma$  that does not depend on the values of the  $x$  variables. *These are assumptions that we can check by examining the residuals in the same way that we did for simple linear regression.*



### MULTIPLE LINEAR REGRESSION MODEL

The **statistical model for multiple linear regression** is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i$$

for  $i = 1, 2, \dots, n$ .

The **mean response  $\mu_y$**  is a linear function of the explanatory variables:

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

The **deviations  $\varepsilon_i$**  are assumed to be independent and Normally distributed with mean 0 and standard deviation  $\sigma$ . In other words, they are an SRS from the  $N(0, \sigma)$  distribution.

The **parameters of the model** are  $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ , and  $\sigma$

The assumption that the subpopulation means are related to the regression coefficients  $\beta$  by the equation

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

implies that we can estimate all subpopulation means from estimates of the  $\beta$ 's. To the extent that this equation is accurate, we have a useful tool for describing how the mean of  $y$  varies with the collection of  $x$ 's.

We do, however, need to be cautious when interpreting each of the regression coefficients in a multiple regression. First, the  $\beta_0$  coefficient represents the mean of  $y$  when *all* the  $x$  variables equal zero. Even more so than in simple linear regression, this subpopulation is rarely of interest. Second, the description provided by the regression coefficient of each  $x$  variable is similar to that provided by the slope in simple linear regression but only in a specific situation, namely, *when all other  $x$  variables are held constant*. We need this extra condition because with multiple  $x$  variables, it is quite possible that a unit change in one  $x$  variable may be associated with changes in other  $x$  variables. If that occurs, then the change in the mean of  $y$  is not described by just a single regression coefficient.

## USE YOUR KNOWLEDGE

### 11.2 Understanding the fitted regression line.

The fitted regression equation for a multiple regression is

$$\hat{y} = -1.8 + 6.1x_1 - 1.1x_2$$

- (a) If  $x_1 = 3$  and  $x_2 = 1$  what is the predicted value of  $y$ ?
- (b) For the answer to part (a) to be valid, is it necessary that the values  $x_1 = 3$  and  $x_2 = 1$  correspond to a case in the data set? Explain why or why not.
- (c) If you hold  $x_2$  at a fixed value, what is the effect of an increase of two units in  $x_1$  on the predicted value of  $y$ ?

## Estimation of the multiple regression parameters



least squares, p. 113

Similar to simple linear regression, we use the method of least squares to obtain estimators of the regression coefficients  $\beta$ . The details, however, are more complicated. Let

$$b_0, b_1, b_2, \dots, b_p$$

denote the estimators of the parameters

$$\beta_0, \beta_1, \beta_2, \dots, \beta_p$$

For the  $i$ th observation, the predicted response is

$$\hat{y}_i = b_0 + b_1x_{i1} + b_2x_{i2} + \dots + b_px_{ip}$$



**residual, p. 569**

The  $i$ th residual, the difference between the observed and the predicted response, is therefore

$$e_i = \text{observed response} - \text{predicted response}$$

$$= y_i - \hat{y}_i$$

$$= y_i - b_0 - b_1x_{i1} - b_2x_{i2} - \dots - b_px_{ip}$$

The method of least squares chooses the values of the  $b$ 's that make the sum of the squared residuals as small as possible. In other words, the parameter estimates  $b_0, b_1, b_2, \dots, b_p$  minimize the quantity

$$\Sigma(y_i - b_0 - b_1x_{i1} - b_2x_{i2} - \dots - b_px_{ip})^2$$



*The formula for the least-squares estimates is complicated. We will be content to understand the principle on which it is based and to let software do the computations.*

The parameter  $\sigma^2$  measures the variability of the responses about the population regression equation. As in the case of simple linear regression, we estimate  $\sigma^2$  by an average of the squared residuals. The estimator is

$$s^2 = \frac{\Sigma e_i^2}{n-p-1}$$

$$= \frac{\Sigma (y_i - \hat{y}_i)^2}{n-p-1}$$

## ← LOOK BACK

### degrees of freedom, p. 44

The quantity  $n - p - 1$  is the degrees of freedom associated with  $s^2$ . The degrees of freedom equal the sample size,  $n$  minus  $p + 1$ , the number of  $\beta$ 's we must estimate to fit the model. In the simple linear regression case there is just one explanatory variable, so  $p = 1$  and the degrees of freedom are  $n - 2$ . To the model standard deviation  $\sigma$  we use

$$s = s^2$$

### Confidence intervals and significance tests for regression coefficients

We can obtain confidence intervals and perform significance tests for each of the regression coefficients  $\beta_j$  as we did in simple linear regression. The standard errors of the  $b$ 's have more complicated formulas, but all are multiples of  $s$ . We again rely on statistical software to do the calculations.

#### CONFIDENCE INTERVALS AND SIGNIFICANCE TESTS FOR $\beta_j$

A level  $C$  confidence interval for  $\beta_j$  is

$$b_j \pm t^* SE_{b_j}$$

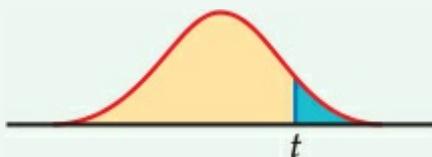
where  $SE_{b_j}$  is the standard error of  $b_j$  and  $t^*$  is the value for the  $t(n - p - 1)$  density curve with area  $C$  between  $-t^*$  and  $t^*$ .

To test the hypothesis  $H_0: \beta_j = 0$ , compute the  **$t$  statistic**

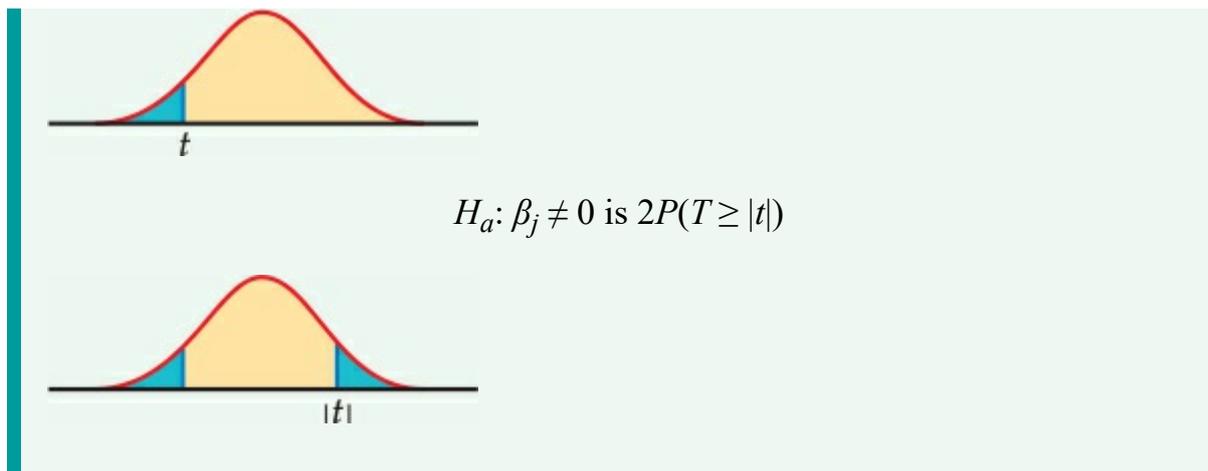
$$t = b_j / SE_{b_j}$$

In terms of a random variable  $T$  having the  $t(n - p - 1)$  distribution, the  $P$ -value for a test of  $H_0$  against

$$H_a: \beta_j > 0 \text{ is } P(T \geq t)$$



$$H_a: \beta_j < 0 \text{ is } P(T \leq t)$$



← LOOK BACK

**confidence intervals for mean response, p. 577**

**prediction intervals, p. 579**

Because regression is often used for prediction, we may wish to use multiple regression models to construct confidence intervals for a mean response and prediction intervals for a future observation. The basic ideas are the same as in the simple linear regression case.

In most software systems, the same commands that give confidence and prediction intervals for simple linear regression work for multiple regression. The only difference is that we specify a list of explanatory variables rather than a single variable. Modern software allows us to perform these rather complex calculations without an intimate knowledge of all the computational details. This frees us to concentrate on the meaning and appropriate use of the results.

### ANOVA table for multiple regression

In simple linear regression the  $F$  test from the ANOVA table is equivalent to the two-sided  $t$  test of the hypothesis that the slope of the regression line is 0. For multiple regression there is a corresponding ANOVA  $F$  test, but it tests the hypothesis that *all* the regression coefficients (with the exception of the intercept) are 0. Here is the general form of the ANOVA table for multiple regression:

← LOOK BACK

**ANOVA  $F$  test, p. 588**

Source	Degrees of freedom	Sum of squares	Mean square	$F$
Model	$p$	$\Sigma(\hat{y}_i - \bar{y})^2$	SSM/DFM	MSM/MSE
Error	$n - p - 1$	$\Sigma(y_i - \hat{y}_i)^2$	SSE/DFE	

Total	$n - 1$	$\Sigma(y_i - \bar{y})^2$	SST/DFT
-------	---------	---------------------------	---------

The ANOVA table is similar to that for simple linear regression. The degrees of freedom for the model increase from 1 to  $p$  to reflect the fact that we now have  $p$  explanatory variables rather than just one. As a consequence, the degrees of freedom for error decrease by the same amount. *It is always a good idea to calculate the degrees of freedom by hand and then check that your software agrees with your calculations. In this way you can verify that your software is using the number of cases and number of explanatory variables that you intended.*



The sums of squares represent sources of variation. Once again, both the sums of squares and their degrees of freedom add:

$$SST = SSM + SSE$$

$$DFT = DFM + DFE$$

The estimate of the variance  $\sigma^2$  for our model is again given by the MSE in the ANOVA table. That is,  $s^2 = \text{MSE}$ .



### ***F* statistic, p. 588**

The ratio  $\text{MSM}/\text{MSE}$  is an  $F$  statistic for testing the null hypothesis

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$$

against the alternative hypothesis

$$H_a: \text{at least one of the } \beta_j \text{ is not } 0$$

The null hypothesis says that none of the explanatory variables are predictors of the response variable when used in the form expressed by the multiple regression equation. The alternative states that *at least one* of them is a predictor of the response variable.

As in simple linear regression, large values of  $F$  give evidence against  $H_0$ . When  $H_0$  is true,  $F$  has the  $F(p, n - p - 1)$  distribution. The degrees of freedom for the  $F$  distribution are those associated with the model and error in the ANOVA table.



A common error in the use of multiple regression is to assume that all the regression coefficients are statistically different from zero whenever the  $F$  statistic has a small  $P$ -value. Be sure that you understand the difference between the  $F$  test and the  $t$  tests for individual coefficients.

### ANALYSIS OF VARIANCE $F$ Test

In the multiple regression model, the hypothesis

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$$

is tested against the alternative hypothesis

$$H_a: \text{at least one of the } \beta_j \text{ is not } 0$$

by the analysis of variance  **$F$  statistic**

$$F = \frac{MSM}{MSE}$$

The  $P$ -value is the probability that a random variable having the  $F(p, n - p - 1)$  distribution is greater than or equal to the calculated value of the  $F$  statistic.

### Squared multiple correlation $R^2$

For simple linear regression we noted that the square of the sample correlation could be written as the ratio of  $SSM$  to  $SST$  and could be interpreted as the proportion of variation in  $y$  explained by  $x$ . A similar statistic is routinely calculated for multiple regression.

### THE SQUARED MULTIPLE CORRELATION

The statistic

$$R^2 = \frac{SSM}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

is the proportion of the variation of the response variable  $y$  that is explained by the explanatory variables  $x_1, x_2, \dots, x_p$  in a multiple linear regression.

Often,  $R^2$  is multiplied by 100 and expressed as a percent. The square root of  $R^2$  called the **multiple correlation coefficient**, is the correlation between the observations  $y_i$  and the predicted values  $\hat{y}_i$ .

## USE YOUR KNOWLEDGE

### 11.3 Significance tests for regression coefficients.

As part of a study on undergraduate success among actuarial students a multiple regression was run using 82 students.<sup>3</sup> The following table contains the estimated coefficients and standard errors:

Variable	Estimate	SE
Intercept	-0.764	0.651
SAT Math	0.00156	0.00074
SAT Verbal	0.00164	0.00076
High school rank	1.470	0.430
College placement exam	0.889	0.402

- All the estimated coefficients for the explanatory variables are positive. Is this what you would expect? Explain.
- What are the degrees of freedom for the model and error?
- Test the significance of each coefficient and state your conclusions.

### 11.4 ANOVA table for multiple regression.

Use the following information and the general form of the ANOVA table for multiple regression on page 617 to perform the ANOVA  $F$  test and compute  $R^2$ .

Source	Degrees of freedom	Sum of squares
Model		75
Error	53	
Total	57	594

## 11.2 A Case Study

### Preliminary analysis

In this section we illustrate multiple regression by analyzing the data from the study described in Example 11.1. The response variable is the cumulative GPA, on a 4-point scale, after three semesters. The explanatory variables previously mentioned are average high school grades, represented by HSM, HSS, and HSE. We also examine the SAT Mathematics (SATM), SAT Critical Reading (SATCR), and SAT Writing (SATW) scores as explanatory variables. We have data for  $n = 150$  students in the study. We use SAS, Excel, and Minitab to illustrate the outputs that are given by most software.

The first step in the analysis is to carefully examine each of the variables. Means, standard deviations, and minimum and maximum values appear in Figure 11.2. The minimum value for high school mathematics (HSM) appears to be rather extreme; it is  $(8.59 - 2.00)/1.46 = 4.51$  standard deviations below the mean. Similarly, the minimum value for GPA is 3.43 standard deviations below the mean. We do not discard either of these cases at this time but will take care in our subsequent analyses to see if they have an excessive influence on our results.

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
GPA	GPA	150	2.8421333	0.8178992	0.0300000	4.0000000
SATM	SATM	150	623.6000000	74.8356589	460.0000000	800.0000000
SATCR	SATCR	150	573.8000000	87.6208274	330.0000000	800.0000000
SATW	SATW	150	562.6000000	80.0874522	350.0000000	770.0000000
HSM	HSM	150	8.5866667	1.4617571	2.0000000	10.0000000
HSS	HSS	150	8.8000000	1.3951017	4.0000000	10.0000000
HSE	HSE	150	8.8333333	1.2660601	4.0000000	10.0000000

**FIGURE 11.2**

Descriptive statistics for the College of Science student case study.

The mean for the SATM score is higher than the means for the Critical Reading (SATCR) and Writing (SATW) scores, as we might expect for a group of science majors. The three SAT standard deviations are all about the same.

Although mathematics scores were higher on the SAT, the means and standard deviations of the three high school grade variables are very similar. Since the level and difficulty of high school courses vary within and across schools, this may not be that surprising. The mean GPA is 2.842 on a 4-point scale, with standard deviation 0.818.

Because the variables GPA, SATM, SATCR, and SATW have many possible values, we could use stemplots or histograms to examine the shapes of their distributions. Normal quantile plots indicate whether or not the distributions look Normal. *It is important to note that the multiple regression model does not require any of these distributions to be Normal.* Only the deviations of the responses  $y$  from their means are assumed to be Normal.



The purpose of examining these plots is to understand something about each variable alone before attempting to use it in a complicated model. *Extreme values of any variable should be noted and checked for accuracy.* If found to be correct, the cases with these values should be carefully examined to see if they are truly exceptional and perhaps do not belong in the same analysis with the other cases. When our data on science majors are examined in this way, no obvious problems are evident.



The high school grade variables HSM, HSS, and HSE have relatively few values and are best summarized by giving the relative frequencies for each possible value. The output in Figure 11.3 provides these summaries. The distributions are all skewed, with a large proportion of high grades (10 = A and 9 = A-.) Again we emphasize that these distributions need not be Normal.

SAS

HSM				
HSM	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2	1	0.67	1	0.67
5	2	1.33	3	2.00
6	13	8.67	16	10.67
7	14	9.33	30	20.00
8	35	23.33	65	43.33
9	30	20.00	95	63.33
10	55	36.67	150	100.00

HSS				
HSS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
4	2	1.33	2	1.33
5	3	2.00	5	3.33
6	4	2.67	9	6.00
7	20	13.33	29	19.33
8	19	12.67	48	32.00
9	39	26.00	87	58.00
10	63	42.00	150	100.00

HSE				
HSE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
4	1	0.67	1	0.67
5	1	0.67	2	1.33
6	7	4.67	9	6.00
7	13	8.67	22	14.67
8	28	18.67	50	33.33
9	41	27.33	91	60.67
10	59	39.33	150	100.00

Done

### FIGURE 11.3

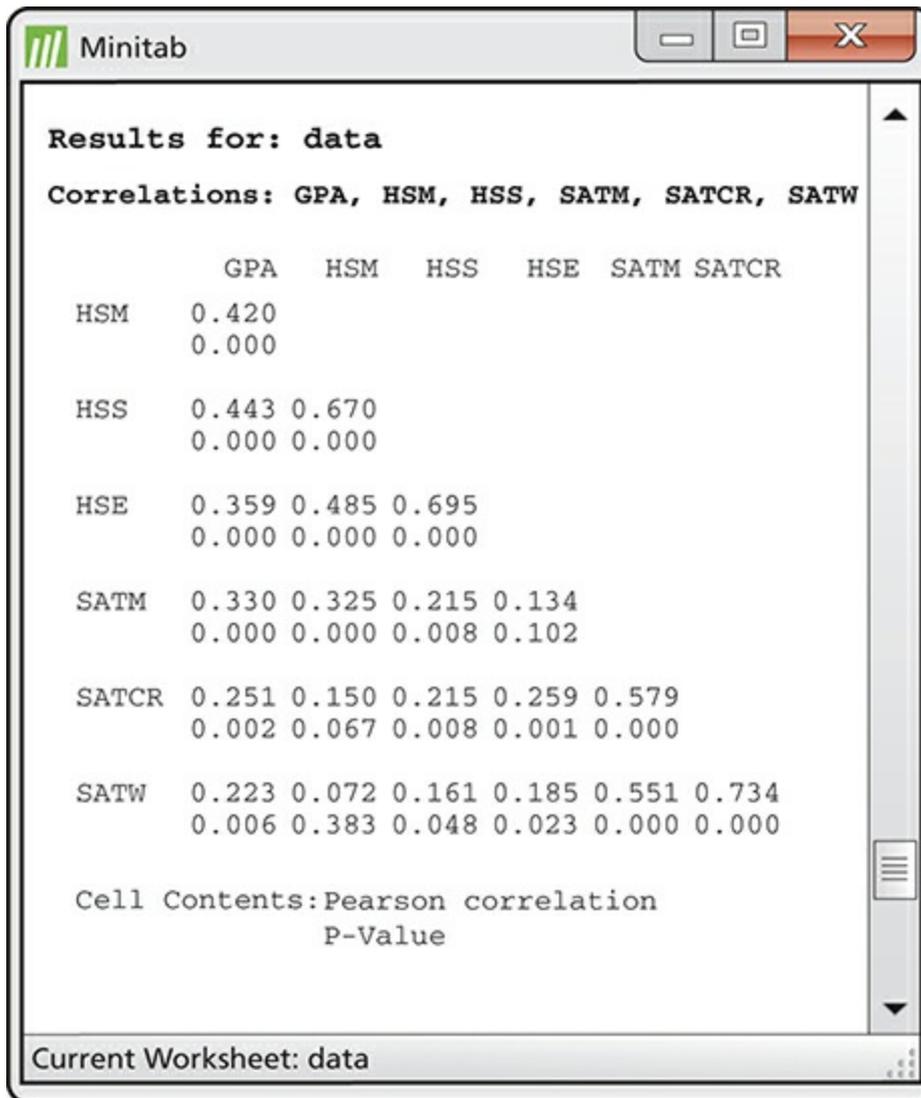
The distributions of the high school grade variables.

## Relationships between pairs of variables



### correlation, p. 103

The second step in our analysis is to examine the relationships between all pairs of variables. Scatterplots and correlations are our tools for studying two-variable relationships. The correlations appear in Figure 11.4. The output includes the  $P$ -value for the test of the null hypothesis that the population correlation is 0 versus the two-sided alternative for each pair. Thus, we see that the correlation between GPA and HSM is 0.42, with a  $P$ -value of 0.000 (that is,  $P < 0.0005$ ), whereas the correlation between GPA and SATW is 0.22, with a  $P$ -value of 0.006. Because of the large sample size, even somewhat weak associations are found to be statistically significant.



**FIGURE 11.4**

Correlations among the case study variables.

As we might expect, math and science grades have the highest correlation with GPA ( $r = 0.42$  and  $r = 0.44$ ), followed by English grades (0.36) and then SAT Mathematics (0.33). SAT Critical Reading (SATCR) and SAT Writing (SATW) have comparable, somewhat weak, correlations with GPA. On the other hand, SATCR and SATW have a high correlation with each other (0.73). The high school grades also correlate well with each other (0.49 to 0.70). SATM correlates well with the other SAT scores (0.58 and 0.55), somewhat with HSM (0.32), less with HSS (0.22), and poorly with HSE (0.13). SATCR and SATW do not correlate well with any of the high school grades (0.07 to 0.26).



It is important to keep in mind that by examining pairs of variables we are

seeking a better understanding of the data. *The fact that the correlation of a particular explanatory variable with the response variable does not achieve statistical significance does not necessarily imply that it will not be a useful (and statistically significant) predictor in a multiple regression.*

Numerical summaries such as correlations are useful, but plots are generally more informative when seeking to understand data. Plots tell us whether the numerical summary gives a fair representation of the data.



For a multiple regression, each pair of variables should be plotted. For the seven variables in our case study, this means that we should examine 21 plots. In general, there are  $p + 1$  variables in a multiple regression analysis with  $p$  explanatory variables, so that  $p(p + 1)/2$  plots are required. *Multiple regression is a complicated procedure. If we do not do the necessary preliminary work, we are in serious danger of producing useless or misleading results.* We leave the task of making these plots as an exercise.

## USE YOUR KNOWLEDGE

### 11.5 Pairwise relationships among variables in the GPA data set.



Using a statistical package, generate the pairwise correlations and scatterplots discussed previously. Comment on any unusual patterns or observations.

## Regression on high school grades

To explore the relationship between the explanatory variables and our response variable GPA, we run several multiple regressions. The explanatory variables fall into two classes. High school grades are represented by HSM, HSS, and HSE, and standardized tests are represented by the three SAT scores. We begin our analysis by using the high school grades to predict GPA. Figure 11.5 gives the multiple

regression output.

The output contains an ANOVA table, some additional descriptive statistics, and information about the parameter estimates. When examining any ANOVA table, it is a good idea to first verify the degrees of freedom. This ensures that we have not made some serious error in specifying the model for the software or in entering the data. Because there are  $n = 150$  cases, we have  $DFT = n - 1 = 149$ . The three explanatory variables give  $DFM = p = 3$  and  $DFE = n - p - 1 = 150 - 3 - 1 = 146$ .

The ANOVA  $F$  statistic is 14.35, with a  $P$ -value of  $< 0.0001$ . Under the null hypothesis

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

the  $F$  statistic has an  $F(3, 146)$  distribution. According to this distribution, the chance of obtaining an  $F$  statistic of 14.35 or larger is less than 0.0001. We therefore conclude that at least one of the three regression coefficients for the high school grades is different from 0 in the population regression equation.

In the descriptive statistics that follow the ANOVA table we find that Root MSE is 0.726. This value is the square root of the MSE given in the ANOVA table and is  $s$ , the estimate of the parameter  $\sigma$  of our model. The value of  $R^2$  is 0.23. That is, 23% of the observed variation in the GPA scores is explained by linear regression on high school grades.

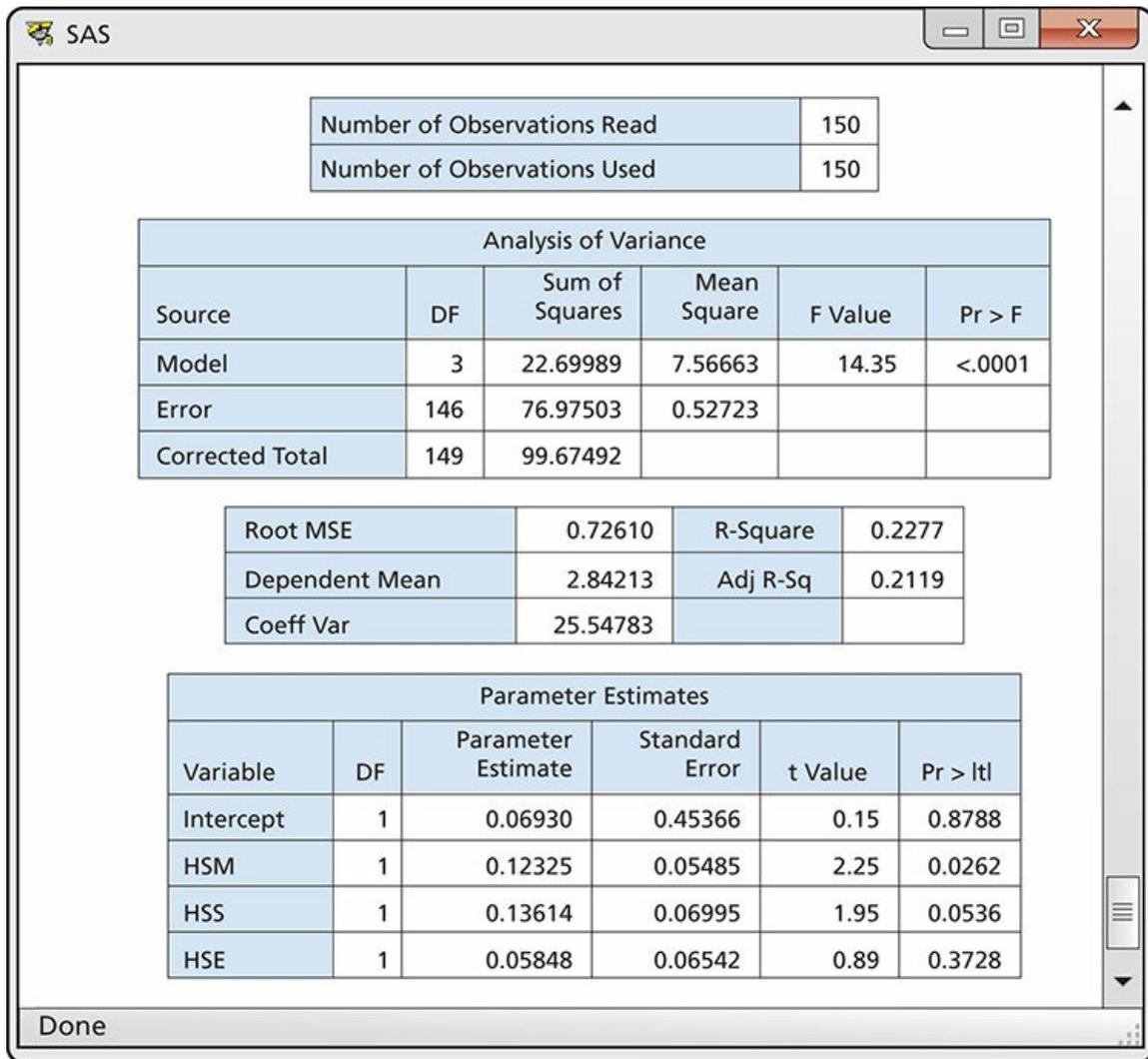
Although the  $P$ -value of the  $F$  test is very small, the model does not explain very much of the variation in GPA. Remember, a small  $P$ -value does not necessarily tell us that we have a strong predictive relationship, particularly when the sample size is large.

From the Parameter Estimates section of the computer output we obtain the fitted regression equation

$$\widehat{\text{GPA}} = 0.069 + 0.123\text{HSM} + 0.136\text{HSS} + 0.058\text{HSE}$$

Let's find the predicted GPA for a student with an A- average in HSM, B+ in HSS, and B in HSE. The explanatory variables are  $\text{HSM} = 9$ ,  $\text{HSS} = 8$ , and  $\text{HSE} = 7$ . The predicted GPA is

$$\begin{aligned}\widehat{\text{GPA}} &= 0.069 + 0.123(9) + 0.136(8) + 0.058(7) \\ &= 2.67\end{aligned}$$



**FIGURE 11.5**

Multiple regression output for regression using high school grades to predict GPA.

Recall that the  $t$  statistics for testing the regression coefficients are obtained by dividing the estimates by their standard errors. Thus, for the coefficient of HSM we obtain the  $t$ -value given in the output by calculating

$$t = b / SE_b = 0.12325 / 0.05485 = 2.25$$

The  $P$ -values appear in the last column. Note that these  $P$ -values are for the two-sided alternatives. HSM has a  $P$ -value of 0.0262, and we conclude that the regression coefficient for this explanatory variable is significantly different from 0. The  $P$ -values for the other explanatory variables (0.0536 for HSS and 0.3728 for HSE) do not achieve statistical significance.

### Interpretation of results

The significance tests for the individual regression coefficients seem to contradict the impression obtained by examining the correlations in Figure 11.4. In that

display we see that the correlation between GPA and HSS is 0.44 and the correlation between GPA and HSE is 0.36. The  $P$ -values for both of these correlations are  $< 0.0005$ . In other words, if we used HSS alone in a regression to predict GPA, or if we used HSE alone, we would obtain statistically significant regression coefficients.

This phenomenon is not unusual in multiple regression analysis. Part of the explanation lies in the correlations between HSM and the other two explanatory variables. These are rather high (at least compared with most other correlations in Figure 11.4). The correlation between HSM and HSS is 0.67, and that between HSM and HSE is 0.49. Thus, when we have a regression model that contains all three high school grades as explanatory variables, there is considerable overlap of the predictive information contained in these variables.



*The significance tests for individual regression coefficients assess the significance of each predictor variable assuming that all other predictors are included in the regression equation.* Given that we use a model with HSM and HSS as predictors, the coefficient of HSE is not statistically significant. Similarly, given that we have HSM and HSE in the model, HSS does not have a significant regression coefficient. HSM, however, adds significantly to our ability to predict GPA even after HSS and HSE are already in the model.

Unfortunately, we cannot conclude from this analysis that the *pair* of explanatory variables HSS and HSE contribute nothing significant to our model for predicting GPA once HSM is in the model. Questions like these require fitting additional models.

*The impact of relations among the several explanatory variables on fitting models for the response is the most important new phenomenon encountered in moving from simple linear regression to multiple regression.* In this chapter, we can only illustrate some of the many complicated problems that can arise.

## Residuals

As in simple linear regression, we should always examine the residuals as an aid to determining whether the multiple regression model is appropriate for the data. Because there are several explanatory variables, we must examine several residual plots. It is usual to plot the residuals versus the predicted values  $\hat{y}$  and also versus each of the explanatory variables. Look for outliers, influential observations, evidence of a curved (rather than linear) relation, and anything else unusual. Again, we leave the task of making these plots as an exercise. The plots all appear to show more or less random noise above and below the center value of 0.

If the deviations  $\varepsilon$  in the model are Normally distributed, the residuals should be

Normally distributed. Figure 11.6 presents a Normal quantile plot and histogram of the residuals. Both suggest some skewness (shorter right tail) in the distribution. However, given our large sample size, we do not think this skewness is strong enough to invalidate this analysis.

## USE YOUR KNOWLEDGE

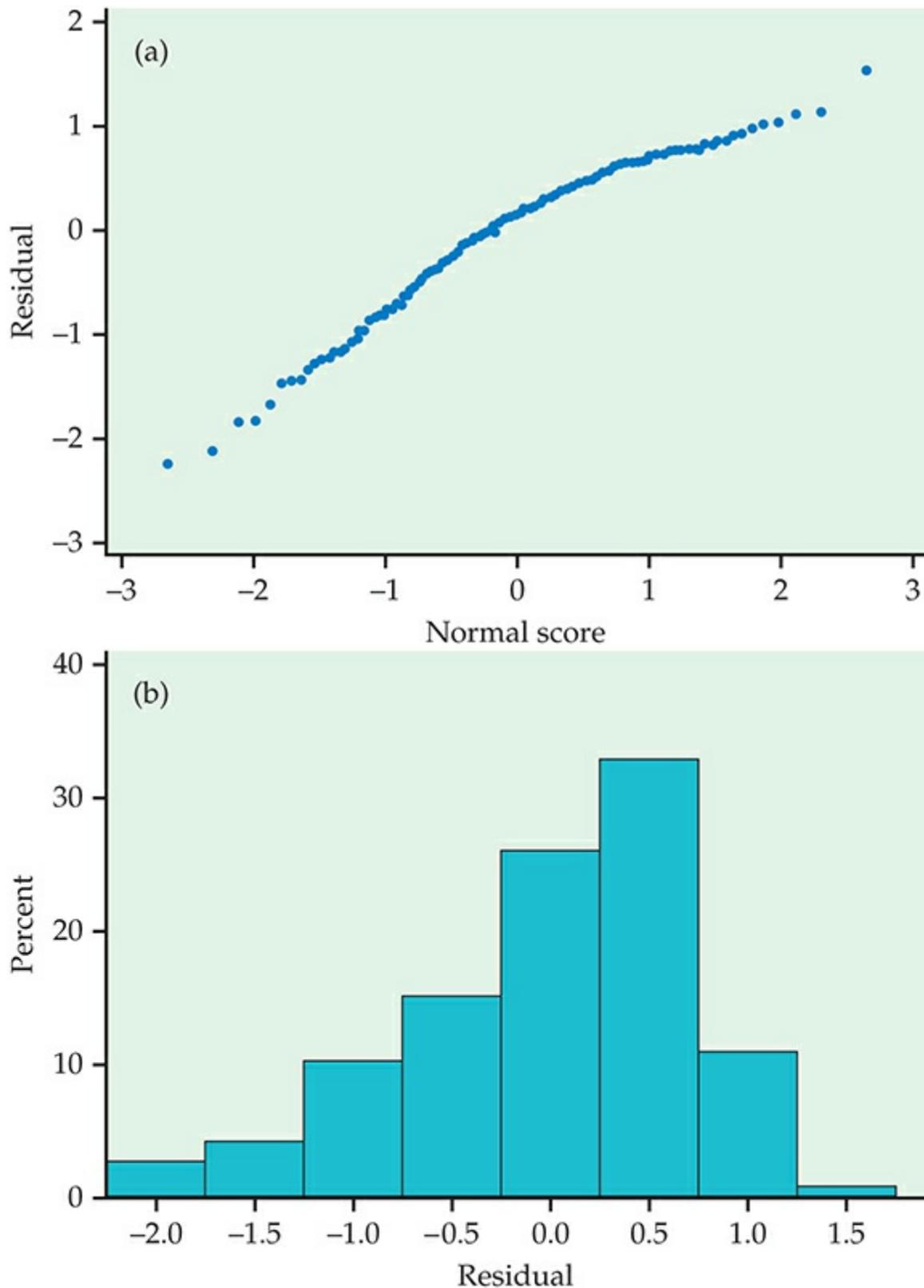
### 11.6 Residual plots for the GPA analysis.



Using a statistical package, fit the linear model with HSM and HSE as predictors and obtain the residuals and predicted values. Plot the residuals versus the predicted values, HSM, and HSE. Are the residuals more or less randomly dispersed around zero? Comment on any unusual patterns.

## Refining the model

Because the variable HSE has the largest  $P$ -value of the three explanatory variables (see Figure 11.5) and therefore appears to contribute the least to our explanation of GPA, we rerun the regression using only HSM and HSS as explanatory variables. The SAS output appears in Figure 11.7. The  $F$  statistic indicates that we can reject the null hypothesis that the regression coefficients for the two explanatory variables are both 0. The  $P$ -value is still  $<0.0001$ . The value of  $R^2$  has dropped very slightly compared with our previous run, from 0.2277 to 0.2235. Thus, dropping HSE from the model resulted in the loss of very little explanatory power.



**FIGURE 11.6**

(a) Normal quantile plot and (b) histogram of the residuals from the high school grades model. There are no important deviations from Normality.

The measure  $s$  of variation about the fitted equation (Root MSE in the printout) is nearly identical for the two regressions, another indication that we lose very little

when we drop HSE. The  $t$  statistics for the individual regression coefficients indicate that HSM is still significant ( $P = 0.0240$ ), while the statistic for HSS is larger than before (2.99 versus 1.95) and is now statistically significant ( $P = 0.0032$ ).

Comparison of the fitted equations for the two multiple regression analyses tells us something more about the intricacies of this procedure. For the first run we have

$$\text{GPA}^{\hat{}} = 0.069 + 0.123\text{HSM} + 0.136\text{HSS} + 0.058\text{HSE}$$

whereas the second gives us

$$\text{GPA}^{\hat{}} = 0.257 + 0.125\text{HSM} + 0.172\text{HSS}$$

Eliminating HSE from the model changes the regression coefficients for all the remaining variables and the intercept. This phenomenon occurs quite generally in multiple regression. *Individual regression coefficients, their standard errors, and significance tests are meaningful only when interpreted in the context of the other explanatory variables in the model.*



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	22.27859	11.13930	21.16	<.0001
Error	147	77.39633	0.52651		
Corrected Total	149	99.67492			

Root MSE	0.72561	R-Square	0.2235
Dependent Mean	2.84213	Adj R-Sq	0.2129
Coeff Var	25.53037		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	0.25696	0.40189	0.64	0.5236
HSM	HSM	1	0.12498	0.05478	2.28	0.0240
HSS	HSS	1	0.17182	0.05740	2.99	0.0032

Done

## FIGURE 11.7

Multiple regression output for regression using HSM and HSS to predict GPA.

### Regression on SAT scores

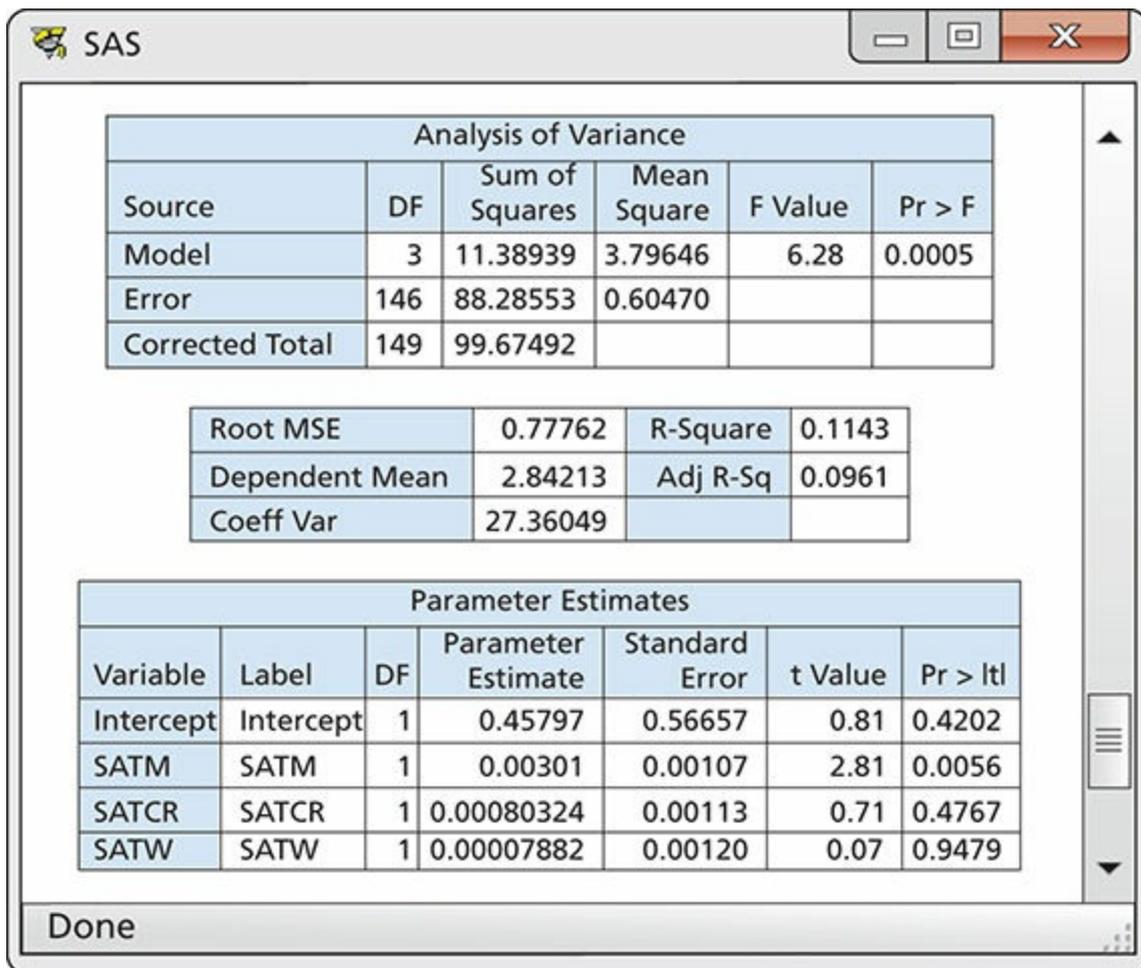
We now turn to the problem of predicting GPA using the three SAT scores. Figure 11.8 gives the output. The fitted model is

$$\widehat{\text{GPA}} = 0.45797 + 0.00301\text{SATM} + 0.00080\text{SATCR} + 0.00008\text{SATW}$$

The degrees of freedom are as expected: 3, 146, and 149. The  $F$  statistic is 6.28, with a  $P$ -value of 0.0005. We conclude that the regression coefficients for SATM, SATCR, and SATW are not all 0. Recall that we obtained the  $P$ -value  $< 0.0001$  when we used high school grades to predict GPA. Both multiple regression equations are highly significant, but this obscures the fact that the two models have quite different explanatory power. For the SAT regression,  $R^2 = 0.1143$ , whereas for the high school grades model even with only HSM and HSS (Figure 11.7), we have  $R^2 = 0.2235$ , a value almost twice as large. *Stating that we have a statistically significant result is quite different from saying that an effect is large or important.*



Further examination of the output in Figure 11.8 reveals that the coefficient of SATM is significant ( $t = 2.81$ ,  $P = 0.0056$ ), and that SATCR ( $t = 0.71$ ,  $P = 0.4767$ ) and SATW ( $t = 0.07$ ,  $P = 0.9479$ ) are not. For a complete analysis we should carefully examine the residuals. Also, we might want to run the analysis without SATW and the analysis with SATM as the only explanatory variable.



**FIGURE 11.8**

Multiple regression output for regression using SAT scores to predict GPA.

## Regression using all variables

We have seen that fitting a model using either the high school grades or the SAT scores results in a highly significant regression equation. The mathematics component of each of these groups of explanatory variables appears to be a key predictor. Comparing the values of  $R^2$  for the two models indicates that high school grades are better predictors than SAT scores. Can we get a better prediction equation using all the explanatory variables together in one multiple regression?

To address this question we run the regression with all six explanatory variables. The output from SAS, Minitab, and Excel appears in Figure 11.9. Although the format and organization of outputs differ among software packages, the basic results that we need are easy to find.

The degrees of freedom are as expected: 6, 143, and 149. The  $F$  statistic is 8.95, with a  $P$ -value  $< 0.0001$ , so at least one of our explanatory variables has a nonzero regression coefficient. This result is not surprising, given that we have already seen that HSM and SATM are strong predictors of GPA. The value of  $R^2$  is 0.2730, which is about 0.05 higher than the value of 0.2235 that we found for the high

school grades regression.

Examination of the  $t$  statistics and the associated  $P$ -values for the individual regression coefficients reveals a surprising result. None of the variables are significant! At first, this result may appear to contradict the ANOVA results. How can the model explain over 27% of the variation and have  $t$  tests that suggest none of the variables make a significant contribution?

Once again it is important to understand that these  $t$  tests assess the contribution of each variable when it is added to a model that already has the other five explanatory variables. This result does not necessarily mean that the regression coefficients for the six explanatory variables are *all* 0. It simply means that the contribution of each variable overlaps considerably with the contribution of the other five variables already in the model.

SAS

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	27.21030	4.53505	8.95	<.0001
Error	143	72.46462	0.50675		
Corrected Total	149	99.67492			

Root MSE	0.71186	R-Square	0.2730
Dependent Mean	2.84213	Adj R-Sq	0.2425
Coeff Var	25.04670		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	-1.18678	0.61641	-1.93	0.0562
SATM	SATM	1	0.00199	0.00106	1.88	0.0619
SATCR	SATCR	1	0.00015701	0.00105	0.15	0.8813
SATW	SATW	1	0.00047398	0.00112	0.42	0.6719
HSM	HSM	1	0.09148	0.05718	1.60	0.1119
HSS	HSS	1	0.13010	0.06877	1.89	0.0605
HSE	HSE	1	0.05679	0.06568	0.86	0.3887

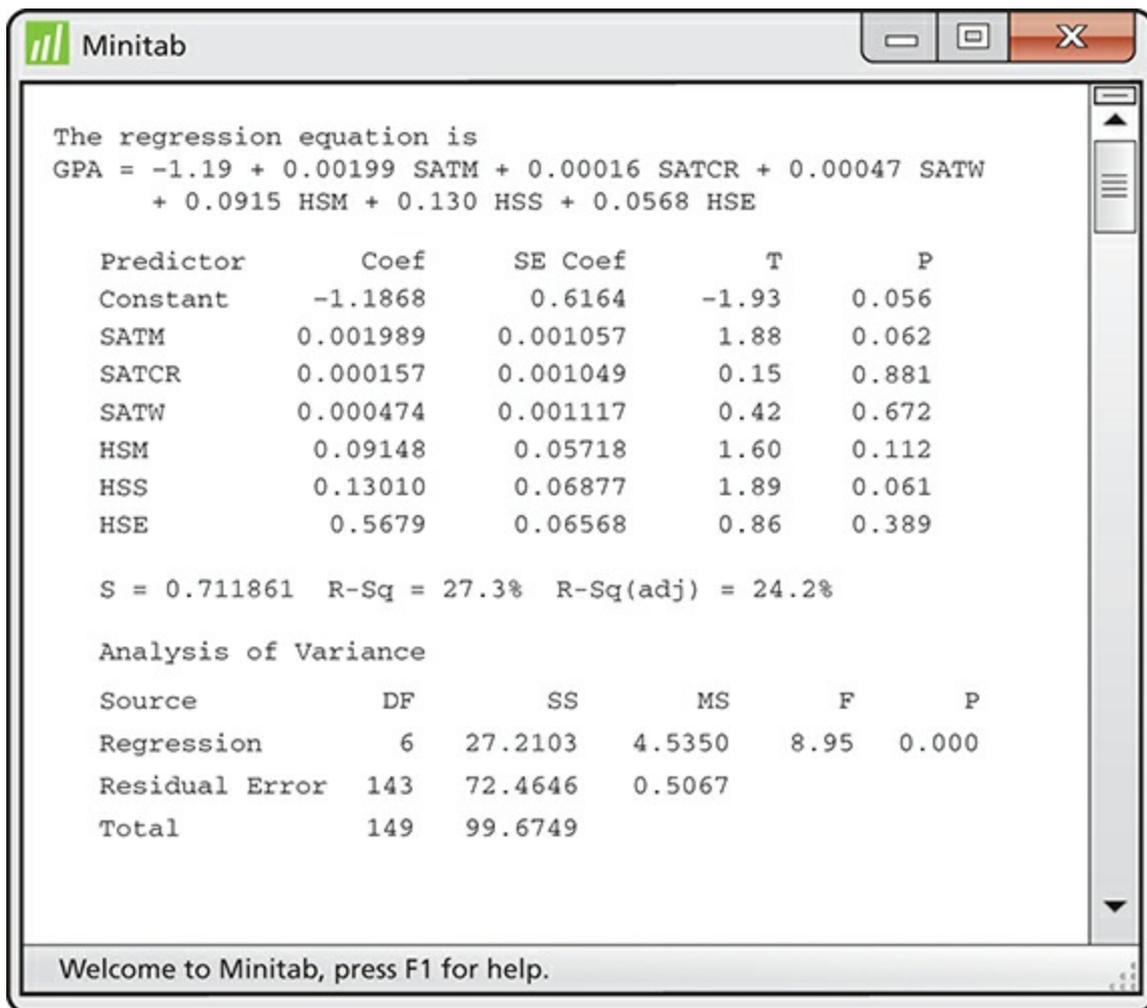
  

Test SAT Results for Dependent Variable GPA				
Source	DF	Mean Square	F Value	Pr > F
Numerator	3	1.50347	2.97	0.0341
Denominator	143	0.50675		

Test HS Results for Dependent Variable GPA				
Source	DF	Mean Square	F Value	Pr > F
Numerator	3	5.27364	10.41	<.0001
Denominator	143	0.50675		

Done



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	<i>Regression Statistics</i>						
4	Multiple R	0.522484872					
5	R Square	0.272990441					
6	Adjusted R Square	0.242486543					
7	Standard Error	0.711860645					
8	Observations	150					
9							
10	<i>ANOVA</i>						
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	6	27.21029964	4.53505	8.949363	2.69075E-08	
13	Residual	143	72.46461769	0.506746			
14	Total	149	99.67491733				
15							
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-Value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	-1.18678259	0.616408136	-1.92532	0.056174	-2.40523174	0.031666565
18	SATM	0.001988735	0.001056719	1.881996	0.061869	-0.00010007	0.004077537
19	SATCR	0.000157014	0.001049462	0.149614	0.88128	-0.00191745	0.002231478
20	SATW	0.000473983	0.001116795	0.424414	0.671902	-0.00173358	0.002681542
21	HSM	0.091476942	0.057181116	1.599775	0.111855	-0.02155252	0.204506408
22	HSS	0.130096576	0.068767787	1.891824	0.060536	-0.00583617	0.266029324
23	HSE	0.056790708	0.065681417	0.864639	0.388685	-0.07304124	0.186622652

**FIGURE 11.9**

Multiple regression output for regression using all variables to predict GPA.

When a model has a large number of insignificant variables, it is common to refine the model. We prefer smaller models to larger models because they are easier to work with and understand. However, given the many complications that can arise in multiple regression, there is no universal “best” approach to refine a model. There is also no guarantee that there is just one acceptable refined model.

Many statistical software packages now provide the capability of summarizing all possible models from a set of  $P$  variables. We suggest using this capability to reduce the number of candidate models (for example, there are a total of 63 models when  $p = 6$ ) and then carefully studying the remaining models before making a decision as to a best model or set of best models. If in doubt, consult an expert.

## Test for a collection of regression coefficients

Many statistical software packages also provide the capability for testing whether a collection of regression coefficients in a multiple regression model are *all* 0. We use this approach to address two interesting questions about our data set. We did not discuss such tests in the outline that opened this section, but the basic idea is quite simple and discussed in Exercise 11.26 (page 637).

In the context of the multiple regression model with all six predictors, we ask first whether or not the coefficients for the three SAT scores are all 0. In other words, do the SAT scores add any significant predictive information to that already contained in the high school grades? To be fair, we also ask the complementary question: Do the high school grades add any significant predictive information to that already contained in the SAT scores?

The answers are given in the last two parts of the SAS output in Figure 11.9. For the first test we see that  $F = 2.97$ . Under the null hypothesis that the three SAT coefficients are 0, this statistic has an  $F(3, 143)$  distribution and the  $P$ -value is 0.0341. We conclude that the SAT scores (as a group) are significant predictors of GPA in a regression that already contains the high school scores as predictor variables. This means that we cannot just focus on refined models that involve the high school grades. Both high school grades and SAT scores appear to contribute to our explanation of GPA.

The test statistic for the three high school grade variables is  $F = 10.41$ . Under the null hypothesis that these three regression coefficients are 0, the statistic has an  $F(3, 143)$  distribution and the  $P$ -value is  $< 0.0001$ . Again this means that high school grades contain useful information for predicting GPA that is not contained in the SAT scores.

---

## BEYOND THE BASICS

### Multiple logistic regression

Many studies have yes/no or success/failure response variables. A surgery patient lives or dies; a consumer does or does not purchase a product after viewing an advertisement. Because the response variable in a multiple regression is assumed to have a Normal distribution, this methodology is not suitable for predicting such responses. However, there are models that apply the ideas of regression to response variables with only two possible outcomes.

One type of model that can be used is called **logistic regression**. We think in terms of a binomial model for the two possible values of the response variable and use one or more explanatory variables to explain the probability of

success. Details are more complicated than those for multiple regression and are given in Chapter 14. However, the fundamental ideas are very much the same. Here is an example.

## logistic regression

### EXAMPLE

#### 11.2 Tipping behavior in Canada.



The Consumer Report on Eating Share Trends (CREST) contains data spanning all provinces of Canada and details away-from-home food purchases by roughly 4000 households per quarter. Some researchers accessed these data but restricted their attention to restaurants at which tips would normally be given.<sup>4</sup> From a total of 73,822 observations, “high” and “low” tipping variables were created based on whether the observed tip rate was above 20% or below 10%, respectively. They then used logistic regression to identify explanatory variables associated with either “high” or “low” tips.

The model consisted of over 25 explanatory variables, grouped as “control” variables and “stereotype-related” variables. The stereotype-related explanatory variables were  $x_1$ , a variable having the value 1 if the age of the diner was greater than 65 years, and 0 otherwise;  $x_2$ , coded as 1 if the meal was on Sunday, and 0 otherwise;  $x_3$ , coded as 1 to indicate English was a second language;  $x_4$ , a variable coded 1 if the diner was a French-speaking Canadian;  $x_5$ , a variable coded 1 if alcoholic drinks were served with the meal;

and  $x_6$ , a variable coded 1 if the meal involved a lone male.



### chi-square distribution, p. 538

Similar to the  $F$  test in multiple regression, there is a chi-square test for multiple logistic regression that tests the null hypothesis that *all* coefficients of the explanatory variables are zero. These results were not presented in the article because the focus was more on comparing the high- and low-tip models. In place of the  $t$  tests for individual coefficients in multiple regression, chi-square tests, each with 1 degree of freedom, are used to test whether individual coefficients are zero. The article does report these tests. A majority of the variables considered in the models have  $P$ -values less than 0.01.

Interpretation of the coefficients is a little more difficult in multiple logistic regression because of the form of the model. For example, the high-tip model (using only the stereotype-related variables) is

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_6 x_6$$

The expression  $p/(1 - p)$  is the **odds** that the tip was above 20%. Logistic regression models the “log odds” as a linear combination of the explanatory variables. Positive coefficients are associated with a higher probability that the tip is high. These coefficients are often transformed back ( $e^{\beta_j}$ ) to the odds scale, giving us an **odds ratio**. An odds ratio greater than 1 is associated with a higher probability that the tip is high. Here is the table of odds ratios reported in the article for the high-tip model:

odds

odds ratio

Explanatory variable	Odds ratio
Senior adult	0.7420*
Sunday	0.9970*
English as second language	0.7360*
French-speaking Canadian	0.7840*
Alcoholic drinks	1.1250*
Lone male	1.0220

The starred values were significant at the 0.01 level. We see that the probability of a high tip is reduced (odds ratio less than 1) when the diner is over 65 years old, speaks English as a second language, and is a French-speaking Canadian. The probability of a high tip is increased (odds ratio greater than 1) if alcohol is served with the meal.

## CHAPTER 11 Summary

**Data for multiple linear regression** consist of the values of a response variable  $y$  and  $P$  explanatory variables  $x_1, x_2, \dots, x_p$ , for  $n$  cases. We write the data and enter them into software in the form

Individual	Variables				
	$x_1$	$x_2$	...	$x_p$	$y$
1	$x_{11}$	$x_{12}$	...	$x_{1p}$	$y_1$
1	$x_{21}$	$x_{22}$	...	$x_{2p}$	$y_2$
⋮					
$n$	$x_{n1}$	$x_{n2}$	...	$x_{np}$	$y_n$

The statistical model for **multiple linear regression** with response variable  $y$  and  $P$  explanatory variables  $x_1, x_2, \dots, x_p$  is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i$$

where  $i = 1, 2, \dots, n$ . The  $\varepsilon_i$  are assumed to be independent and Normally distributed with mean 0 and standard deviation  $\sigma$ . The **parameters** of the model are  $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ , and  $\sigma$ .

The **multiple regression equation** predicts the response variable by a linear relationship with all the explanatory variables:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p$$

The  $\beta$ 's are estimated by  $b_0, b_1, b_2, \dots, b_p$ , which are obtained by the **method of least squares**. The parameter  $\sigma$  is estimated by

$$s = \text{MSE} = \frac{\sum e_i^2}{n-p-1}$$

where the  $e_i$  are the **residuals**,

$$e_i = y_i - \hat{y}_i$$

Always examine the **distribution of the residuals** and plot them against the explanatory variables prior to inference.

A **level  $C$  confidence interval for  $\beta_j$**  is

$$b_j \pm t^* \text{SE}_{b_j}$$

where  $t^*$  is the value for the  $t(n-p-1)$  density curve with area  $C$  between  $-t^*$  and  $t^*$ .

The test of the hypothesis  $H_0: \beta_j = 0$  is based on the  **$t$  statistic**

$$t = \frac{b_j}{\text{SE}_{b_j}}$$

and the  $t(n - p - 1)$  distribution.

The estimate  $b_j$  of  $\beta_j$  and the test and confidence interval for  $\beta_j$  are all based on a specific multiple linear regression model. The results of all these procedures change if other explanatory variables are added to or deleted from the model.

The **ANOVA table** for a multiple linear regression gives the degrees of freedom, sum of squares, and mean squares for the model, error, and total sources of variation. The **ANOVA  $F$  statistic** is the ratio  $MSM/MSE$  and is used to test the null hypothesis

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$$

If  $H_0$  is true, this statistic has an  $F(p, n - p - 1)$  distribution.

The **squared multiple correlation** is given by the expression

$$R^2 = \frac{SSM}{SST}$$

and is interpreted as the proportion of the variability in the response variable  $y$  that is explained by the explanatory variables  $x_1, x_2, \dots, x_p$  in the multiple linear regression.

## CHAPTER 11 Exercises

*For Exercise 11.1, see page 614; for Exercise 11.2, see page 615; for Exercises 11.3 and 11.4, see page 619; for Exercise 11.5, see page 623; and for Exercise 11.6, see page 625.*

### 11.7 95% confidence intervals for regression coefficients.

In each of the following settings, give a 95% confidence interval for the coefficient of  $x_1$ .

- (a)  $n = 26$ ,  $\hat{y} = 1.6 + 6.4x_1 + 5.7x_2$ ,  $SE_{b_1} = 3.1$
- (b)  $n = 53$ ,  $\hat{y} = 1.6 + 6.4x_1 + 5.7x_2$ ,  $SE_{b_1} = 2.9$
- (c)  $n = 26$ ,  $\hat{y} = 1.6 + 4.8x_1 + 3.2x_2 + 5.2x_3$ ,  $SE_{b_1} = 2.2$
- (d)  $n = 124$ ,  $\hat{y} = 1.6 + 4.8x_1 + 3.2x_2 + 5.2x_3$ ,  $SE_{b_1} = 2.1$

### 11.8 Significance tests for regression coefficients.

For each of the settings in the previous exercise, test the null hypothesis that the coefficient of  $x_1$  is zero versus the two-sided alternative.

### 11.9 What's wrong?

In each of the following situations, explain what is wrong and why.

- (a) In a multiple regression with a sample size of 39 and 3 explanatory variables, the test statistic for the null hypothesis  $H_0: b_2 = 0$  is a  $t$  statistic that follows the  $t(35)$  distribution when the null hypothesis is true.
- (b) The multiple correlation coefficient gives the proportion of the variation in the response variable that is explained by the explanatory variables.
- (c) A small  $P$ -value for the ANOVA  $F$  test implies that all explanatory variables are significantly different from zero.

### 11.10 What's wrong?

In each of the following situations, explain what is wrong and why.

- (a) One of the assumptions for multiple regression is that the distribution of each explanatory variable is Normal.
- (b) The smaller the  $P$ -value for the ANOVA  $F$  test, the greater the explanatory power of the model.
- (c) All explanatory variables that are significantly correlated with the response variable will have a

statistically significant regression coefficient in the multiple regression model.

(d) The multiple correlation coefficient gives the average correlation between the response variable and each explanatory variable in the model.

### 11.11 Constructing the ANOVA table.

Seven explanatory variables are used to predict a response variable using a multiple regression. There are 142 observations.

- (a) Write the statistical model that is the foundation for this analysis. Also include a description of all assumptions.
- (b) Outline the analysis of variance table giving the sources of variation and numerical values for the degrees of freedom.

### 11.12 More on constructing the ANOVA table.

A multiple regression analysis of 78 cases was performed with 5 explanatory variables. Suppose that  $SSM = 16.5$  and  $SSE = 100.8$ .

- (a) Find the value of the  $F$  statistic for testing the null hypothesis that the coefficients of all the explanatory variables are zero.
- (b) What are the degrees of freedom for this statistic?
- (c) Find bounds on the  $P$ -value using Table E. Show your work.
- (d) What proportion of the variation in the response variable is explained by the explanatory variables?

### 11.13 Refining the GPA model using all variables.

Figure 11.9 (page 629) summarizes the regression model using all variables. Let's now compare several reduced models. For each of the following models, report the fitted model, MSE, percent explained variation, and the  $P$ -values for each of the individual coefficients. Based on these results, which model do you think is "best"? Explain your answer.  GPA

- (a) SATM and HSS
- (b) SATM, HSM, and HSS
- (c) SATM, HSM, HSS, and HSE
- (d) HSM and HSS

### 11.14 Predicting college debt: combining measures.

Refer to Exercises 10.10 (page 601) and 10.14 (page 602) for a description of the problem. Let's now consider fitting a model using all the explanatory variables.  BESTVALUE

- (a) Write out the statistical model for this analysis, making sure to specify all assumptions.
- (b) Run the multiple regression model and specify the fitted regression equation.

(c) Obtain the residuals from part (b) and check assumptions. Comment on any unusual residuals or patterns in the residuals.

(d) What percent of the variability in average debt is explained by this model?

### 11.15 Predicting college debt: a simpler model.

Refer to the previous exercise. In the multiple regression analysis using all seven variables, only one variable, StudPerFac, is significant at the 0.05 level. Remove the variable with the highest  $P$ -value one at a time until you end up with a multiple regression model that has only significant predictors.

Summarize your final model in a short paragraph.  **BESTVALUE**

### 11.16 Comparison of prediction intervals.

Refer to the previous two exercises. The Ohio State University has Admit = 68, Yr4Grad = 49, StudPerFac = 19, InAfterAid = 12,680, OutAfterAid = 27,575, AvgAid = 7789, and PercBorrow = 52. Use your software to construct

(a) a 95% prediction interval based on the model with all the predictors.

(b) a 95% prediction interval based on the model using your simpler model.

(c) Compare the two intervals. Do the models give similar predictions and intervals?

### 11.17 Predicting energy-drink consumption.

Energy-drink advertising consistently emphasizes a physically active lifestyle and often features extreme sports and risk taking. Are these typical characteristics of an energy-drink consumer? A researcher decided to examine the links between energy-drink consumption, sport-related (jock) identity, and risk taking.<sup>5</sup> She invited over 1500 undergraduate students enrolled in large introductory-level courses at a public university to participate. Each participant had to complete a 45-minute anonymous questionnaire. From this questionnaire jock identity and risk-taking scores were obtained, where the higher the score, the stronger the trait. She ended up with 795 respondents. The following table summarizes the results of a multiple regression analysis using the frequency of energy-drink consumption in the past 30 days as the response variable:

Explanatory variable	$b$
Age	-0.02
Sex (1 = female, 0 = male)	-0.11 <sup>**</sup>
Race (1 = nonwhite, 0 = white)	-0.02
Ethnicity (1 = Hispanic, 0 = non-Hispanic)	0.10 <sup>**</sup>
Parental education	0.02
College GPA	-0.01
Jock identity	0.05
Risk taking	0.19 <sup>***</sup>

A superscript of <sup>\*\*</sup> means that the individual coefficient  $t$  test had a  $P$ -value less than 0.01, and a superscript of <sup>\*\*\*</sup> means that the test had a  $P$ -value less than 0.001. All other  $P$ -values were greater than 0.05.

- (a) The overall  $F$  statistic is reported to be 8.11. What are the degrees of freedom associated with this statistic?
- (b)  $R$  is reported to be 0.28. What percent of the variation in energy-drink consumption is explained by the model? Is this a highly predictive model? Explain.
- (c) Interpret each of the regression coefficients that are significant.
- (d) The researcher states, “Controlling for gender, age, race, ethnicity, parental educational achievement, and college GPA, each of the predictors (risk taking and jock identity) was positively associated with energy-drink consumption frequency.” Explain what is meant by “controlling for” these variables and how this helps strengthen her assertion that jock identity and risk taking are positively associated with energy-drink consumption.

### 11.18 Consider the gender of the students.

Refer to Exercise 11.13. The seventh predictor variable provided in the GPA data set is a gender indicator variable. This variable (SEX) takes the value 1 for males and 2 for females (see Figure 11.1). If we include it in our model, it allows the intercept to differ for the two genders. If we plug in the indicator values, we see that the estimated male intercept is  $b_0 + b_7(1)$  and the estimated female intercept is  $b_0 + b_7(2)$ . The estimate  $b_0 + b_7$  represents the fitted coefficient for the SEX indicator variable. Also notice that if we take the difference between these two estimates,  $b_0 + b_7(2) - (b_0 + b_7(1)) = b_7$ , this coefficient is also an estimate of the difference in intercepts.  GPA

- (a) Add the variable SEX to each of the models in Exercise 11.13 and repeat the exercise.
- (b) Does this indicator variable appear to contribute to our explanation of GPA? If it does, do males or females have higher GPA scores? Explain your answer.
- (c) Given your results in Exercise 11.13 and the ones here, which model do you think is the best? Explain your answer.

### 11.19 A mechanistic explanation of popularity.

In Exercise 10.59 (page 609) correlations between an adolescent’s “popularity,” expression of a serotonin receptor gene, and rule-breaking behaviors were assessed. An additional portion of the analysis looked at the relationship between the gene expression level and popularity, after adjusting for rule-breaking (RB) behaviors. This adjustment was necessary because RB is positively associated both with this gene expression and with popularity in adolescents. The following summarizes these regression analyses using the composite (questionnaire and video) RB score. A total of 202 individuals were included in this analysis.

	Estimate	Standard error
Model 1		
Gene expression	0.204	0.066
Model 2		
Gene expression	0.161	0.066
RB.composite	0.100	0.030

For all analysis use the 0.05 significance level.

- (a) What are the error degrees of freedom for Model 1 and Model 2?
- (b) Test the null hypothesis that the serotonin gene receptor coefficient is equal to 0 in Model 1.

State the test statistic and  $P$ -value.

(c) Perform both individual-variable  $t$  tests for Model 2. Again state the test statistics and  $P$ -values.

(d) Is there still a positive relationship between the serotonin gene receptor expression level and popularity after adjusting for RB? If yes, compare the increase in popularity for a unit increase in gene expression (while RB remains unchanged) in the two models.

Results such as these suggest not only that adolescents with high serotonin receptor gene expression are predisposed to increased RB behaviors, but also that such behaviors are socially advantageous.

## 11.20 Is the number of tornadoes increasing?

In Exercise 10.29, data on the number of tornadoes in the United States between 1953 and 2012 were analyzed to see if there was a linear trend over time. Many argue that the probability of sighting a tornado has increased over time because there are more people living in the United States.

Let's investigate this by including the U.S. census count as an additional explanatory variable. 

### TWISTER

(a) Using numerical and graphical summaries, describe the relationship between each pair of variables.

(b) Perform a multiple regression using both year and population count as explanatory variables. Write down the fitted model.

(c) Obtain the residuals from part (b). Plot them versus the two explanatory variables and generate a Normal quantile plot. What do you conclude?

(d) Test the hypothesis that there is a linear increase over time. State the null and alternative hypotheses, test statistic, and  $P$ -value. What is your conclusion?

## 11.21 Checking for a polynomial relationship.

When looking at the residuals from the simple linear model of BMI versus physical activity (PA), Figure 10.5 (page 572) suggested a possible curvilinear relationship. Let's investigate this further.

Multiple regression can be used to fit the polynomial curve of degree  $q$ ,  $y = \beta_0 + \beta_1x + \beta_2x^2 + \dots + \beta_qx^q$ , through the creation of additional explanatory variables  $x^2, x^3$ , etc. Let's investigate a quadratic fit ( $q = 2$ ) for the physical activity problem.  **PABMI**

(a) It is often best to subtract the sample mean  $\bar{x}$  before creating the necessary explanatory variables. In this case, the average number of steps per day is 8.614. Create new explanatory variables  $x_1 = (PA - 8.614)$  and  $x_2 = (PA - 8.614)^2$  and run a multiple regression for BMI using the explanatory variables  $x_1$  and  $x_2$ . Write down the fitted regression line.

(b) The regression model that included only PA had  $R^2 = 14.9$ . What is  $R^2$  with the inclusion of this quadratic term?

(c) Obtain the residuals from part (a) and check the multiple regression assumptions. Are there any remaining patterns in the data? Are the residuals approximately Normal? Explain.

(d) Test the hypothesis that the coefficient of the variable  $(PA - 8.614)^2$  is equal to 0. Report the  $t$  statistic, degrees of freedom, and  $P$ -value. Does the quadratic term contribute significantly to the fit? Explain your answer.

## 11.22 Architectural firm billings.

A summary of firms engaged in commercial architecture in the Indianapolis, Indiana, area provides firm characteristics such as total annual billing in billions of dollars and the number of architects, engineers, and staff employed by the firm.<sup>6</sup> Consider developing a model to predict total billing. 

### BILLING

- Using numerical and graphical summaries, describe the distributions of total billing and the number of architects, engineers, and staff.
- For each of the 6 pairs of variables, use graphical and numerical summaries to describe the relationship.
- Carry out a multiple regression. Report the fitted regression equation and the value of the regression standard error  $s$ .
- Analyze the residuals from the multiple regression. Are there any concerns?
- The firm HCO did not report its total billing but employs 3 architects, 1 engineer, and 17 staff members. What is the predicted total billing for this firm?

*The following six exercises use the MOVIES data file. This data set contains an SRS of 35 movies released in the same year. This sample was collected from the Internet Movie Database (IMDb) to see if information available soon after a movie's theatrical release can successfully predict total revenue.*<sup>7</sup> All dollar amounts are measured in millions of U.S. dollars.  MOVIES

## 11.23 Predicting movie revenue—preliminary analysis.

The response variable is a movie's total U.S. revenue (USRevenue). Let's consider as explanatory variables the movie's budget (Budget); opening-weekend revenue (Opening); the number of theaters (Theaters) the movie was in for the opening weekend; and the movie's IMDb rating (Opinion), which is on a 1 to 10 scale (10 being best). While this rating is updated continuously, we'll assume that the current rating is the rating at the end of the first week.

- Using numerical and graphical summaries, describe the distribution of each explanatory variable. Are there any unusual observations that should be monitored?
- Using numerical and graphical summaries, describe the relationship between each pair of explanatory variables.

## 11.24 Predicting movie revenue—simple linear regressions.

Now let's look at the response variable and its relationship with each explanatory variable.

- Using numerical and graphical summaries, describe the distribution of the response variable, USRevenue.
- This variable is not Normally distributed. Does this violate one of the key model assumptions? Explain.
- Generate scatterplots of each explanatory variable and USRevenue. Do all these relationships look linear? Explain what you see.

## 11.25 Predicting movie revenue—multiple linear regression.

Now consider fitting a model using all the explanatory variables.

- (a) Write out the statistical model for this analysis, making sure to specify all assumptions.
- (b) Run the multiple regression model and specify the fitted regression equation.
- (c) Obtain the residuals from part (b) and check assumptions. Comment on any unusual residuals or patterns in the residuals.
- (d) What percent of the variability in USRevenue is explained by this model?

### 11.26 A simpler model.

In the multiple regression analysis using all four explanatory variables, Theaters and Budget appear to be the least helpful (given that the other two explanatory variables are in the model).

- (a) Perform a new analysis using only the movie's openingweekend revenue and IMDb rating. Give the estimated regression equation for this analysis.
- (b) What percent of the variability in USRevenue is explained by this model?
- (c) In this chapter we discussed the  $F$  test for a collection of regression coefficients. In most cases, this capability is provided by the software. When it is not, the test can be performed using the  $R^2$ -values from the full and reduced models. The test statistic is

$$F = \frac{(n-p-1)(R^2_{full} - R^2_{reduced})}{R^2_{reduced}}$$

with  $q$  and  $n - p - 1$  degrees of freedom.  $R^2_{full}$  is the value for the full model and  $R^2_{reduced}$  is the value for the reduced model. Here  $n = 35$  movies,  $p = 4$  variables in the full model, and  $q = 2$  variables were removed to form the reduced model. Plug in the values of  $R^2$  from part (b) of this exercise and part (d) of the previous exercise, and compute the test statistic and  $P$ -value. Do Theaters and Budget combined add any significant predictive information beyond what is already contained in Opening and Opinion?

### 11.27 Predicting U.S. movie revenue.

Refer to the previous two exercises. *Get Smart* was released in the same year, had a budget of \$80.0 million dollars, was shown in 3911 theaters during the first weekend, grossing \$38.7 million dollars, and had an IMDb rating of 6.8. Use your software to construct

- (a) a 95% prediction interval based on the model with all four predictors.
- (b) a 95% prediction interval based on the model using only opening-weekend revenue and IMDb rating.
- (c) Compare the two intervals. Do the models give similar predictions?

### 11.28 Effect of potential outliers.

Consider the simpler model of Exercise 11.26 for this analysis.

- (a) Two movies have much larger U.S. revenues than predicted. Which are they and how much more revenue did they earn than predicted?
- (b) Remove these two movies and redo the multiple regression. Make a table giving the regression

coefficients and their standard errors,  $t$  statistics, and  $P$ -values.

(c) Compare these results with those from Exercise 11.26. How does the removal of these outlying movies impact the estimated model?

(d) Obtain the residuals from this reduced data set and graphically examine their distribution. Do the residuals appear approximately Normal? Explain your answer.

*The following three exercises use the RANKINGS data file. Since 2004, The Times Higher Education Supplement has provided an annual ranking of the world universities. A total score for each university is calculated based on the scores for the following explanatory variables: Peer Review (40%); Faculty-to-Student Ratio (20%); Citations-to-Faculty Ratio (20%); Recruiter Review (10%); Percent International Faculty (5%); and Percent International Students (5%). The percents represent the contributions of each score to the total. For our purposes, we will assume that these weights are unknown and will focus on the development of a model for the total score based on the first three explanatory variables. The report includes a table for the top 200 universities.<sup>8</sup> The RANKINGS data file contains a random sample of 75 of these universities. This is not a random sample of all universities but for our purposes here we will consider it to be.*  **RANKINGS**

### **11.29 Annual ranking of world universities.**

Let's consider developing a model to predict total score based on the peer review score (PEER), faculty-to-student ratio (FtoS), and citations-to-faculty ratio (CtoF).

(a) Using numerical and graphical summaries, describe the distribution of each explanatory variable.

(b) Using numerical and graphical summaries, describe the relationship between each pair of explanatory variables.

### **11.30 Looking at the simple linear regressions.**

Now let's look at the relationship between each explanatory variable and the total score.

(a) Generate scatterplots for each explanatory variable and the total score. Do these relationships all look linear?

(b) Compute the correlation between each explanatory variable and the total score. Are certain explanatory variables more strongly associated with the total score?

### **11.31 Multiple linear regression model.**

Now consider a regression model using all three explanatory variables.

(a) Write out the statistical model for this analysis, making sure to specify all assumptions.

(b) Run the multiple regression model and specify the fitted regression equation.

(c) Generate a 95% confidence interval for each coefficient. Should any of these intervals contain 0? Explain.

(d) What percent of the variation in total score is explained by this model? What is the estimate for  $\sigma^2$ ?

### **11.32 Predicting GPA of seventh-graders.**

Refer to the educational data for 78 seventh-grade students given in Table 1.3 (page 29). We view GPA as the response variable. IQ, gender, and self-concept are the explanatory variables.  **SEVENGR**

- (a) Find the correlation between GPA and each of the explanatory variables. What percent of the total variation in student GPAs can be explained by the straight-line relationship with each of the explanatory variables?
- (b) The importance of IQ in explaining GPA is not surprising. The purpose of the study is to assess the influence of self-concept on GPA. So we will include IQ in the regression model and ask, “How much does self-concept contribute to explaining GPA after the effect of IQ on GPA is taken into account?” Give a model that can be used to answer this question.
- (c) Run the model and report the fitted regression equation. What percent of the variation in GPA is explained by the explanatory variables in your model?
- (d) Translate the question of interest into appropriate null and alternative hypotheses about the model parameters. Give the value of the test statistic and its  $P$ -value. Write a short summary of your analysis with an emphasis on your conclusion.

*The following three exercises use the HAPPY data file. The World Database of Happiness is an online registry of scientific research on the subjective appreciation of life. It is available at [worlddatabaseofhappiness.eur.nl](http://worlddatabaseofhappiness.eur.nl), and the project is directed by Dr. Ruut Veenhoven, Erasmus University, Rotterdam. One inventory presents the “average happiness” score for various nations. This average is based on individual responses from numerous general population surveys to a general life satisfaction (well-being) question. Scores range from 0 (dissatisfied) to 10 (satisfied). The NationMaster website, [www.nationmaster.com](http://www.nationmaster.com), contains a collection of statistics associated with various nations. For our analysis, we will consider the GINI index, which measures the degree of inequality in the distribution of income (higher score = greater inequality); the degree of corruption in government (higher score = less corruption); average life expectancy; and the degree of democracy (higher score = more civil and political liberties).*  **HAPPY**

### 11.33 Predicting a nation’s “average happiness” score.

Consider the five statistics for each nation: LSI, the average life-satisfaction score; GINI, the GINI index; CORRUPT, the degree of government corruption; LIFE, the average life expectancy; and DEMOCRACY, a measure of civil and political liberties.

- (a) Using numerical and graphical summaries, describe the distribution of each variable.
- (b) Using numerical and graphical summaries, describe the relationship between each pair of variables.

### 11.34 Building a multiple linear regression model.

Let’s now build a model to predict the life-satisfaction score, LSI.

- (a) Consider a simple linear regression using GINI as the explanatory variable. Run the regression and summarize the results. Be sure to check assumptions.
- (b) Now consider a model using GINI and LIFE. Run the multiple regression and summarize the results. Again be sure to check assumptions.
- (c) Now consider a model using GINI, LIFE, and DEMOCRACY. Run the multiple regression and summarize the results. Again be sure to check assumptions.

(d) Now consider a model using all four explanatory variables. Again summarize the results and check assumptions.

### 11.35 Selecting from among several models.

Refer to the results from the previous exercise.

- Make a table giving the estimated regression coefficients, standard errors,  $t$  statistics, and  $P$ -values.
- Describe how the coefficients and  $P$ -values change for the four models.
- Based on the table of coefficients, suggest another model. Run that model, summarize the results, and compare it with the other ones. Which model would you choose to explain LSI? Explain.

*The following six exercises use the BIOMARK data file. Healthy bones are continually being renewed by two processes. Through bone formation, new bone is built; through bone resorption, old bone is removed. If one or both of these processes are disturbed, by disease, aging, or space travel, for example, bone loss can be the result. The variables  $VO+$  and  $VO-$  measure bone formation and bone resorption, respectively. Osteocalcin (OC) is a biochemical marker for bone formation: higher levels of bone formation are associated with higher levels of OC. A blood sample is used to measure OC, and it is much less expensive to obtain than direct measures of bone formation. The units are milligrams of OC per milliliter of blood (mg/ml). Similarly, tartrate-resistant acid phosphatase (TRAP) is a biochemical marker for bone resorption that is also measured in blood. It is measured in units per liter (U/l). These variables were measured in a study of 31 healthy women aged 11 to 32 years.<sup>9</sup> Variables with the first letter "L" are the logarithms of the measured variables.*

**BIOMARK**

### 11.36 Bone formation and resorption.

Consider the following four variables:  $VO+$ , a measure of bone formation;  $VO-$ , a measure of bone resorption; OC, a biomarker of bone formation; and TRAP, a biomarker of bone resorption.

- Using numerical and graphical summaries, describe the distribution of each of these variables.
- Using numerical and graphical summaries, describe the relationship between each pair of variables.

### 11.37 Predicting bone formation.

Let's use regression methods to predict  $VO+$ , the measure of bone formation.

- Since OC is a biomarker of bone formation, we start with a simple linear regression using OC as the explanatory variable. Run the regression and summarize the results. Be sure to include an analysis of the residuals.
- Because the processes of bone formation and bone resorption are highly related, it is possible that there is some information in the bone resorption variables that can tell us something about bone formation. Use a model with both OC and TRAP, the biomarker of bone resorption, to predict  $VO+$ . Summarize the results. In the context of this model, it appears that TRAP is a better predictor of bone formation,  $VO+$ , than the biomarker of bone formation, OC. Is this view consistent with the pattern of relationships that you described in the previous exercise? One possible explanation is that, although all these variables are highly related, TRAP is measured with more precision than OC.

### 11.38 More on predicting bone formation.

Now consider a regression model for predicting VO<sup>+</sup> using OC, TRAP, and VO<sup>-</sup>.

- Write out the statistical model for this analysis including all assumptions.
- Run the multiple regression to predict VO<sup>+</sup> using OC, TRAP, and VO<sup>-</sup>. Summarize the results.
- Make a table giving the estimated regression coefficients, standard errors, and *t* statistics with *P*-values for this analysis and for the two that you ran in the previous exercise. Describe how the coefficients and the *P*-values differ for the three analyses.
- Give the percent of variation in VO<sup>+</sup> explained by each of the three models and the estimate of  $\sigma$ . Give a short summary.
- The results you found in part (b) suggest another model. Run that model, summarize the results, and compare them with the results in part (b).

### 11.39 Predicting bone formation using transformed variables.

Because the distributions of VO<sup>+</sup>, VO<sup>-</sup>, OC, and TRAP tend to be skewed, it is common to work with logarithms rather than the measured values. Using the questions in the previous three exercises as a guide, analyze the log data.

### 11.40 Predicting bone resorption.

Refer to Exercises 11.36 to 11.38. Answer these questions with the roles of VO<sup>+</sup> and VO<sup>-</sup> reversed; that is, run models to predict VO<sup>-</sup>, with VO<sup>+</sup> as an explanatory variable.

### 11.41 Predicting bone resorption using transformed variables.

Refer to the previous exercise. Rerun using logs.

*The following 11 exercises use the PCB data file. Polychlorinated biphenyls (PCBs) are a collection of synthetic compounds, called congeners, that are particularly toxic to fetuses and young children. Although PCBs are no longer produced in the United States, they are still found in the environment. Since human exposure to these PCBs is primarily through the consumption of fish, the Environmental Protection Agency (EPA) monitors the PCB levels in fish. Unfortunately, there are 209 different congeners, and measuring all of them in a fish specimen is an expensive and time-consuming process. You've been asked to see if the total amount of PCBs in a specimen can be estimated with only a few, easily quantifiable congeners.<sup>10</sup> If this can be done, costs can be greatly reduced.*  PCB

### 11.42 Relationships among PCB congeners.

Consider the following variables: PCB (the total amount of PCB) and four congeners: PCB52, PCB118, PCB138, and PCB180.

- Using numerical and graphical summaries, describe the distribution of each of these variables.
- Using numerical and graphical summaries, describe the relationship between each pair of variables.

### 11.43 Predicting the total amount of PCB.

Use the four congeners PCB52, PCB118, PCB138, and PCB180 in a multiple regression to predict PCB.

- Write the statistical model for this analysis. Include all assumptions.
- Run the regression and summarize the results.
- Examine the residuals. Do they appear to be approximately Normal? When you plot them versus each of the explanatory variables, are any patterns evident?

### 11.44 Adjusting the analysis for potential outliers.

The examination of the residuals in part (c) of the previous exercise suggests that there may be two outliers, one with a high residual and one with a low residual.

- Because of safety issues, we are more concerned about underestimating PCB in a specimen than about overestimating. Give the specimen number for each of the two suspected outliers. Which one corresponds to an overestimate of PCB?
- Rerun the analysis with the two suspected outliers deleted, summarize these results, and compare them with those you obtained in the previous exercise.

### 11.45 More on predicting the total amount of PCB.

Run a regression to predict PCB using the variables PCB52, PCB118, and PCB138. Note that this is similar to the analysis that you did in Exercise 11.43, with the change that PCB180 is not included as an explanatory variable.

- Summarize the results.
- In this analysis, the regression coefficient for PCB118 is not statistically significant. Give the estimate of the coefficient and the associated  $P$ -value.
- Find the estimate of the coefficient for PCB118 and the associated  $P$ -value for the model analyzed in Exercise 11.43.
- Using the results in parts (b) and (c), write a short paragraph explaining how the inclusion of other variables in a multiple regression can have an effect on the estimate of a particular coefficient and the results of the associated significance test.

### 11.46 Multiple regression model for total TEQ.

Dioxins and furans are other classes of chemicals that can cause undesirable health effects similar to those caused by PCB. The three types of chemicals are combined using toxic equivalent scores (TEQs), which attempt to measure the health effects on a common scale. The PCB data file contains TEQs for PCB, dioxins, and furans. The variables are called TEQPCB, TEQDIOXIN, and TEQFURAN. The data file also includes the total TEQ, defined to be the sum of these three variables.

- Consider using a multiple regression to predict TEQ using the three components TEQPCB, TEQDIOXIN, and TEQFURAN as explanatory variables. Write the multiple regression model in the form

$$\text{TEQ} = \beta_0 + \beta_1 \text{TEQPCB} + \beta_2 \text{TEQDIOXIN} + \beta_3 \text{TEQFURAN} + \varepsilon$$

Give numerical values for the parameters  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$ .

(b) The multiple regression model assumes that the  $\varepsilon$ 's are Normal with mean zero and standard deviation  $\sigma$ . What is the numerical value of  $\sigma$ ?

(c) Use software to run this regression and summarize the results.

### 11.47 Multiple regression model for total TEQ, continued.

The information summarized in TEQ is used to assess and manage risks from these chemicals. For example, the World Health Organization (WHO) has established the tolerable daily intake (TDI) as 1 to 4 TEQs per kilogram of body weight per day. Therefore, it would be very useful to have a procedure for estimating TEQ using just a few variables that can be measured cheaply. Use the four PCB congeners PCB52, PCB118, PCB138, and PCB180 in a multiple regression to predict TEQ. Give a description of the model and assumptions, summarize the results, examine the residuals, and write a summary of what you have found.

### 11.48 Predicting total amount of PCB using transformed variables.

Because distributions of variables such as PCB, the PCB congeners, and TEQ tend to be skewed, researchers frequently analyze the logarithms of the measured variables. Create a data set that has the logs of each of the variables in the PCB data file. Note that zero is a possible value for PCB126; most software packages will eliminate these cases when you request a log transformation.

(a) If you do not do anything about the 16 zero values of PCB126, what does your software do with these cases? Is there an error message of some kind?

(b) If you attempt to run a regression to predict the log of PCB using the log of PCB126 and the log of PCB52, are the cases with the zero values of PCB126 eliminated? Do you think that this is a good way to handle this situation?

(c) The smallest nonzero value of PCB126 is 0.0052. One common practice when taking logarithms of measured values is to replace the zeros by one-half of the smallest observed value. Create a logarithm data set using this procedure; that is, replace the 16 zero values of PCB126 by 0.0026 before taking logarithms. Use numerical and graphical summaries to describe the distributions of the log variables.

### 11.49 Predicting total amount of PCB using transformed variables, continued.

Refer to the previous exercise.

(a) Use numerical and graphical summaries to describe the relationships between each pair of log variables.

(b) Compare these summaries with the summaries that you produced in Exercise 11.42 for the measured variables.

### 11.50 Even more on predicting total amount of PCB using transformed variables.

Use the log data set that you created in Exercise 11.48 to find a good multiple regression model for predicting the log of PCB. Use only log PCB variables for this analysis. Write a report summarizing your results.

### **11.51 Predicting total TEQ using transformed variables.**

Use the log data set that you created in Exercise 11.48 to find a good multiple regression model for predicting the log of TEQ. Use only log PCB variables for this analysis. Write a report summarizing your results and comparing them with the results that you obtained in the previous exercise.

### **11.52 Interpretation of coefficients in log PCB regressions.**

Use the results of your analysis of the log PCB data in Exercise 11.50 to write an explanation of how regression coefficients, standard errors of regression coefficients, and tests of significance for explanatory variables can change depending on what other explanatory variables are included in the multiple regression analysis.

*The following nine exercises use the CHEESE data file. As cheddar cheese matures, a variety of chemical processes take place. The taste of matured cheese is related to the concentration of several chemicals in the final product. In a study of cheddar cheese from the LaTrobe Valley of Victoria, Australia, samples of cheese were analyzed for their chemical composition and were subjected to taste tests. The variable “Case” is used to number the observations from 1 to 30. “Taste” is the response variable of interest. The taste scores were obtained by combining the scores from several tasters. Three of the chemicals whose concentrations were measured were acetic acid, hydrogen sulfide, and lactic acid. For acetic acid and hydrogen sulfide (natural) log transformations were taken. Thus, the explanatory variables are the transformed concentrations of acetic acid (“Acetic”) and hydrogen sulfide (“H2S”) and the untransformed concentration of lactic acid (“Lactic”).<sup>11</sup>*

**CHEESE**

### **11.53 Describing the explanatory variables.**

For each of the four variables in the CHEESE data file, find the mean, median, standard deviation, and interquartile range. Display each distribution by means of a stemplot and use a Normal quantile plot to assess Normality of the data. Summarize your findings. Note that when doing regressions with these data, we do not assume that these distributions are Normal. Only the residuals from our model need to be (approximately) Normal. The careful study of each variable to be analyzed is nonetheless an important first step in any statistical analysis.

### **11.54 Pairwise scatterplots of the explanatory variables.**

Make a scatterplot for each pair of variables in the CHEESE data file (you will have six plots). Describe the relationships. Calculate the correlation for each pair of variables and report the  $P$ -value for the test of zero population correlation in each case.

### **11.55 Simple linear regression model of Taste.**

Perform a simple linear regression analysis using Taste as the response variable and Acetic as the explanatory variable. Be sure to examine the residuals carefully. Summarize your results. Include a plot of the data with the least-squares regression line. Plot the residuals versus each of the other two chemicals. Are any patterns evident? (The concentrations of the other chemicals are lurking variables for the simple linear regression.)

### **11.56 Another simple linear regression model of Taste.**

Repeat the analysis of Exercise 11.55 using Taste as the response variable and H2S as the explanatory variable.

### **11.57 The final simple linear regression model of Taste.**

Repeat the analysis of Exercise 11.55 using Taste as the response variable and Lactic as the explanatory variable.

### **11.58 Comparing the simple linear regression models.**

Compare the results of the regressions performed in the three previous exercises. Construct a table with values of the  $F$  statistic, its  $P$ -value,  $R^2$  and the estimate  $s$  of the standard deviation for each model. Report the three regression equations. Why are the intercepts in these three equations different?

### **11.59 Multiple regression model of Taste.**

Carry out a multiple regression using Acetic and H2S to predict Taste. Summarize the results of your analysis. Compare the statistical significance of Acetic in this model with its significance in the model with Acetic alone as a predictor (Exercise 11.55). Which model do you prefer? Give a simple explanation for the fact that Acetic alone appears to be a good predictor of Taste, but with H2S in the model, it is not.

### **11.60 Another multiple regression model of Taste.**

Carry out a multiple regression using H2S and Lactic to predict Taste. When we compare the results of this analysis with the simple linear regressions using each of these explanatory variables alone, it is evident that a better result is obtained by using both predictors in a model. Support this statement with explicit information obtained from your analysis.

### **11.61 The final multiple regression model of Taste.**

Use the three explanatory variables Acetic, H2S, and Lactic in a multiple regression to predict Taste. Write a short summary of your results, including an examination of the residuals. Based on all the regression analyses you have carried out on these data, which model do you prefer and why?

### **11.62 Finding a multiple regression model on the Internet.**

Search the Internet to find an example of the use of multiple regression. Give the setting of the example, describe the data, give the model, and summarize the results. Explain why the use of multiple regression in this setting was appropriate or inappropriate.

# 12 One-Way Analysis of Variance<sub>CHAPTER</sub>



**12.1 Inference for One-Way Analysis of Variance**

**12.2 Comparing the Means**

## Introduction

Many of the most effective statistical studies are comparative. For example, we may wish to compare customer satisfaction of men and women who use an online fantasy football site or compare the responses to various treatments in a clinical trial. With a quantitative response, we display these comparisons with back-to-back stemplots or side-by-side boxplots, and we measure them with five-number summaries or with means and standard deviations.

When only two groups are compared, Chapter 7 provides the tools we need to answer the question “Is the difference between groups statistically significant?” Two-sample  $t$  procedures compare the means of two Normal populations, and we saw that these procedures, unlike comparisons of spread, are sufficiently robust to be widely useful.

In this chapter, we will compare any number of means by techniques that generalize the two-sample  $t$  test and share its robustness and usefulness. These methods will allow us to address comparisons such as

- How does a user’s number of Facebook friends affect his or her social attractiveness?
- On average, which of 5 brands of automobile tires wears longest?
- Among three therapies for lung cancer, is there a difference in average progression-free survival?

## 12.1 Inference for One-Way Analysis of Variance

When you complete this section, you will be able to

- Describe the one-way ANOVA model and when it is used for inference.
- Provide a description of the underlying idea of the ANOVA  $F$  test in terms of the variation between population means and the variation within populations.
- Summarize what the ANOVA  $F$  test can tell you about the population means and what it cannot.
- Construct an ANOVA table in terms of sources of variation and degrees of freedom. Compute mean squares and the  $F$  statistic when provided various sums of squares.
- Interpret statistical software ANOVA output to obtain the ANOVA  $F$  test results and the coefficient of determination.
- Use diagnostic plots and population sample statistics to check the assumptions of the one-way ANOVA model.

When comparing different populations or treatments, the data are subject to sampling variability. For example, we would not expect to observe exactly the same sales data if we mailed an advertising offer to different random samples of households. We also wouldn't expect a new group of cancer patients to provide the same set of progression-free survival times. We therefore pose the question for inference in terms of the *mean* response.



**comparing two means, p. 447**

In Chapter 7 we met procedures for comparing the means of two populations. We now extend those methods to problems involving more than two populations. The statistical methodology for comparing several means is called **analysis of variance**, or simply **ANOVA**. In this and the following section, we will examine the basic ideas and assumptions that are needed for ANOVA. Although the details differ, many of the concepts are similar to those discussed in the two-sample case.

### ANOVA

We will consider two ANOVA techniques. When there is only one way to classify the populations of interest, we use **one-way ANOVA** to analyze the data.

We call this categorical explanatory variable a **factor**. For example, to compare the average tread lifetimes of 5 specific brands of tires we use one-way ANOVA with tire brand as our factor. This chapter presents the details for one-way ANOVA.

### one-way ANOVA

#### factor

In many other comparative studies, there is more than one way to classify the populations. For the tire study, the researcher may also want to consider temperature. Are there brands that do relatively better in the heat? Analyzing the effect of two factors, brand and temperature, requires **two-way ANOVA**. This technique will be discussed in Chapter 13.

### two-way ANOVA

## Data for one-way ANOVA

One-way analysis of variance is a statistical method for comparing several population means. We draw a simple random sample (SRS) from each population and use the data to test the null hypothesis that the population means are all equal. Consider the following two examples.

### EXAMPLE

#### 12.1 Does haptic feedback improve performance?



A group of technology students is interested in whether haptic feedback (forces and vibrations applied through a joystick) is helpful in navigating a simulated game environment they created. To investigate this, they randomly assign each of 60 students to one of three joystick controller types and record the time it takes to complete a navigation mission. The joystick types are (1) a standard video game joystick, (2) a game joystick with force feedback, and (3) a game joystick with vibration feedback.

## EXAMPLE

### 12.2 Average age of coffeehouse customers.

How do five coffeehouses around campus differ in the demographics of their customers? Are certain coffeehouses more popular among graduate students? Do professors tend to favor one coffeehouse? A market researcher asks 50 customers of each coffeehouse to respond to a questionnaire. One variable of interest is the customer's age.

These two examples are similar in that

- There is a single quantitative response variable measured on many units; the units are students in the first example and customers in the second.

- The goal is to compare several populations: students using different joystick types in the first example and customers of five coffeehouses in the second.



### observation versus experiment, p. 172

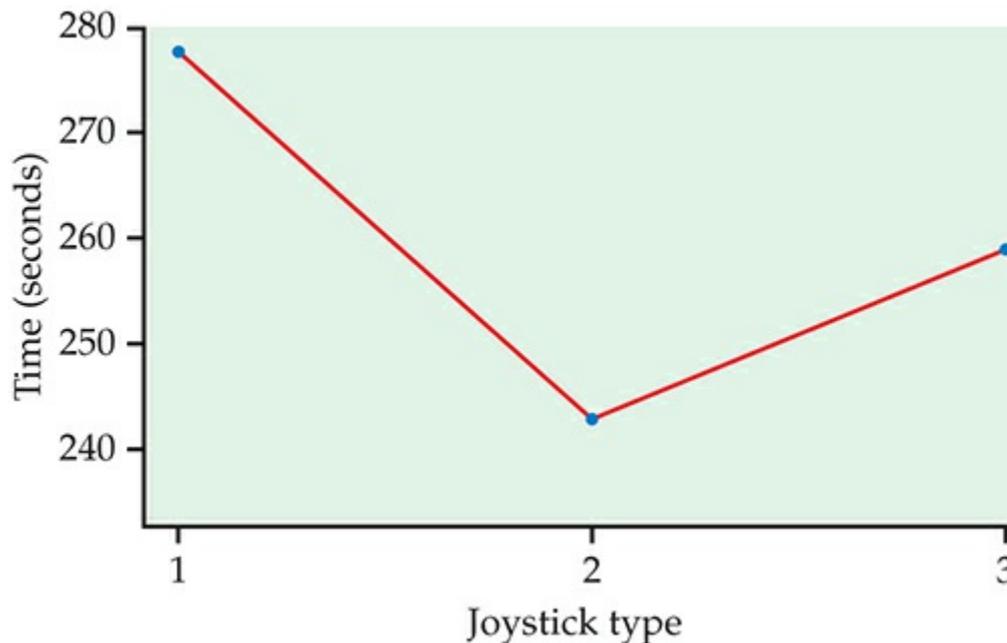
There is, however, an important difference. Example 12.1 describes an experiment in which each student is randomly assigned to a joystick type. Example 12.2 is an observational study in which customers are selected during a particular time period and not all agree to provide data. These samples of customers are not random samples, but we will treat them as such because we believe that the selective sampling and nonresponse are ignorable sources of bias. This will not always be the case. *Always consider the various sources of bias in an observational study.*



In both examples, we will use ANOVA to compare the mean responses. The same ANOVA methods apply to data from random samples and to data from randomized experiments. *It is important to keep the data-production method in mind when interpreting the results.* A strong case for causation is best made by a randomized experiment.

## Comparing means

The question we ask in ANOVA is “Do all groups have the same population mean?” We will often use the term “groups” for the populations to be compared in a one-way ANOVA. To answer this question we compare the sample means. Figure 12.1 displays the sample means for Example 12.1. It appears that a joystick with force feedback has the shortest average completion time. But is the observed difference in sample means just the result of chance variation? We should not expect sample means to be equal even if the population means are all identical.



**FIGURE 12.1**  
Average completion time for three different joystick types.

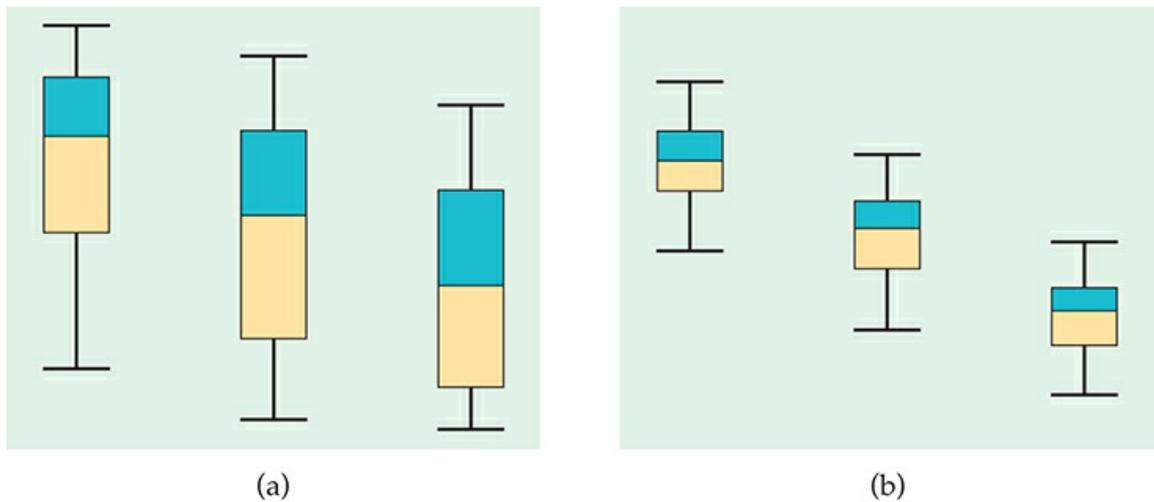


**standard deviation of  $\bar{x}$ , p. 306**

The purpose of ANOVA is to assess whether the observed differences among sample means are *statistically significant*. Could a variation among the three sample means this large be plausibly due to chance, or is it good evidence for a difference among the population means? This question can't be answered from the sample means alone. Because the standard deviation of a sample mean  $\bar{x}$  is the population standard deviation  $\sigma$  divided by  $n$ , the answer also depends upon both the variation within the groups of observations and the sizes of the samples.

Side-by-side boxplots help us see the within-group variation. Compare Figures 12.2(a) and 12.2(b). The sample medians are the same in both figures, but the large variation within the groups in Figure 12.2(a) suggests that the differences among the sample medians could be due simply to chance variation. The data in Figure 12.2(b) are much more convincing evidence that the populations differ.

Even the boxplots omit essential information, however. To assess the observed differences, we must also know how large the samples are. Nonetheless, boxplots are a good preliminary display of the data.



**FIGURE 12.2** (a) Side-by-side boxplots for three groups with large within-group variation. The differences among centers may be just chance variation. (b) Side-by-side boxplots for three groups with the same centers as in panel (a) but with small within-group variation. The differences among centers are more likely to be significant.

Although ANOVA compares means and boxplots display medians, these two measures of center will be close together for distributions that are nearly symmetric. If the distributions are not symmetric, we may consider a transformation prior to displaying the data.

← **LOOK BACK**

transforming data, p. 436

### The two-sample $t$ statistic

Two-sample  $t$  statistics compare the means of two populations. If the two populations are assumed to have equal but unknown standard deviations and the sample sizes are both equal to  $n$ , the  $t$  statistic is

← **LOOK BACK**

pooled two-sample  $t$  statistic, p. 461

$$t = \frac{\bar{x}_1 - \bar{x}_2 - \mu_0}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{\bar{x}_1 - \bar{x}_2 - \mu_0}{\sqrt{s_p^2 \left( \frac{2}{n} \right)}}$$

The square of this  $t$  statistic is

$$t^2 = \frac{n(\bar{x}_1 - \bar{x}_2 - \mu_0)^2}{2s_p^2}$$

If we use ANOVA to compare two populations, the ANOVA  $F$  statistic is exactly equal to this  $t^2$ . We can therefore learn something about how ANOVA works by

looking carefully at the statistic in this form.

The numerator in the  $t^2$  statistic measures the variation **between** (or among) the groups in terms of the difference between their sample means  $\bar{x}_1$  and  $\bar{x}_2$  and the common sample size  $n$ . The numerator can be large because of a large difference between the sample means or because the common sample size is large.

### between group variation

The denominator measures the variation **within** groups by  $sp^2$ , the pooled estimator of the common variance. If the within-group variation is small, the same variation between the groups produces a larger statistic and thus a more significant result.

### within group variation

Although the general form of the  $F$  statistic is more complicated, the idea is the same. To assess whether several populations all have the same mean, we compare the variation *among* the means of several groups with the variation *within* groups. Because we are comparing variation, the method is called *analysis of variance*.

## An overview of ANOVA

ANOVA tests the null hypothesis that the population means are *all* equal. The alternative is that they are not all equal. This alternative could be true because all the means are different or simply because one of them differs from the rest. This is a more complex situation than comparing just two populations. If we reject the null hypothesis, we need to perform some further analysis to draw conclusions about which population means differ from which others and by how much.

The computations needed for an ANOVA are more lengthy than those for the  $t$  test. For this reason we generally use computer programs to perform the calculations. Automating the calculations frees us from the burden of arithmetic and allows us to concentrate on interpretation.



*Complicated computations do not guarantee a valid statistical analysis.* We should always start our ANOVA with a careful examination of the data using graphical and numerical summaries. Just as in linear regression, outliers and extreme deviations from Normality can invalidate the computed results.

## EXAMPLE

### 12.3 Number of Facebook friends.

A feature of each Facebook user's profile is the number of Facebook "friends," an indicator of the user's social network connectedness. Among college students on Facebook, the average number of Facebook friends has been estimated to be around 281.<sup>1</sup>

Offline, having more friends is associated with higher ratings of positive attributes such as likability and trustworthiness. Is this also the case with Facebook friends?



An experiment was run to examine the relationship between the number of Facebook friends and the user's perceived social attractiveness.<sup>2</sup> A total of 134 undergraduate participants were randomly assigned to observe one of five Facebook profiles. Everything about the profile was the same except the number of friends, which appeared on the profile as 102, 302, 502, 702, or 902.

After viewing the profile, each participant was asked to fill out a questionnaire on the physical and social attractiveness of the profile user. Each attractiveness score is an average of several seven-point questionnaire items, ranging from 1 (strongly disagree) to 7 (strongly agree). Here is a summary of the data for the social attractiveness score:

Number of friends	<i>n</i>	$\bar{x}$	<i>s</i>
102	24	3.82	1.00
302	33	4.88	0.85
502	26	4.56	1.07
702	30	4.41	1.43
902	21	3.99	1.02

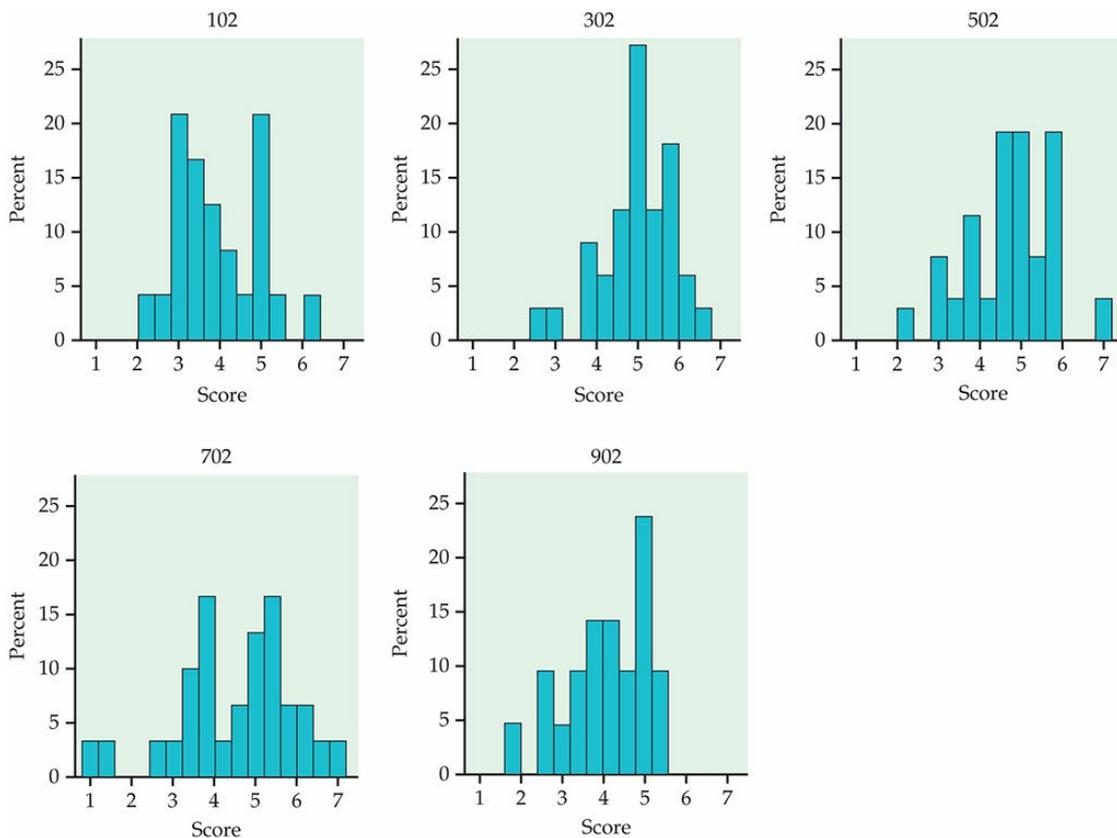
Histograms for the five groups are given in Figure 12.3. Note that the heights of the bars in the histograms are percents rather than counts. This is commonly done when the group sample sizes vary. Figure 12.4 gives side-by-side boxplots for these data. We see that the scores covered the entire range of possible values, from 1.0 to 7.0. We also see a lot of overlap in scores across groups. The histograms are relatively symmetric, and with the group sample sizes all more than 15, we can feel confident that the sample means are

approximately Normal.

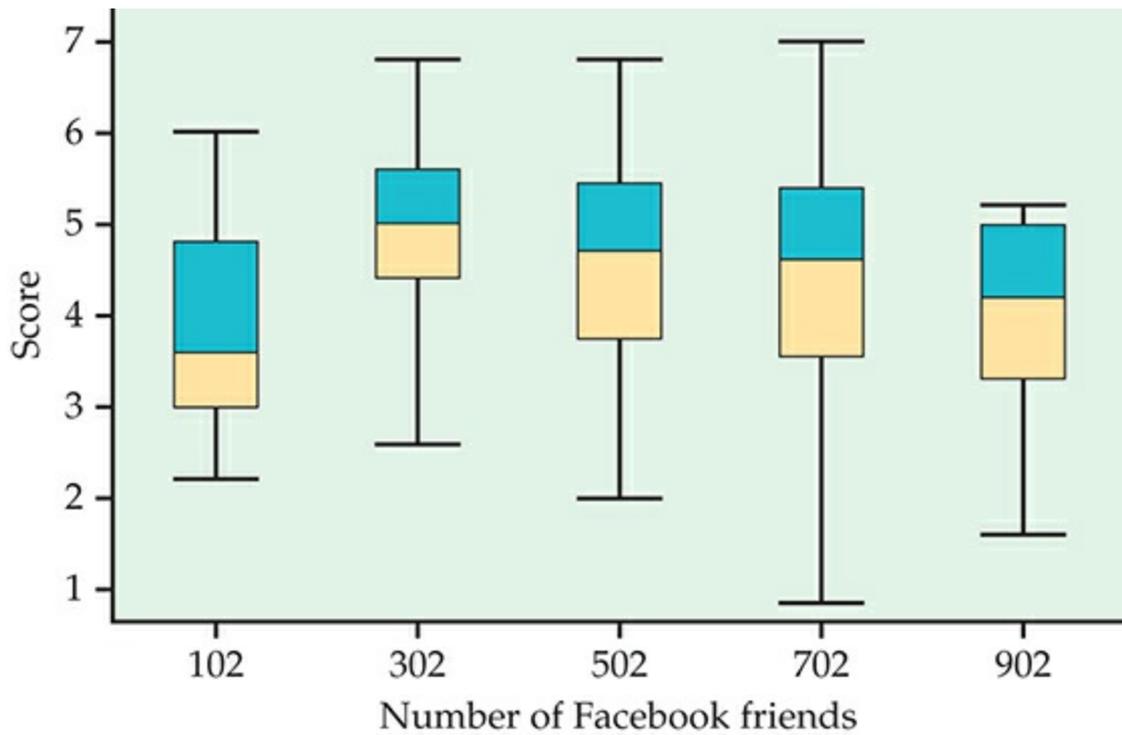


**guidelines for two-sample  $t$  procedures, p. 456**

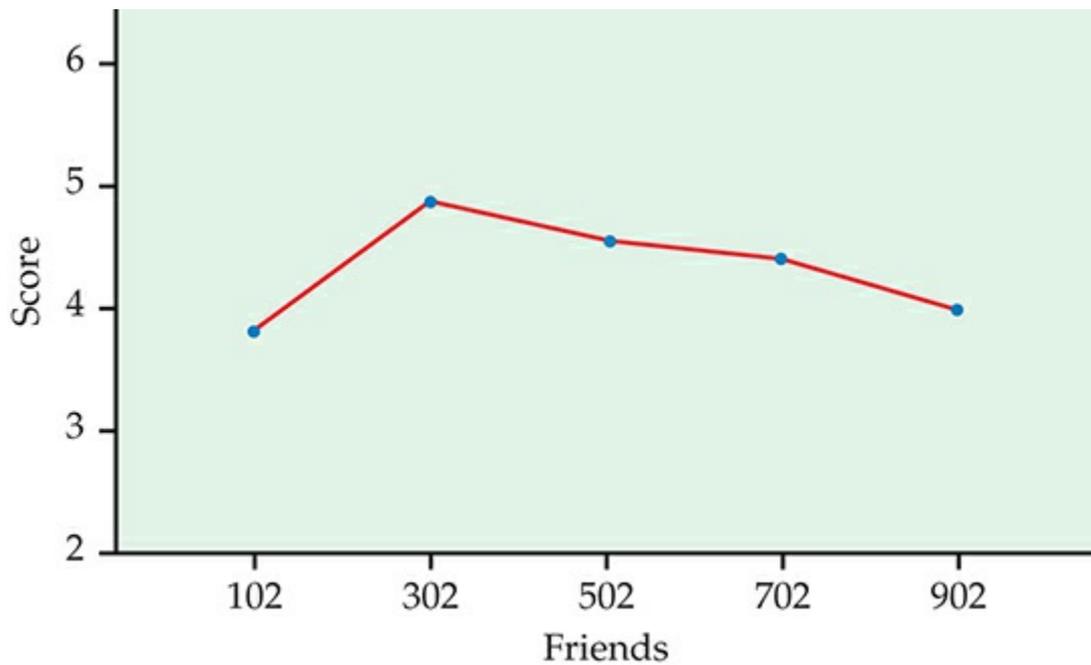
The five sample means are plotted in Figure 12.5 (page 650). They rise and then fall as the number of friends increases. This suggests that having too many Facebook friends can harm a user's social attractiveness. However, given the variability in the data, this pattern could also just be the result of chance variation. We will use ANOVA to make this determination.



**FIGURE 12.3**  
Histograms for the Facebook friends example.



**FIGURE 12.4**  
Side-by-side boxplots for the Facebook friends example.



**FIGURE 12.5**  
Social attractiveness means for the Facebook friends example.

In this setting, we have an experiment in which undergraduate Facebook users were randomly assigned to view one of five Facebook profiles. Each of these profile populations has a mean, and our inference asks questions about these means. The undergraduates in this study were all from the same university. They also volunteered in exchange for course credit.



*Formulating a clear definition of the populations being compared with ANOVA can be difficult.* Often some expert judgment is required, and different consumers of the results may have differing opinions. Whether one can consider the samples in this study as SRSs from the population of undergraduates at the university or from the population of all undergraduates is open for debate. Regardless, we are more confident in generalizing our conclusions to similar populations when the results are clearly significant than when the level of significance just barely passes the standard of  $P = 0.05$ .

We first ask whether or not there is sufficient evidence in the data to conclude that the corresponding population means are not all equal. Our null hypothesis here states that the population mean score is the same for all five Facebook profiles. The alternative is that they are not all the same.



Our inspection of the data for our example suggests that the means may follow a curvilinear relationship. *Rejecting the null hypothesis that the means are all the same using ANOVA is not the same as concluding that all the means are different from one another.* The ANOVA null hypothesis can be false in many different ways. Additional analysis is required to distinguish among these possibilities.



When there are particular versions of the alternative hypothesis that are of interest, we use **contrasts** to examine them. In our example, we might want to test whether there is a curvilinear relationship between the number of friends and attractiveness score. *Note that, to use contrasts, it is necessary that the questions of interest be formulated before examining the data.* It is cheating to make up these questions after analyzing the data.

#### contrasts

If we have no specific relations among the means in mind before looking at the data, we instead use a **multiple-comparisons** procedure to determine which pairs of population means differ significantly. In the next section we will explore both contrasts and multiple comparisons in detail.

## USE YOUR KNOWLEDGE

### 12.1 What's wrong?

For each of the following, explain what is wrong and why.

- (a) ANOVA tests the null hypothesis that the sample means are all equal.
- (b) A strong case for causation is best made in an observational study.
- (c) You use ANOVA to compare the variances of the populations.
- (d) A multiple-comparisons procedure is used to compare a relation among means that was specified prior to looking at the data.

### 12.2 What's wrong?

For each of the following, explain what is wrong and why.

- (a) In rejecting the null hypothesis, one can conclude that all the means are different from one another.
- (b) A one-way ANOVA can be used only when there are two means to be compared.
- (c) The ANOVA  $F$  statistic will be large when the within-group variation is much larger than the between-group variation.

## The ANOVA model

When analyzing data, the following equation reminds us that we look for an overall pattern and deviations from it:

$$\text{DATA} = \text{FIT} + \text{RESIDUAL}$$



**DATA = FIT + RESIDUAL, p. 567**

In the regression model of Chapter 10, the FIT was the population regression line, and the RESIDUAL represented the deviations of the data from this line. We now apply this framework to describe the statistical models used in ANOVA. These models provide a convenient way to summarize the assumptions that are the foundation for our analysis. They also give us the necessary notation to describe

the calculations needed.



## Normal distributions, p. 58

First, recall the statistical model for a random sample of observations from a single Normal population with mean  $\mu$  and standard deviation  $\sigma$ . If the observations are

$$x_1, x_2, \dots, x_n$$

we can describe this model by saying that the  $x_j$  are an SRS from the  $N(\mu, \sigma)$  distribution. Another way to describe the same model is to think of the  $x$ 's varying about their population mean. To do this, write each observation  $x_j$  as

$$x_j = \mu + \varepsilon_j$$

The  $\varepsilon_j$  are then an SRS from the  $N(0, \sigma)$  distribution. Because  $\mu$  is unknown, the  $\varepsilon$ 's cannot actually be observed. This form more closely corresponds to our

$$\text{DATA} = \text{FIT} + \text{RESIDUAL}$$

way of thinking. The FIT part of the model is represented by  $\mu$ . It is the systematic part of the model, like the line in a regression. The RESIDUAL part is represented by  $\varepsilon_j$ . It represents the deviations of the data from the fit and is due to random, or chance, variation.

There are two unknown parameters in this statistical model:  $\mu$  and  $\sigma$ . We estimate  $\mu$  by  $\bar{x}$ , the sample mean, and  $\sigma$  by  $s$ , the sample standard deviation. The differences  $\varepsilon_j = x_j - \bar{x}$  are the residuals and correspond to the  $\varepsilon_j$  in the statistical model.

The model for one-way ANOVA is very similar. We take random samples from each of  $I$  different populations. The sample size is  $n_i$  for the  $i$ th population. Let  $x_{ij}$  represent the  $j$ th observation from the  $i$ th population. The  $I$  population means are the FIT part of the model and are represented by  $\mu_i$ . The random variation, or RESIDUAL, part of the model is represented by the deviations  $\varepsilon_{ij}$  of the observations from the means.

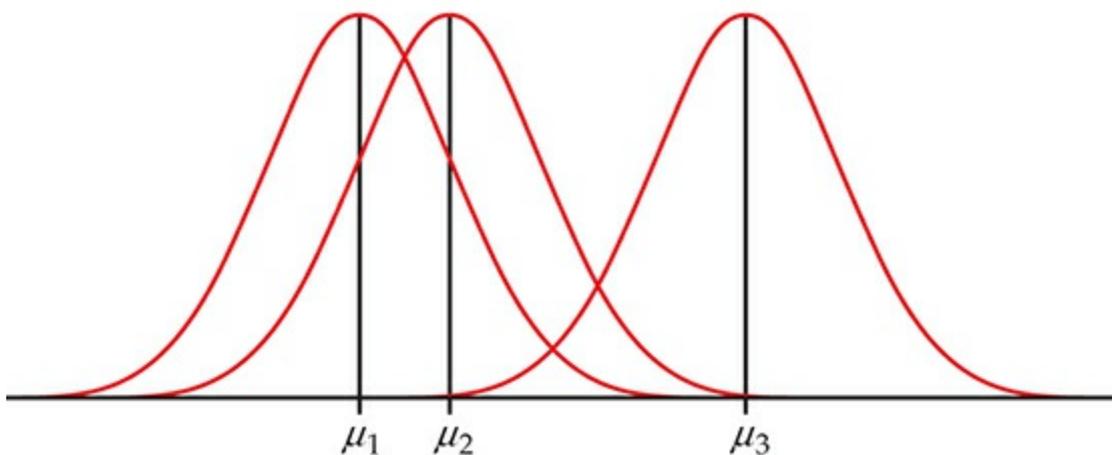
## THE ONE-WAY ANOVA MODEL

The **one-way ANOVA model** is

$$x_{ij} = \mu_i + \varepsilon_{ij}$$

for  $i = 1, \dots, I$  and  $j = 1, \dots, n_i$ . The  $\varepsilon_{ij}$  are assumed to be from an  $N(0, \sigma)$  distribution. The **parameters of the model** are the population means  $\mu_1, \mu_2, \dots, \mu_I$  and the common standard deviation  $\sigma$ .

Note that the sample sizes  $n_i$  may differ, but the standard deviation  $\sigma$  is assumed to be the same in all the populations. Figure 12.6 pictures this model for  $I = 3$ . The three population means  $\mu_i$  are different, but the shapes of the three Normal distributions are the same, reflecting the assumption that all three populations have the same standard deviation.



**FIGURE 12.6**

Model for one-way ANOVA with three groups. The three populations have Normal distributions with the same standard deviation.

## EXAMPLE

### 12.4 ANOVA model for the Facebook friends study.

In the Facebook friends example, there are five profiles that we want to compare, so  $I = 5$ . The population means  $\mu_1, \mu_2, \dots, \mu_5$  are the mean social attractiveness scores for the profiles with 102, 302, 502, 702, and 902 friends, respectively. The sample sizes  $n_i$  are 24, 33, 26, 30, and 21. It is common to use numerical subscripts to distinguish the different means, and some software requires that levels of factors in ANOVA be specified as numerical values. In this situation, it is very important to keep track of what each numerical value represents when drawing conclusions. In our example, we could use numerical

values to suggest the actual groups by replacing  $\mu_1$  with  $\mu_{102}$ ,  $\mu_2$  with  $\mu_{302}$ , and so on.

The observation  $x_{1,1}$  is the social attractiveness score for the first participant who observed the profile with 102 friends. The data for the other participants assigned to this profile are denoted by  $x_{1,2}, x_{1,3}, \dots, x_{1,24}$ . Similarly, the data for the other four groups have a first subscript indicating the profile and a second subscript indicating the participants assigned to that profile.

According to our model, the score for the first participant is  $x_{1,1} = \mu_1 + \varepsilon_{1,1}$ , where  $\mu_1$  is the average score for *all* undergraduates after viewing Profile 1 and  $\varepsilon_{1,1}$  is the chance variation due to this particular participant. Similarly, the score for the last participant who observed the profile with 902 friends is  $x_{5,21} = \mu_5 + \varepsilon_{5,21}$ , where  $\mu_5$  is the average score for all undergraduates after viewing Profile 5, and  $\varepsilon_{5,21}$  is the chance variation due to this participant.



**central limit theorem, p. 307**

The ANOVA model assumes that these chance variations  $\varepsilon_{ij}$  are independent and Normally distributed with mean 0 and standard deviation  $\sigma$ . For our example, we have clear evidence that the data are non-Normal. The observed scores are numbers ranging from 1.0 to 7.0 by increments of 0.2. However, because our inference is based on the sample means, which will be approximately Normally distributed, we are not overly concerned about this violation of model assumptions.

## Estimates of population parameters

The unknown parameters in the statistical model for ANOVA are the  $I$  population means  $\mu_i$  and the common population standard deviation  $\sigma$ . To estimate  $\mu_i$  we use the sample mean for the  $i$ th group:

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

The residuals  $e_{ij} = x_{ij} - \bar{x}_i$  reflect the variation about the sample means that we see in the data and are used in the calculations of the sample standard deviations

$$s_i = \sqrt{\frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{n_i - 1}}$$

The ANOVA model assumes that the population standard deviations are all equal. Before estimating  $\sigma$ , it is important to check this equality assumption using the sample standard deviations. Unfortunately, formal tests for the equality of standard deviations in several groups share the lack of robustness against non-

Normality that we noted in Chapter 7 for the case of two groups.



**F test for equality of spread, p. 474**

ANOVA procedures, however, are not extremely sensitive to unequal standard deviations provided the group sample sizes are the same or similar. Thus, we do not recommend a formal test of equality of standard deviations as a preliminary to the ANOVA. Instead, we will use the following rule as a guideline.

**RULE FOR EXAMINING STANDARD DEVIATIONS IN ANOVA**

If the largest standard deviation is less than twice the smallest standard deviation, we can use methods based on the assumption of equal standard deviations, and our results will still be approximately correct.<sup>3</sup>

When we assume that the population standard deviations are equal, each sample standard deviation is an estimate of  $\sigma$ . To combine these into a single estimate, we use a generalization of the pooling method introduced in Chapter 7 (page 461).

**POOLED ESTIMATOR OF  $\sigma$**

Suppose that we have sample variances  $s_1^2, s_2^2, \dots, s_I^2$  from  $I$  independent SRSs of sizes  $n_1, n_2, \dots, n_I$  from populations with common variance  $\sigma^2$ . The **pooled sample variance**

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \dots + (n_I - 1)s_I^2}{(n_1 - 1) + (n_2 - 1) + \dots + (n_I - 1)}$$

is an unbiased estimator of  $\sigma^2$ . The **pooled standard deviation**

$$s_p = \sqrt{s_p^2}$$

is the estimate of  $\sigma$ .



Pooling gives more weight to groups with larger sample sizes. If the sample

sizes are equal,  $s_p^2$  is just the average of the  $I$  sample variances. *Note that  $s_p$  is not the average of the  $I$  sample standard deviations.*

If it appears that we have unequal standard deviations, we generally try to transform the data so that they are approximately equal. We might, for example, work with  $x_{ij}$  or  $\log x_{ij}$ . Fortunately, we can often find a transformation that *both* makes the group standard deviations more nearly equal and also makes the distributions of observations in each group more nearly Normal. If the standard deviations are markedly different and cannot be made similar by a transformation, inference requires different methods such as the bootstrap described in Chapter 16.

## EXAMPLE

### 12.5 Population estimates for the Facebook friends study.

In the Facebook friends study there are  $I = 5$  groups and the sample sizes are  $n_1 = 24$ ,  $n_2 = 33$ ,  $n_3 = 26$ ,  $n_4 = 30$ , and  $n_5 = 21$ . The sample standard deviations are  $s_1 = 1.00$ ,  $s_2 = 0.85$ ,  $s_3 = 1.07$ ,  $s_4 = 1.43$  and  $s_5 = 1.02$ .

Because the largest standard deviation (1.43) is less than twice the smallest ( $2 \times 0.85 = 1.70$ ), our rule indicates that we can use the assumption of equal population standard deviations.

The pooled variance estimate is

$$\begin{aligned} s_p^2 &= \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + (n_3-1)s_3^2 + (n_4-1)s_4^2 + (n_5-1)s_5^2}{(n_1-1) + (n_2-1) + (n_3-1) + (n_4-1) + (n_5-1)} \\ &= \frac{(23)(1.00)^2 + (32)(0.85)^2 + (25)(1.07)^2 + (29)(1.43)^2 + (20)(1.02)^2}{23+32+25+29+20} \\ &= 154.85129 = 1.20 \end{aligned}$$

The pooled standard deviation is

$$s_p = 1.20 = 1.10$$

This is our estimate of the common standard deviation  $\sigma$  of the social attractiveness scores for the five profiles.

## USE YOUR KNOWLEDGE

### 12.3 Computing the pooled standard deviation.

An experiment was run to compare three timed-release fertilizers in terms of plant growth. The sample sizes were 23, 21, and 27 plants, and the corresponding estimated standard deviations were 4, 5, and 7 centimeters.

- Is it reasonable to use the assumption of equal standard deviations when we analyze these data? Give a reason for your answer.
- Give the values of the variances for the three groups.
- Find the pooled variance.
- What is the value of the pooled standard deviation?

### 12.4 Visualizing the ANOVA model.

For each of the following situations, draw a picture of the ANOVA model similar to Figure 12.6 (page 652). Use the numerical values for the  $\mu_i$ . To sketch the Normal curves, you may want to review the 68–95–99.7 rule on page 59.

- $\mu_1 = 18$ ,  $\mu_2 = 13$ ,  $\mu_3 = 14$ , and  $\sigma = 5$ .
- $\mu_1 = 18$ ,  $\mu_2 = 14$ ,  $\mu_3 = 16$ ,  $\mu_4 = 24$ , and  $\sigma = 7$ .
- $\mu_1 = 18$ ,  $\mu_2 = 13$ ,  $\mu_3 = 14$ , and  $\sigma = 2$ .

## Testing hypotheses in one-way ANOVA

Comparison of several means is accomplished by using an  $F$  statistic to compare the variation among groups with the variation within groups. We now show how the  $F$  statistic expresses this comparison. Calculations are organized in an ANOVA table, which contains numerical measures of the variation among groups and within groups.



### ANOVA table, p. 589

First, we must specify our hypotheses for one-way ANOVA. As before,  $I$  represents the number of populations to be compared.

## HYPOTHESES FOR ONE-WAY ANOVA

The **null and alternative hypotheses** for one-way ANOVA are

$$H_0: \mu_1 = \mu_2 = \dots = \mu_I$$

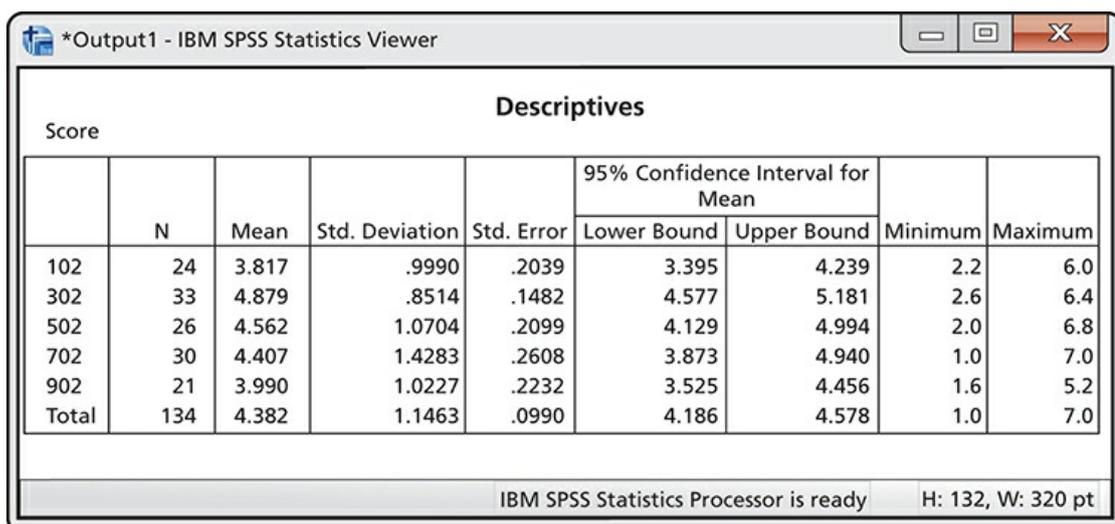
$H_a$ : not all of the  $\mu_i$  are equal

We will now use the Facebook friends example to illustrate how to do a one-way ANOVA. Because the calculations are generally performed using statistical software, we focus on interpretation of the output.

## EXAMPLE

### 12.6 Reading software output.

Figure 12.7 gives descriptive statistics generated by SPSS for the ANOVA of the Facebook friends example. Summaries for each profile are given on the first five lines. In addition to the sample size, the mean, and the standard deviation, this output also gives the minimum and maximum observed value, standard error of the mean, and the 95% confidence interval for the mean of each profile. The five sample means  $\bar{x}_i$  given in the output are estimates of the five unknown population means  $\mu_i$ .



The screenshot shows a window titled '\*Output1 - IBM SPSS Statistics Viewer'. Inside, there is a table titled 'Descriptives' for the variable 'Score'. The table has 10 columns: N, Mean, Std. Deviation, Std. Error, 95% Confidence Interval for Mean (with sub-columns for Lower Bound and Upper Bound), Minimum, and Maximum. There are six rows of data, including a 'Total' row.

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
102	24	3.817	.9990	.2039	3.395	4.239	2.2	6.0
302	33	4.879	.8514	.1482	4.577	5.181	2.6	6.4
502	26	4.562	1.0704	.2099	4.129	4.994	2.0	6.8
702	30	4.407	1.4283	.2608	3.873	4.940	1.0	7.0
902	21	3.990	1.0227	.2232	3.525	4.456	1.6	5.2
Total	134	4.382	1.1463	.0990	4.186	4.578	1.0	7.0

**FIGURE 12.7**

SPSS output with descriptive statistics for the Facebook friends example.

The output gives the estimates of the standard deviations,  $s_j$ , for each group

but does not provide  $s_p$ , the pooled estimate of the model standard deviation,  $\sigma$ . We could perform the calculation using a calculator, as we did in Example 12.5. We will see an easier way to obtain this quantity from the ANOVA table in Figure 12.8.



Some software packages report  $s_p$  as part of the standard ANOVA output. *Sometimes you are not sure whether or not a quantity given by software is what you think it is.* A good way to resolve this dilemma is to do a sample calculation with a simple example to check the numerical results.



*Note that  $s_p$  is not the standard deviation given in the “Total” row of Figure 12.7.* This quantity is the standard deviation that we would obtain if we viewed the data as a single sample of 134 participants and ignored the possibility that the profile means could be different. As we have mentioned many times before, it is important to use care when reading and interpreting software output.

## EXAMPLE

### 12.7 Reading software output, continued.

Additional output generated by SPSS for the ANOVA of the Facebook friends example is given in Figure 12.8. We will discuss the construction of this output next. For now, we observe that the results of our significance test are given in the last two columns of the output. The null hypothesis that the five population means are the same is tested by the statistic  $F = 4.142$ , and the associated  $P$ -value is reported as  $P = 0.003$ . The data provide clear evidence to support the claim that there are some differences among the five profile population means.

The screenshot shows a window titled '\*Output1 - IBM SPSS Statistics Viewer'. Inside, there is an ANOVA table for the variable 'Score'. The table has six columns: Source, Sum of Squares, df, Mean Square, F, and Sig. The rows are: Between Groups, Within Groups, and Total. Below the table, a status bar indicates 'IBM SPSS Statistics Processor is ready' and 'H: 132, W: 320 pt'.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	19.890	4	4.973	4.142	.003
Within Groups	154.867	129	1.201		
Total	174.757	133			

**FIGURE 12.8**  
SPSS output giving the ANOVA table for the Facebook friends example.

## The ANOVA table

The information in an analysis of variance is organized in an ANOVA table. To understand the table, it is helpful to think in terms of our

$$\text{DATA} = \text{FIT} + \text{RESIDUAL}$$

view of statistical models. For one-way ANOVA, this corresponds to

$$x_{ij} = \mu_i + \varepsilon_{ij}$$

We can think of these three terms as sources of variation. The ANOVA table separates the variation in the data into two parts: the part due to the fit and the remainder, which we call residual.

## EXAMPLE

### 12.8 ANOVA table for the Facebook friends study.

The SPSS output in Figure 12.8 gives the sources of variation in the first column. Here, FIT is called Between Groups, RESIDUAL is called Within Groups, and DATA is the last entry, Total. Different software packages use different terms for these sources of variation but the basic concept is common to all. In place of FIT, some software packages use Between Groups, Model, or the name of the factor. Similarly, terms like Within Groups or Error are

frequently used in place of RESIDUAL.

The Between Groups row in the table gives information related to the variation **among** group means. In writing ANOVA tables, for this row we will use the generic label “groups” or some other term that describes the factor being studied.

#### variation among groups

The Within Groups row in the table gives information related to the variation **within** groups. We noted that the term “error” is frequently used for this source of variation, particularly for more general statistical models. This label is most appropriate for experiments in the physical sciences where the observations within a group differ because of measurement error. In business and the biological and social sciences, on the other hand, the within-group variation is often due to the fact that not all firms or plants or people are the same. This sort of variation is not due to errors and is better described as “residual” or “within-group” variation. Nevertheless, we will use the generic label “error” for this source of variation in writing ANOVA tables.

#### variation within groups

Finally, the Total row in the ANOVA table corresponds to the DATA term in our  $DATA = FIT + RESIDUAL$  framework. So, for analysis of variance,

$$DATA = FIT + RESIDUAL$$

translates into

$$\text{Total} = \text{Between Groups} + \text{Within Groups}$$



#### sum of squares, p. 587

The second column in the software output given in Figure 12.8 is labeled Sum of Squares. As you might expect, each sum of squares is a sum of squared deviations. We use SSG, SSE, and SST for the entries in this column, corresponding to groups, error, and total. Each sum of squares measures a different type of variation. SST measures variation of the data around the overall mean,  $x_{ij} - \bar{x}$ . Variation of the group means around the overall mean,  $\bar{x}_i - \bar{x}$  is measured by SSG. Finally, SSE measures variation of each observation around its group mean,  $x_{ij} - \bar{x}_i$ .

## EXAMPLE

### 12.9 ANOVA table for the Facebook friends study, continued.

The Sum of Squares column in Figure 12.8 gives the values for the three sums of squares. They are

$$SST = 174.757$$

$$SSG = 19.890$$

$$SSE = 154.867$$

Verify that  $SST = SSG + SSE$  for this example.

This fact is true in general. The total variation is always equal to the among-group variation plus the within-group variation. Note that software output frequently gives many more digits than we need, as in this case.

In this example it appears that most of the variation is coming from within groups. However, to assess whether the observed differences in sample means are statistically significant, some additional calculations are needed.



### degrees of freedom, p. 44

Associated with each sum of squares is a quantity called the degrees of freedom. Because SST measures the variation of all  $N$  observations around the overall mean, its degrees of freedom are  $DFT = N - 1$ . This is the same as the degrees of freedom for the ordinary sample variance with sample size  $N$ . Similarly, because SSG measures the variation of the  $I$  sample means around the overall mean, its degrees of freedom are  $DFG = I - 1$ . Finally, SSE is the sum of squares of the deviations  $x_{ij} - \bar{x}_i$ . Here we have  $N$  observations being compared with  $I$  sample means, and  $DFE = N - I$ .

## EXAMPLE

### 12.10 Degrees of freedom for the Facebook friends study.

In the Facebook friends example, we have  $I = 5$  and  $N = 134$ . Therefore,

$$DFT = N - 1 = 134 - 1 = 133$$

$$DFG = I - 1 = 5 - 1 = 4$$

$$DFE = N - I = 134 - 5 = 129$$

These are the entries in the df column of Figure 12.8.

Note that the degrees of freedom add in the same way that the sums of squares add. That is,  $DFT = DFG + DFE$ .



**mean square, p. 587**

For each source of variation, the mean square is the sum of squares divided by the degrees of freedom. You can verify this by doing the divisions for the values given on the output in Figure 12.8. We compare these mean squares to test whether the population means are all the same.

## SUMS OF SQUARES, DEGREES OF FREEDOM, AND MEAN SQUARES

**Sums of squares** represent variation present in the data. They are calculated by summing squared deviations. In one-way ANOVA there are three **sources of variation**: groups, error, and total. The sums of squares are related by the formula

$$SST = SSG + SSE$$

Thus, the total variation is composed of two parts, one due to groups and one due to error.

**Degrees of freedom** are related to the deviations that are used in the sums of squares. The degrees of freedom are related in the same way as the sums of squares are:

$$DFT = DFG + DFE$$

To calculate each **mean square**, divide the corresponding sum of squares by its degrees of freedom.

We can use the mean square for error to find  $s_p$ , the pooled estimate of the parameter  $\sigma$  of our model. It is true in general that

$$s_p^2 = MSE = SSE/DFE$$

In other words, the mean square for error is an estimate of the within-group

variance,  $\sigma^2$ . The estimate of  $\sigma$  is therefore the square root of this quantity. So,

$$s_p = \sqrt{\text{MSE}}$$

## EXAMPLE

### 12.11 MSE for the Facebook friends study.

From the SPSS output in Figure 12.8 we see that the MSE is reported as 1.201. The pooled estimate of  $\sigma$  is therefore

$$\begin{aligned} s_p &= \sqrt{\text{MSE}} \\ &= \sqrt{1.201} = 1.10 \end{aligned}$$

This estimate is equal to our calculations of  $s_p$  in Example 12.5.

## The $F$ test

If  $H_0$  is true, there are no differences among the group means. The ratio  $\text{MSG}/\text{MSE}$  is a statistic that is approximately 1 if  $H_0$  is true and tends to be larger if  $H_a$  is true. This is the ANOVA  $F$  statistic. In our example,  $\text{MSG} = 4.973$  and  $\text{MSE} = 1.201$ , so the ANOVA  $F$  statistic is

$$F = \frac{\text{MSG}}{\text{MSE}} = \frac{4.973}{1.201} = 4.142$$

When  $H_0$  is true, the  $F$  statistic has an  $F$  distribution that depends upon two numbers: the *degrees of freedom for the numerator* and the *degrees of freedom for the denominator*. These degrees of freedom are those associated with the mean squares in the numerator and denominator of the  $F$  statistic. For one-way ANOVA, the degrees of freedom for the numerator are  $\text{DFG} = I - 1$ , and the degrees of freedom for the denominator are  $\text{DFE} = N - I$ . We use the notation  $F(I - 1, N - I)$  for this distribution.

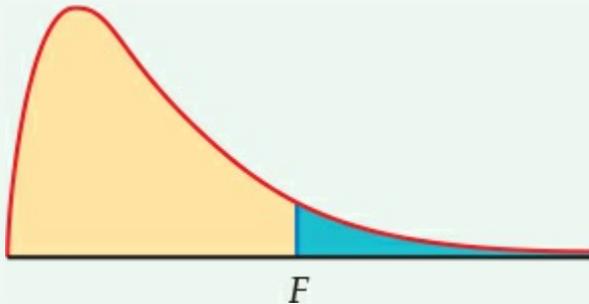


The *One-Way ANOVA* applet is an excellent way to see how the value of the  $F$  statistic and the  $P$ -value depend upon the variability of the data within the groups, the sample sizes, and the differences between the means. See Exercises 12.28 to 12.30 (page 682) for use of this applet.

## THE ANOVA $F$ TEST

To test the null hypothesis in a one-way ANOVA, calculate the  **$F$  statistic**

$$F = \text{MSG} / \text{MSE}$$



When  $H_0$  is true, the  $F$  statistic has the  $F(I - 1, N - I)$  distribution. When  $H_a$  is true, the  $F$  statistic tends to be large. We reject  $H_0$  in favor of  $H_a$  if the  $F$  statistic is sufficiently large.

The  **$P$ -value** of the  $F$  test is the probability that a random variable having the  $F(I - 1, N - I)$  distribution is greater than or equal to the calculated value of the  $F$  statistic.



Tables of  $F$  critical values are available for use when software does not give the  $P$ -value. Table E in the back of the book contains the  $F$  critical values for probabilities  $p = 0.100, 0.050, 0.025, 0.010, \text{ and } 0.001$ . For one-way ANOVA we use critical values from the table corresponding to  $I - 1$  degrees of freedom in the numerator and  $N - I$  degrees of freedom in the denominator. *When determining the  $P$ -value, remember that the  $F$  test is always one-sided because any differences among the group means tend to make  $F$  large.*

### EXAMPLE

#### 12.12 The ANOVA $F$ test for the Facebook friends study.

In the Facebook friends study, we found  $F = 4.14$ . (Note that it is standard

practice to round  $F$  statistics to two places after the decimal point.) There were five populations, so the degrees of freedom in the numerator are  $DFG = I - 1 = 4$ . For this example the degrees of freedom in the denominator are  $DFE = N - I = 134 - 5 = 129$ . Software provided a  $P$ -value of 0.003, so at the 0.05 significance level, we reject  $H_0$  and conclude that the population means are not all the same.

Suppose that  $P = 0.003$  was not provided. We'll now run through the process of using the table of  $F$  critical values to approximate the  $P$ -value. Although you will rarely need to do this in practice, the process will help you to understand the  $P$ -value calculation.

In Table E we first find the column corresponding to 4 degrees of freedom in the numerator. For the degrees of freedom in the denominator, we see that there are entries for 100 and 200. The values for these entries are very close. To be conservative we use critical values corresponding to 100 degrees of freedom in the denominator since these are slightly larger.

$p$	Critical value
0.100	2.00
0.050	2.46
0.025	2.92
0.010	3.51
0.001	5.02

We have  $F = 4.14$ . This is in between the critical value for  $P = 0.010$  and  $P = 0.001$ . Using the table, we can conclude only that  $0.001 < P < 0.010$ .

The following display shows the general form of a one-way ANOVA table with the  $F$  statistic. The formulas in the sum of squares column can be used for calculations in small problems. There are other formulas that are more efficient for hand or calculator use, but ANOVA calculations are usually done by computer software.

Source	Degrees of freedom	Sum of squares	Mean square	$F$
Groups	$I - 1$	$\sum_{\text{groups}} n_i(\bar{x}_i - \bar{x})^2$	$SSG/DFG$	$MSG/MSE$
Error	$N - I$	$\sum_{\text{groups}} (n_i - 1)s_i^2$	$SSE/DFE$	
Total	$N - 1$	$\sum_{\text{obs}} (x_{ij} - \bar{x})^2$		

One other item given by some software for ANOVA is worth noting. For an analysis of variance, we define the **coefficient of determination** as

coefficient of determination

$$R^2 = \frac{SSG}{SST}$$



## multiple correlation coefficient, p. 618

The coefficient of determination plays the same role as the squared multiple correlation  $R^2$  in a multiple regression. We can easily calculate the value from the ANOVA table entries.

### EXAMPLE

#### 12.13 Coefficient of determination for the Facebook friends study.

The software-generated ANOVA table for the Facebook friends study is given in Figure 12.8. From that display, we see that  $SSG = 19.890$  and  $SST = 174.757$ . The coefficient of determination is

$$R^2 = \frac{SSG}{SST} = \frac{19.890}{174.757} = 0.11$$

About 11% of the variation in social attractiveness scores is explained by the different profiles. The other 89% of the variation is due to participant-to-participant variation within each of the profile groups. We can see this in the histograms of Figure 12.3. Each of the groups has a large amount of variation, and there is a substantial amount of overlap in the distributions. *The fact that we have strong evidence ( $P < 0.003$ ) against the null hypothesis that the five population means are all the same does not tell us that the distributions of values are far apart.*



### USE YOUR KNOWLEDGE

#### 12.5 What's wrong?

For each of the following, explain what is wrong and why.

- (a) Within-group variation is the variation in the data due to the differences in the sample means.

- (b) The mean squares in an ANOVA table will add, that is,  $MST = MSG + MSE$ .
- (c) The pooled estimate  $s_p$  is a parameter of the ANOVA model.
- (d) A very small  $P$ -value implies that the group distributions of responses are far apart.

### **12.6 Determining the critical value of $F$ .**

For each of the following situations, state how large the  $F$  statistic needs to be for rejection of the null hypothesis at the 0.05 level.

- (a) Compare 4 groups with 4 observations per group.
- (b) Compare 5 groups with 4 observations per group.
- (c) Compare 5 groups with 5 observations per group.
- (d) Summarize what you have learned about  $F$  distributions from this exercise.

## 12.2 Comparing the Means

When you complete this section, you will be able to

- Distinguish between the use of contrasts to examine particular versions of the alternative hypothesis and the use of a multiple-comparisons method to compare pairs of means.
- Construct a level  $C$  confidence interval for a comparison of means expressed as a contrast.
- Perform a  $t$  significance test for a contrast and summarize the results.
- Summarize the trade-off of a multiple-comparisons method in terms of controlling false rejections and not detecting true differences in means.
- Describe what is done when one uses the Bonferroni method to control the probability of a false rejection.
- Interpret statistical software ANOVA output and draw conclusions regarding differences in population means.

### Contrasts

The ANOVA  $F$  test gives a general answer to a general question: are the differences among observed group means statistically significant? Unfortunately, a small  $P$ -value simply tells us that the group means are not all the same. It does not tell us specifically which means differ from each other. Plotting and inspecting the means give us some indication of where the differences lie, but we would like to supplement inspection with formal inference.

In the ideal situation, specific questions regarding comparisons among the means are posed before the data are collected. We can answer specific questions of this kind and attach a level of confidence to the answers we give. We now explore these ideas through the Facebook friends example.

### EXAMPLE

#### 12.14 Reporting the results.

In the Facebook friends study we compared the social attractiveness scores for five profiles, which varied only in the number of friends. Let's use  $\bar{x}_{102}, \bar{x}_{302}, \bar{x}_{502}, \bar{x}_{702}$ , and  $\bar{x}_{902}$  to represent the five sample means and a similar notation for the population means. From Figure 12.7 we see that the five sample means are

$$\bar{x}_{102}=3.82, \bar{x}_{302}=4.88, \bar{x}_{502}=4.56, \bar{x}_{702}=4.41, \text{ and } \bar{x}_{902}=3.99$$

The null hypothesis we tested was

$$H_0: \mu_{102} = \mu_{302} = \mu_{502} = \mu_{702} = \mu_{902}$$

versus the alternative that the five population means are not all the same. We would report these results as  $F(4, 129) = 4.14$  with  $P = 0.003$ . Note that we have given the degrees of freedom for the  $F$  statistic in parentheses. Because the  $P$ -value is very small, we conclude that the data provide clear evidence that the five population means are not all the same.

However, having evidence that the five population means are not the same does not tell us all we'd like to know. We would really like our analysis to provide us with more specific information. For example, the alternative hypothesis is true if

$$\mu_{102} < \mu_{302} = \mu_{502} = \mu_{702} = \mu_{902}$$

or if

$$\mu_{102} = \mu_{302} = \mu_{502} > \mu_{702} = \mu_{902}$$

or if

$$\mu_{102} \neq \mu_{302} \neq \mu_{502} \neq \mu_{702} \neq \mu_{902}$$



*When you reject the ANOVA null hypothesis, additional analyses are required to clarify the nature of the differences between the means.*

In terms of offline social networks, previous research has shown that the bigger one's social network, the higher one's social attractiveness. In fact, the relationship between the number of friends and social attractiveness appears linear. Therefore, a reasonable question to ask is whether or not this same sort of pattern exists within an online social network. We can take this question and translate it into a testable hypothesis.

## EXAMPLE

### 12.15 An additional comparison of interest.

The researchers hypothesize that, unlike an offline social network, the positive association between the number of friends and social attractiveness weakens as the number of friends increases. Specifically, the average increase in social attractiveness for an increase of 400 friends is different if starting at 102 friends versus starting at 502 friends. This results in the following null hypothesis:

$$H_{01}: \mu_{502} - \mu_{102} = \mu_{902} - \mu_{502}$$

We could use the two-sided alternative

$$H_{a1}: \mu_{502} - \mu_{102} \neq \mu_{902} - \mu_{502}$$

but we could also argue that the one-sided alternative

$$H_{a1}: \mu_{502} - \mu_{102} > \mu_{902} - \mu_{502}$$

is appropriate for this problem because we expect there to be a leveling off.

In the example above we used  $H_{01}$  and  $H_{a1}$  to designate the null and alternative hypotheses. The reason for this is that there is an additional set of hypotheses to assess if there is a general linear trend. We use  $H_{02}$  and  $H_{a2}$  for this set.

## EXAMPLE

### 12.16 Another comparison of interest.

This comparison tests if there is a general linear trend across the factor levels. Here are the null and alternative hypotheses:

$$H_{02}: -2\mu_{102} - \mu_{302} + \mu_{702} + 2\mu_{902} = 0$$

$$H_{a2}: -2\mu_{102} - \mu_{302} + \mu_{702} + 2\mu_{902} \neq 0$$

Each of  $H_{01}$  and  $H_{02}$  says that a combination of population means is 0. These combinations of means are called contrasts because the coefficients sum to zero.

We use  $\psi$ , the Greek letter psi, for contrasts among population means. For our first comparison, we have

$$\begin{aligned}\psi_1 &= -\mu_{102} + 2\mu_{502} + \mu_{902} \\ &= (-1)\mu_{102} + (2)\mu_{502} + (-1)\mu_{902}\end{aligned}$$

and for the second comparison

$$\psi_2 = (-2)\mu_{102} + (-1)\mu_{302} + (1)\mu_{702} + (2)\mu_{902}$$



In each case, the value of the contrast is 0 when  $H_0$  is true. *Note that we have chosen to define the contrasts so that they will be positive when the alternative of interest (what we expect) is true.* Whenever possible, this is a good idea because it makes some computations easier.

A contrast expresses an effect in the population as a combination of population means. To estimate the contrast, form the corresponding **sample contrast** by using sample means in place of population means. Under the ANOVA assumptions, a sample contrast is a linear combination of independent Normal variables and therefore has a Normal distribution. We can obtain the standard error of a contrast by using the rules for variances. Inference is based on  $t$  statistics. Here are the details.

sample contrast



rules for variances, p. 275

## CONTRASTS

A **contrast** is a combination of population means of the form

$$\psi = \sum a_i \mu_i$$

where the coefficients  $a_i$  sum to 0. The corresponding **sample contrast** is

$$c = \sum a_i \bar{x}_i$$

The **standard error of  $c$**  is

$$SE_c = s_p \sqrt{\sum a_i^2 / n_i}$$

To test the null hypothesis

$$H_0: \psi = 0$$

use the ***t* statistic**

$$t = c / SE_c$$

with degrees of freedom DFE that are associated with  $s_p$ . The alternative hypothesis can be one-sided or two-sided.

A level ***C*** confidence interval for  $\psi$  is

$$c \pm t^* SE_c$$

where  $t^*$  is the value for the  $t(\text{DFE})$  density curve with area  $C$  between  $-t^*$  and  $t^*$ .



rules for means, p. 272

Because each  $\bar{x}_i$  estimates the corresponding  $\mu_i$ , the addition rule for means tells us that the mean  $\mu_c$  of the sample contrast  $c$  is  $\psi$ . In other words,  $c$  is an unbiased estimator of  $\psi$ . Testing the hypothesis that a contrast is 0 assesses the significance of the effect measured by the contrast. It is often more informative to estimate the size of the effect using a confidence interval for the population contrast.

## EXAMPLE

### 12.17 The contrast coefficients.

In our example the coefficients in the contrasts are

$$a_1 = -1, a_2 = 0, a_3 = 2, a_4 = 0, a_5 = -1 \text{ for } \psi_1$$

and

$$a_1 = -2, a_2 = -1, a_3 = 0, a_4 = 1, a_5 = 2 \text{ for } \psi_2$$

where the subscripts 1, 2, 3, 4, and 5 correspond to the profiles with 102, 302, 502, 702, and 902 friends, respectively. In each case the sum of the  $a_i$  is 0. We

look at inference for each of these contrasts in turn.

## EXAMPLE

### 12.18 Testing the first contrast of interest.

The sample contrast that estimates  $\psi_1$  is

$$\begin{aligned}c_1 &= (-1)\bar{x}_{102} + (2)\bar{x}_{502} + (-1)\bar{x}_{902} \\ &= -3.82 + (2)4.56 - 3.99 = 1.31\end{aligned}$$

with standard error

$$\begin{aligned}SE_{c_1} &= 1.10\sqrt{(-1)^2 24 + (2)^2 26 + (-1)^2 21} \\ &= 0.54\end{aligned}$$

The  $t$  statistic for testing  $H_{01}: \psi_1 = 0$  versus  $H_{a1}: \psi_1 > 0$  is

$$\begin{aligned}t &= c_1 / SE_{c_1} \\ &= 1.31 / 0.54 = 2.43\end{aligned}$$

Because  $s_p$  has 129 degrees of freedom, software using the  $t(129)$  distribution gives the one-sided  $P$ -value as  $P = 0.008$ . If we used Table D, we would conclude that  $0.005 < P < 0.01$ . The  $P$ -value is small, so there is strong evidence against  $H_{01}$ .

We have evidence to conclude that the rate of change in the attractiveness score at the lower levels (estimated to be  $4.56 - 3.82 = 0.74$ ) is larger than the rate of increase at the upper levels (estimated to be  $3.99 - 4.56 = -0.57$ ). This suggests either a leveling off or a decrease in the attractiveness score as the number of friends increases. The size of the difference can be described with a confidence interval.

## EXAMPLE

### 12.19 Confidence interval for the first contrast.

To find the 95% confidence interval for  $\psi_1$ , we combine the estimate with its margin of error:

$$\begin{aligned}c_1 \pm t^* SE_{c_1} &= 1.31 \pm (1.984)(0.54) \\ &= 1.31 \pm 1.07\end{aligned}$$

The 1.984 is a conservative estimate of  $t^*$  using 100 degrees of freedom. The interval is (0.24, 2.38). We are 95% confident that the difference is between 0.24 and 2.38 points.

We use the same method for the second contrast.

## EXAMPLE

### 12.20 Testing the second contrast of interest.

The sample contrast that estimates  $\psi_2$  is

$$\begin{aligned}c_2 &= (-2)x^{-102} + (-1)x^{-302} + (1)x^{-702} + (2)x^{-902} \\ &= (-2)3.82 + (-1)4.88 + (1)4.41 + (2)3.99 \\ &= -7.64 - 4.88 + 4.41 + 7.98 \\ &= -0.13\end{aligned}$$

with standard error

$$SE_{c_2} = 1.10(-2)224 + (-1)233 + (1)230 + (2)221$$

The  $t$  statistic for assessing the significance of this contrast is

$$t = -0.13 / 0.71 = -0.18$$

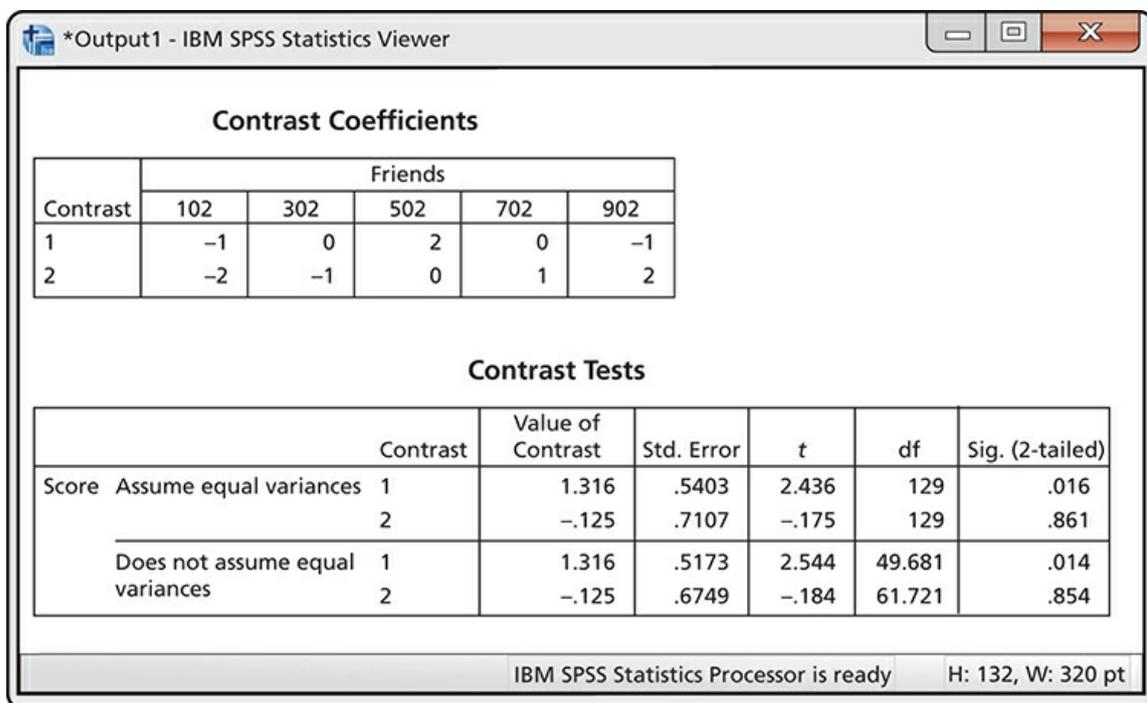
The  $P$ -value for the two-sided alternative is 0.861. The data do not provide much evidence in favor of a linear trend.

This second contrast can be combined with others to assess the various polynomial contributions (for example, linear, quadratic, cubic) to the relationship between attractiveness score and the number of friends. As we saw in Figure 12.5,

a quadratic trend appears most prominent. Further discussion of this contrast can be found in Exercise 12.50 (page 687).

SPSS output for the contrasts is given in Figure 12.9. The results agree with the calculations that we performed in Examples 12.18 and 12.20 except for minor differences due to roundoff error in our calculations. Note that the output does not give the confidence interval that we calculated in Example 12.19. This is easily computed, however, from the contrast estimate and standard error provided in the output.

Some statistical software packages report the test statistics associated with contrasts as  $F$  statistics rather than  $t$  statistics. These  $F$  statistics are the squares of the  $t$  statistics described previously. As with much statistical software output,  $P$ -values for significance tests are reported for the two-sided alternative.



**FIGURE 12.9**

SPSS output giving the contrast analysis for the Facebook friends example.



*If the software you are using gives  $P$ -values for the two-sided alternative, and you are using the appropriate one-sided alternative, divide the reported  $P$ -value by 2. In our example, we argued that a one-sided alternative was appropriate for the first contrast. The software reported the  $P$ -value as 0.016, so we can conclude  $P = 0.008$ . Dividing this value by 2 has no effect on the conclusion.*

Questions about population means are expressed as hypotheses about contrasts. A contrast should express a specific question that we have in mind when designing the study. Because the  $F$  test answers a very general question, it is less powerful

than tests for contrasts designed to answer specific questions.



*When contrasts are formulated before seeing the data, inference about contrasts is valid whether or not the ANOVA  $H_0$  of equality of means is rejected. Specifying the important questions before the analysis is undertaken enables us to use this powerful statistical technique.*

## Multiple comparisons

In many studies, specific questions cannot be formulated in advance of the analysis. If  $H_0$  is not rejected, we conclude that the population means are indistinguishable on the basis of the data given. On the other hand, if  $H_0$  is rejected, we would like to know which pairs of means differ. **Multiple-comparisons methods** address this issue. It is important to keep in mind that multiple-comparisons methods are used only *after rejecting* the ANOVA  $H_0$ .

### multiple-comparisons methods

## EXAMPLE

### 12.21 Comparing each pair of groups.

Return once more to the Facebook friends data with five groups. We can make 10 comparisons between pairs of means. We can write a  $t$  statistic for each of these pairs. For example, the statistic

$$\begin{aligned}t_{12} &= \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{38.2 - 4.881}{1.101 \sqrt{\frac{1}{24} + \frac{1}{33}}} \\ &= -3.59\end{aligned}$$

compares profiles with 102 and 302 friends. The subscripts on  $t$  specify which groups are compared.

The  $t$  statistics for two other pairs are

$$t_{23} = \frac{\bar{x}_2 - \bar{x}_3}{s_p \sqrt{\frac{1}{n_2} + \frac{1}{n_3}}}$$

$$= 4.88 - 4.561 \cdot 1.0133 + 126$$

$$= 1.11$$

and

$$t_{25} = \frac{\bar{x}_2 - \bar{x}_5}{s_p \sqrt{\frac{1}{n_2} + \frac{1}{n_5}}}$$

$$= 4.88 - 3.991 \cdot 1.0133 + 121$$

$$= 2.90$$



### two-sample $t$ procedures, p. 462

These  $t$  statistics are very similar to the pooled two-sample  $t$  statistic for comparing two population means. The difference is that we now have more than two populations, so each statistic uses the pooled estimator  $s_p$  from all groups rather than the pooled estimator from just the two groups being compared. This additional information about the common  $\sigma$  increases the power of the tests. The degrees of freedom for all these statistics are  $DFE = 129$ , those associated with  $s_p$ .

Because we do not have any specific ordering of the means in mind as an alternative to equality, we must use a two-sided approach to the problem of deciding which pairs of means are significantly different.

## MULTIPLE COMPARISONS

To perform a **multiple-comparisons procedure**, compute  **$t$  statistics** for all pairs of means using the formula

$$t_{ij} = \frac{\bar{x}_i - \bar{x}_j}{s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}$$

If

$$|t_{ij}| \geq t^{**}$$

we declare that the population means  $\mu_i$  and  $\mu_j$  are different. Otherwise, we conclude that the data do not distinguish between them. The value of  $t^{**}$  depends upon which multiple-comparisons procedure we choose.

One obvious choice for  $t^{**}$  is the upper  $\alpha/2$  critical value for the  $t(DFE)$  distribution. This choice simply carries out as many separate significance tests of fixed level  $\alpha$  as there are pairs of means to be compared. The procedure based on this choice is called the **least-significant differences method**, or simply LSD.

## LSD method



LSD has some undesirable properties, particularly if the number of means being compared is large. Suppose, for example, that there are  $I = 20$  groups and we use LSD with  $\alpha = 0.05$ . There are 190 different pairs of means. If we perform 190  $t$  tests, each with an error rate of 5%, our overall error rate will be unacceptably large. We expect about 5% of the 190 to be significant even if the corresponding population means are the same. Since 5% of 190 is 9.5, we expect 9 or 10 false rejections.

The LSD procedure fixes the probability of a false rejection for each single pair of means being compared. It does not control the overall probability of *some* false rejection among all pairs. Other choices of  $t^{**}$  control possible errors in other ways. The choice of  $t^{**}$  is therefore a complex problem, and a detailed discussion of it is beyond the scope of this text. Many choices for  $t^{**}$  are used in practice. One major statistical package allows selection from a list of over a dozen choices.

We will discuss only one of these, called the **Bonferroni method**. Use of this procedure with  $\alpha = 0.05$ , for example, guarantees that the probability of *any* false rejection among all comparisons made is no greater than 0.05. This is much stronger protection than controlling the probability of a false rejection at 0.05 for *each separate* comparison.

## Bonferroni method

### EXAMPLE

#### 12.22 Applying the Bonferroni method.

We apply the Bonferroni multiple-comparisons procedure with  $\alpha = 0.05$  to the data from the Facebook friends study. The value of  $t^{**}$  for this procedure uses  $\alpha = 0.05/10 = 0.005$  for each test. From Table D, this value is 2.63. Of the statistics  $t_{12} = -3.59$ ,  $t_{23} = 1.11$ , and  $t_{25} = 2.90$  calculated in Example 12.21, only  $t_{12}$  and  $t_{25}$  are significant. These two statistics compare the profile of 302 friends with the two extreme levels.

Of course, we prefer to use software for the calculations.

## EXAMPLE

### 12.23 Interpreting software output.

The output generated by SPSS for Bonferroni comparisons appears in Figure 12.10. The software uses an asterisk to indicate that the difference in a pair of means is statistically significant. Here, all 10 comparisons are reported. These results agree with the calculations that we performed in Examples 12.21 and 12.22. There are no significant differences except those already mentioned. Note that each comparison is given twice in the output.

**Multiple Comparisons**

Dependent Variable: Score  
Bonferroni

(I) Friends	(J) Friends	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
102	302	-1.0621*	.2939	.004	-1.902	-.223
	502	-.7449	.3102	.177	-1.631	.141
	702	-.5900	.3001	.514	-1.447	.267
	902	-.1738	.3274	1.000	-1.109	.761
302	102	1.0621*	.2939	.004	.223	1.902
	502	.3172	.2873	1.000	-.503	1.138
	702	.4721	.2764	.900	-.317	1.262
	902	.8883*	.3059	.043	.015	1.762
502	102	.7449	.3102	.177	-.141	1.631
	302	-.3172	.2873	1.000	-1.138	.503
	702	.1549	.2936	1.000	-.684	.993
	902	.5711	.3215	.780	-.347	1.489
702	102	.5900	.3001	.514	-.267	1.447
	302	-.4721	.2764	.900	-1.262	.317
	502	-.1549	.2936	1.000	-.993	.684
	902	.4162	.3117	1.000	-.474	1.307
902	102	.1738	.3274	1.000	-.761	1.109
	302	-.8883*	.3059	.043	-1.762	-.015
	502	-.5711	.3215	.780	-1.489	.347
	702	-.4162	.3117	1.000	-1.307	.474

\* The mean difference is significant at the 0.05 level.

IBM SPSS Statistics Processor is ready H: 132, W: 320 pt

**FIGURE 12.10**

SPSS output giving the multiple-comparisons analysis for the Facebook friends example.

The data in the Facebook friends study provide a clear result: the social attractiveness score increases as the number of friends increases to a point and then

decreases. Unfortunately with these data, we cannot accurately describe this relationship in more detail. This lack of clarity is not unusual when performing a multiple-comparisons analysis.

Here, the mean associated with 302 friends is significantly different from the means for the 102- and 902-friend profiles, but it is not found significantly different from the means for the profiles with 502 and 702 friends. To complicate things, the means for profiles with 502 and 702 friends were not found significantly different from the means for the 102- and 902-friend profiles.



*This kind of apparent contradiction points out dramatically the nature of the conclusions of statistical tests of significance.* The conclusion appears to be illogical. If  $\mu_1$  is the same as  $\mu_3$  and if  $\mu_3$  is the same as  $\mu_2$ , doesn't it follow that  $\mu_1$  is the same as  $\mu_2$ ? Logically, the answer must be Yes.

Some of the difficulty can be resolved by noting the choice of words used. In describing the inferences, we talk about failing to detect a difference or concluding that two groups are different. In making logical statements, we say things such as "is the same as." There is a big difference between the two modes of thought. Statistical tests ask, "Do we have adequate evidence to distinguish two means?" It is not illogical to conclude that we have sufficient evidence to distinguish  $\mu_1$  from  $\mu_2$ , but not  $\mu_1$  from  $\mu_3$  or  $\mu_2$  from  $\mu_3$ .

One way to deal with these difficulties of interpretation is to give confidence intervals for the differences. The intervals remind us that the differences are not known exactly. We want to give *simultaneous confidence intervals*, that is, intervals for *all* differences among the population means at once. Again, we must face the problem that there are many competing procedures—in this case, many methods of obtaining simultaneous intervals.

## **SIMULTANEOUS CONFIDENCE INTERVALS FOR DIFFERENCES BETWEEN MEANS**

**Simultaneous confidence intervals** for all differences  $\mu_i - \mu_j$  between population means have the form

$$(\bar{x}_i - \bar{x}_j) \pm t^{**} s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

The critical values  $t^{**}$  are the same as those used for the multiple-comparisons procedure chosen.

The confidence intervals generated by a particular choice of  $t^{**}$  are closely related to the multiple-comparisons results for that same method. If one of the confidence intervals includes the value 0, then that pair of means will not be declared significantly different, and vice versa.

## EXAMPLE

### 12.24 Interpreting software output, continued.

The SPSS output for the Bonferroni multiple-comparisons procedure given in Figure 12.10 includes the simultaneous 95% confidence intervals. We can see, for example, that the interval for  $\mu_1 - \mu_3$  is  $-1.63$  to  $0.14$ . The fact that the interval includes 0 is consistent with the fact that we failed to detect a difference between these two means using this procedure. Note that the interval for  $\mu_3 - \mu_1$  is also provided. This is not really a new piece of information, because it can be obtained from the other interval by reversing the signs and reversing the order, that is,  $-0.14$  to  $1.63$ . So, in fact, we really have only 10 intervals. Use of the Bonferroni procedure provides us with 95% confidence that *all 10* intervals simultaneously contain the true values of the population mean differences.

## Software

We have used SPSS to illustrate the analysis of the Facebook friends data. Other statistical software gives similar output, and you should be able to read it without any difficulty. Here's an example with output from three software packages.

## EXAMPLE

### 12.25 Do eyes affect ad response?

Research from a variety of fields has found significant effects of eye gaze and eye color on emotions and perceptions such as arousal, attractiveness, and honesty. These findings suggest that a model's eyes may play a role in a

viewer's response to an ad.



In one study, students in marketing and management classes of a southern, predominantly Hispanic, university were each presented with one of four portfolios.<sup>4</sup> Each portfolio contained a target ad for a fictional product, Sparkle Toothpaste. Students were asked to view the ad and then respond to questions concerning their attitudes and emotions about the ad and product. All questions were from advertising-effects questionnaires previously used in the literature. Each response was on a seven-point scale.

Although the researchers investigated nine attitudes and emotions, we will focus on the viewer's "attitudes toward the brand." This response was obtained by averaging 10 survey questions.

The target ads were created using two digital photographs of a model. In one picture the model is looking directly at the camera so the eyes can be seen. This picture was used in three target ads. The only difference was the model's eyes, which were made to be either brown, blue, or green. In the second picture, the model is in virtually the same pose but looking downward so the eyes are not visible. A total of 222 surveys were used for analysis. The following table summarizes the responses for the four portfolios. Outputs from Excel, SAS, and Minitab are given in Figure 12.11.

Group	<i>n</i>	Mean	Std. dev.
Blue	67	3.19	1.75
Brown	37	3.72	1.72
Down	41	3.11	1.53
Green	77	3.86	1.67

There is evidence at the 5% significance level to reject the null hypothesis that the four groups have equal means ( $P = 0.036$ ). In Exercises 12.41 and 12.42 (page 685), you are asked to perform further inference using contrasts.

Excel

	A	B	C	D	E	F	G
1	Anova: Single Factor						
2							
3	SUMMARY						
4	<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>		
5	Blue	67	214	3.19403	3.079055		
6	Brown	37	137.8	3.724324	2.942447		
7	Down	41	127.4	3.107317	2.326695		
8	Green	77	297.2	3.85974	2.775332		
9							
10	ANOVA						
11	<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-Value</i>	<i>F crit</i>
12	Between Groups	24.41966	3	8.139886	2.894117	0.036184	2.646014
13	Within Groups	613.1387	218	2.812563			
14							
15	Total	637.5584	221				

SAS

**The GLM Procedure**

Dependent Variable: Score

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	24.4196586	8.1398862	2.89	0.0362
Error	218	613.1387197	2.8125629		
Corrected Total	221	637.5583784			

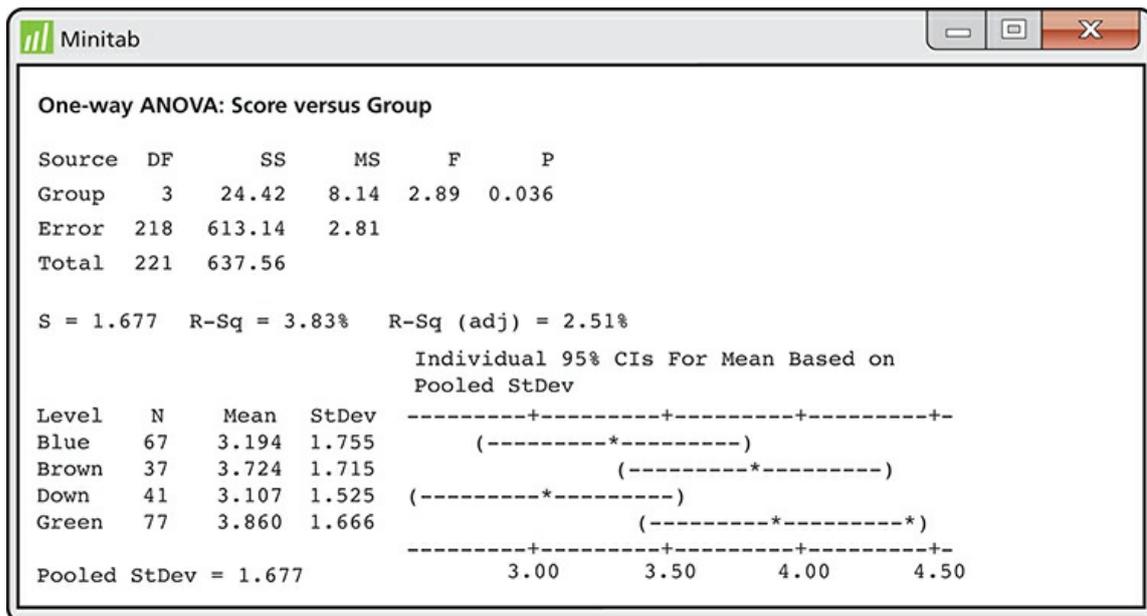
  

R-Square	Coeff Var	Root MSE	Score Mean
0.038302	47.95331	1.677070	3.497297

Level of Group	N	Score	
		Mean	Std Dev
Blue	67	3.19402985	1.75472355
Brown	37	3.72432432	1.71535636
Down	41	3.10731707	1.52535082
Green	77	3.85974026	1.66593262

Done



**FIGURE 12.11**

Excel, SAS, and Minitab output for the advertising study in Example 12.25.

## USE YOUR KNOWLEDGE

### 12.7 Why no multiple comparisons?

Any pooled two-sample  $t$  problem can be run as a one-way ANOVA with  $I = 2$ . Explain why it is inappropriate to analyze the data using contrasts or multiple-comparisons procedures in this setting.

### 12.8 Growth of Douglas fir seedlings.

An experiment was conducted to compare the growth of Douglas fir seedlings under three different levels of vegetation control (0%, 50%, and 100%). Twenty seedlings were randomized to each level of control. The resulting sample means for stem volume were 53, 76, and 110 cubic centimeters ( $\text{cm}^3$ ), respectively, with  $s_p = 28 \text{ cm}^3$ . The researcher hypothesized that the average growth at 50% control would be less than the average of the 0% and 100% levels.

- What are the coefficients for testing this contrast?
- Perform the test and report the test statistic, degrees of freedom, and  $P$ -value. Do the data provide evidence to support this hypothesis?

## Power

Recall that the power of a test is the probability of rejecting  $H_0$  when  $H_a$  is in fact true. Power measures how likely a test is to detect a specific alternative. When planning a study in which ANOVA will be used for the analysis, it is important to perform power calculations to check that the sample sizes are adequate to detect differences among means that are judged to be important.

Power calculations also help evaluate and interpret the results of studies in which  $H_0$  was not rejected. We sometimes find that the power of the test was so low against reasonable alternatives that there was little chance of obtaining a significant  $F$ .



### power, p. 477

In Chapter 7 we found the power for the two-sample  $t$  test. One-way ANOVA is a generalization of the two-sample  $t$  test, so it is not surprising that the procedure for calculating power is quite similar. Here are the steps that are needed:

1. Specify
  - (a) an alternative ( $H_a$ ) that you consider important; that is, values for the true population means  $\mu_1, \mu_2, \dots, \mu_I$ ;
  - (b) sample sizes  $n_1, n_2, \dots, n_I$ ; usually these will all be equal to the common value  $n$ ;
  - (c) a level of significance  $\alpha$ , usually equal to 0.05; and
  - (d) a guess at the standard deviation  $\sigma$ .
2. Use the degrees of freedom  $DFG = I - 1$  and  $DFE = N - I$  to find the critical value that will lead to the rejection of  $H_0$ . This value, which we denote by  $F^*$ , is the upper  $\alpha$  critical value for the  $F(DFG, DFE)$  distribution.

3. Calculate the **noncentrality parameter**<sup>5</sup>

noncentrality parameter

$$\lambda = \sum n_i (\mu_i - \mu^-)^2 / \sigma^2$$

where  $\mu^-$  is a weighted average of the group means

$$\mu^- = \frac{\sum n_i \mu_i}{N}$$

4. Find the power, which is the probability of rejecting  $H_0$  when the alternative hypothesis is true, that is, the probability that the observed  $F$  is greater than  $F^*$ .

Under  $H_a$ , the  $F$  statistic has a distribution known as the **noncentral  $F$  distribution**. SAS, for example, has a function for this distribution. Using this function, the power is

noncentral  $F$  distribution

$$\text{Power} = 1 - \text{PROBF}(F^*, \text{DFG}, \text{DFE}, \lambda)$$

Note that, if the  $n_i$  are all equal to the common value  $n$ , then  $\mu^-$  is the ordinary average of the  $\mu_i$  and

$$\lambda = n \sum (\mu_i - \mu^-)^2 \sigma^2$$

If the means are all equal (the ANOVA  $H_0$ ), then  $\lambda = 0$ . The noncentrality parameter measures how unequal the given set of means is. Large  $\lambda$  points to an alternative far from  $H_0$ , and we expect the ANOVA  $F$  test to have high power. Software makes calculation of the power quite easy, but tables and charts are also available.

## EXAMPLE

### 12.26 Power of a reading comprehension study.

Suppose that a study on reading comprehension for three different teaching methods has 10 students in each group. How likely is this study to detect differences in the mean responses that would be viewed as important? A previous study performed in a different setting found sample means of 41, 47, and 44, and the pooled standard deviation was 7. Based on these results, we will use  $\mu_1 = 41$ ,  $\mu_2 = 47$ ,  $\mu_3 = 44$ , and  $\sigma = 7$  in a calculation of power. The  $n_i$  are equal, so  $\mu^-$  is simply the average of the  $\mu_i$ :

$$\mu^- = \frac{41 + 47 + 44}{3} = 44$$

The noncentrality parameter is therefore

$$\begin{aligned} \lambda &= n \sum (\mu_i - \mu^-)^2 \sigma^2 \\ &= (10)[(41 - 44)^2 + (47 - 44)^2 + (44 - 44)^2] 49 \\ &= (10)(18)49 = 3.67 \end{aligned}$$

Because there are three groups with 10 observations per group,  $\text{DFG} = 2$  and

DFE = 27. The critical value for  $\alpha = 0.05$  is  $F^* = 3.35$ . The power is therefore

$$1 - \text{PROBF}(3.35, 2, 27, 3.67) = 0.3486$$

The chance that we reject the ANOVA  $H_0$  at the 5% significance level is only about 35%.

If the assumed values of the  $\mu_i$  in this example describe differences among the groups that the experimenter wants to detect, then we would want to use more than 10 subjects per group.

## EXAMPLE

### 12.27 Changing the sample size.

To decide on an appropriate sample size for the experiment described in the previous example, we repeat the power calculation for different values of  $n$ , the number of subjects in each group. Here are the results:

$n$	DFG	DFE	$F^*$	$\lambda$	Power
20	2	57	3.16	7.35	0.65
30	2	87	3.10	11.02	0.84
40	2	117	3.07	14.69	0.93
50	2	147	3.06	18.37	0.97
100	2	297	3.03	36.73	$\approx 1$

With  $n = 30$ , the experimenters have an 84% chance of rejecting  $H_0$  with  $\alpha = 0.05$  and thereby demonstrating that the groups have different means. That is, in the long run, 84 out of every 100 such experiments would reject  $H_0$  at the  $\alpha = 0.05$  level of significance. Power of at least 80% is often considered adequate.

Using 50 subjects per group increases the chance of finding significance to 97%. With 100 subjects per group, the experimenters are virtually certain to reject  $H_0$ . The exact power for  $n = 100$  is 0.99989. In most real-life situations the additional cost of increasing the sample size from 50 to 100 subjects per group would not be justified by the relatively small increase in the chance of obtaining statistically significant results.

## CHAPTER 12 Summary

**One-way analysis of variance (ANOVA)** is used to compare several population

means based on independent SRSs from each population. The populations are assumed to be Normal with possibly different means and the same standard deviation.

To do an analysis of variance, first compute sample means and standard deviations for all groups. Side-by-side boxplots give an overview of the data. Examine Normal quantile plots (either for each group separately or for the residuals) to detect outliers or extreme deviations from Normality. Compute the ratio of the largest to the smallest sample standard deviation. If this ratio is less than 2 and the Normal quantile plots are satisfactory, ANOVA can be performed.

The **null hypothesis** is that the population means are *all* equal. The **alternative hypothesis** is true if there are *any* differences among the population means.

ANOVA is based on separating the total variation observed in the data into two parts: variation **among group means** and variation **within groups**. If the variation among groups is large relative to the variation within groups, we have evidence against the null hypothesis.

An **analysis of variance table** organizes the ANOVA calculations. **Degrees of freedom, sums of squares, and mean squares** appear in the table. The ***F* statistic** and its ***P*-value** are used to test the null hypothesis.

The ANOVA ***F* test** shares the **robustness** of the two-sample ***t* test**. It is relatively insensitive to moderate non-Normality and unequal variances, especially when the sample sizes are similar.

Specific questions formulated before examination of the data can be expressed as **contrasts**. Tests and confidence intervals for contrasts provide answers to these questions.

If no specific questions are formulated before examination of the data and the null hypothesis of equality of population means is rejected, **multiple-comparisons** methods are used to assess the statistical significance of the differences between pairs of means.

The **power** of the ***F* test** depends upon the sample sizes, the variation among population means, and the within-group standard deviation.

## CHAPTER 12 Exercises

For Exercises 12.1 and 12.2, see page 651; for Exercises 12.3 and 12.4, see page 655; for Exercises 12.5 and 12.6, see page 662; and for Exercises 12.7 and 12.8, see page 675.

### 12.9 A one-way ANOVA example.

A study compared 4 groups with 6 observations per group. An  $F$  statistic of 3.18 was reported.

- Give the degrees of freedom for this statistic and the entries from Table E that correspond to this distribution.
- Sketch a picture of this  $F$  distribution with the information from the table included.
- Based on the table information, how would you report the  $P$ -value?
- Can you conclude that all pairs of means are different? Explain your answer.

### 12.10 Calculating the ANOVA $F$ test $P$ -value.

For each of the following situations, find the degrees of freedom for the  $F$  statistic and then use Table E to approximate the  $P$ -value.

- Seven groups are being compared with 6 observations per group. The value of the  $F$  statistic is 2.05.
- Five groups are being compared with 11 observations per group. The value of the  $F$  statistic is 2.85.
- Six groups are being compared using 31 total observations. The value of the  $F$  statistic is 4.02.

### 12.11 Calculating the ANOVA $F$ test $P$ -value, continued.

For each of the following situations, find the  $F$  statistic and the degrees of freedom. Then draw a sketch of the distribution under the null hypothesis and shade in the portion corresponding to the  $P$ -value. State how you would report the  $P$ -value.

- Compare 4 groups with 16 observations per group,  $MSE = 50$ , and  $MSG = 127$ .
- Compare 3 groups with 9 observations per group,  $SSG = 58$ , and  $SSE = 172$ .

### 12.12 Calculating the pooled standard deviation.

An experiment was run to compare three groups. The sample sizes were 27, 31, and 122, and the corresponding estimated standard deviations were 37, 28, and 46.

- Is it reasonable to use the assumption of equal standard deviations when we analyze these data? Give a reason for your answer.

- (b) Give the values of the variances for the three groups.
- (c) Find the pooled variance.
- (d) What is the value of the pooled standard deviation?
- (e) Explain why your answer in part (d) is much closer to the standard deviation for the third group than to either of the other two standard deviations.

### 12.13 Describing the ANOVA model.

For each of the following situations, identify the response variable and the populations to be compared, and give  $I$ , the  $n_i$ , and  $N$ .

- (a) A poultry farmer is interested in reducing the cholesterol level in his marketable eggs. He wants to compare two different cholesterol-lowering drugs added to the hens' standard diet as well as an all-vegetarian diet. He assigns 25 of his hens to each of the three treatments.
- (b) A researcher is interested in students' opinions regarding an additional annual fee to support non-income-producing varsity sports. Students were asked to rate their acceptance of this fee on a seven-point scale. She received 94 responses, of which 31 were from students who attend varsity football or basketball games only, 18 were from students who also attend other varsity competitions, and 45 were from students who did not attend any varsity games.
- (c) A professor wants to evaluate the effectiveness of his teaching assistants. In one class period, the 42 students were randomly divided into three equal-sized groups, and each group was taught power calculations from one of the assistants. At the beginning of the next class, each student took a quiz on power calculations, and these scores were compared.

### 12.14 Describing the ANOVA model, continued.

For each of the following situations, identify the response variable and the populations to be compared, and give  $I$ , the  $n_i$ , and  $N$ .

- (a) A developer of a virtual-reality (VR) teaching tool for the deaf wants to compare the effectiveness of different navigation methods. A total of 40 children were available for the experiment, of which equal numbers were randomly assigned to use a joystick, wand, dancemat, or gesture-based pinch gloves. The time (in seconds) to complete a designed VR path is recorded for each child.
- (b) To study the effects of pesticides on birds, an experimenter randomly (and equally) allocated 65 chicks to five diets (a control and four with a different pesticide included). After a month, the calcium content (milligrams) in a 1-centimeter length of bone from each chick was measured.
- (c) A university sandwich shop wants to compare the effects of providing free food with a sandwich order on sales. The experiment will be conducted from 11:00 A.M. to 2:00 P.M. for the next 20 weekdays. On each day, customers will be offered one of the following: a free drink, free chips, a free cookie, or nothing. Each option will be offered five times.

### 12.15 Determining the degrees of freedom.

Refer to Exercise 12.13. For each situation, give the following:

- (a) Degrees of freedom for group, for error, and for the total

- (b) Null and alternative hypotheses
- (c) Numerator and denominator degrees of freedom for the  $F$  statistic

### **12.16 Determining the degrees of freedom, continued.**

Refer to Exercise 12.14. For each situation, give the following:

- (a) Degrees of freedom for group, for error, and for the total
- (b) Null and alternative hypotheses
- (c) Numerator and denominator degrees of freedom for the  $F$  statistic

### **12.17 Data collection and the interpretation of results.**

Refer to Exercise 12.13. For each situation, discuss the method of obtaining the data and how this will affect the extent to which the results can be generalized.

### **12.18 Data collection, continued.**

Refer to Exercise 12.14. For each situation, discuss the method of obtaining the data and how this will affect the extent to which the results can be generalized.

### **12.19 Pain tolerance among sports teams.**

Many have argued that sports such as football require the ability to withstand pain from injury for extended periods of time. To see if there is greater pain tolerance among certain sports teams, a group of researchers assessed 183 male Division II athletes from 5 sports.<sup>6</sup> Each athlete was asked to put his dominant hand and forearm in a 3°C water bath and keep it in there until the pain became intolerable. The total amount of time (in seconds) that each athlete maintained his hand and forearm in the bath was recorded. Following this procedure, each athlete completed a series of surveys on aggression and competitiveness. In their report, the researchers state:

*A univariate between subjects (sports team) ANOVA was performed on the total amount of time athletes were able to keep their hand and forearm in the water bath, and found it to be statistically significant,  $F(4, 146) = 4.96, p < .001$ . The lacrosse and soccer players tolerated the pain for a longer period of time than athletes from the other teams. Swimmers tolerated the pain for a significantly shorter period of time than the other teams.*

- (a) Based on the description of the experiment, what should the degrees of freedom be for this analysis?
- (b) Assuming that the degrees of freedom reported are correct, data from how many athletes were used in this analysis?
- (c) The researchers do not comment on the missing data in their report. List two reasons why these data may not have been used, and for each, explain how the omission could impact or bias the results.

### **12.20 Multitasking with technology in the classroom.**

Laptops and other digital technologies with wireless access to the Internet are becoming more and more common in the classroom. While numerous studies have shown that these technologies can be used effectively as part of teaching, there is concern that these technologies can also distract learners if used for off-task behaviors.

In one study that looked at the effects of off-task multitasking with digital technologies in the classroom, a total of 145 undergraduates were randomly assigned to one of seven conditions.<sup>7</sup> Each condition involved a task that was conducted simultaneously with a class lecture. The study consisted of three 20-minute lectures, each followed by a 15-item quiz. The following table summarizes the conditions and quiz results.

Condition	<i>n</i>	Lecture 1	Lecture 2	Lecture 3
Texting	21	0.57	0.75	0.56
Email	20	0.52	0.69	0.50
Facebook	20	0.50	0.68	0.43
MSN Messaging	21	0.48	0.71	0.42
Natural use control	21	0.50	0.78	0.58
Word-processing control	21	0.55	0.75	0.57
Paper-and-pencil control	21	0.60	0.74	0.53

- For this analysis, let's consider the average of the three quizzes as the response. Compute this mean for each condition.
- The analysis of these average scores results in  $SSG = 0.22178$  and  $SSE = 2.00238$ . Test the null hypothesis that the mean scores across all conditions are equal.
- Using the marginal means from part (a) and the Bonferroni multiple-comparisons method, determine which pairs of means differ significantly at the 0.05 significance level. (*Hint*: There are 21 pairwise comparisons, so the critical  $t$ -value is 3.095.)
- Summarize your results from parts (b) and (c) in a short report.

### 12.21 Contrasts for multitasking.

Refer to the previous exercise. Let  $\mu_1, \mu_2, \dots, \mu_7$  represent the mean scores for the 7 conditions. The first 4 conditions refer to off-task behaviors, while the last 3 conditions represent different sorts of controls.

- The researchers hypothesized that the average score for the off-task behaviors would be lower than that for the paper-and-pencil control condition. Write a contrast that expresses this comparison.
- For this contrast, give  $H_0$  and an appropriate  $H_a$ .
- Calculate the test statistic and approximate  $P$ -value for the significance test. What do you conclude?

### 12.22 Residual analysis.

In this chapter, we considered comparing sample standard deviations to assess the model assumption of constant variance and examining histograms of the group responses to assess the assumption of Normality. As we did in both simple linear regression (Chapter 10) and multiple linear regression (Chapter 11), we can also assess these assumptions by examining the residuals. Let's do that here for the Facebook friends study.  FRIENDS

- (a) Fit the model and obtain the residuals. Generate a scatterplot of the residuals versus the group variable. Does it appear that the residuals are symmetrically scattered above and below 0? Are there any outliers?
- (b) Is the spread of the residuals in each group relatively equal? This is a visual way to assess constant variance.
- (c) Generate a histogram or Normal quantile plot of the residuals. Does it appear that these residuals are reasonably Normal?

### 12.23 Organic foods and morals?

Organic foods are often marketed using moral terms such as “honesty” and “purity.” Is this just a marketing strategy or is there a conceptual link between organic food and morality? In one experiment, 62 undergraduates were randomly assigned to one of three food conditions (organic, comfort, and control).<sup>8</sup> First, each participant was given a packet of four food types from the assigned condition and told to rate the desirability of each food on a 7-point scale. Then, each was presented with a list of six moral transgressions and asked to rate each on a 7-point scale ranging from 1 = not at all morally wrong to 7 = very morally wrong. The average of these six scores was used as the response. 

- (a) Make a table giving the sample size, mean, and standard deviation for each group. Is it reasonable to pool the variances?
- (b) Generate a histogram for each of the groups. Can we feel confident that the sample means are approximately Normal? Explain your answer.

### 12.24 Organic foods and morals, continued.

Refer to the previous exercise. 

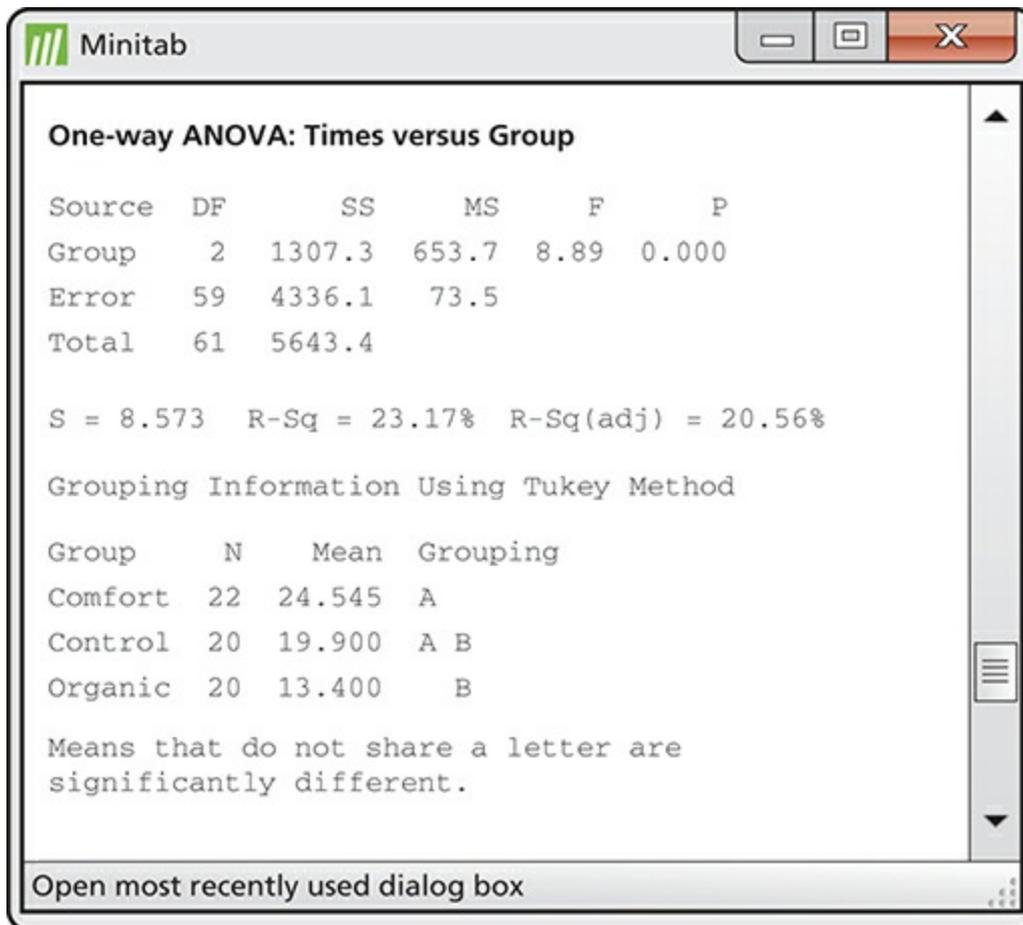
- (a) Analyze the scores using analysis of variance. Report the test statistic, degrees of freedom, and *P*-value.
- (b) Assess the assumptions necessary for inference by examining the residuals. Summarize your findings.
- (c) Compare the groups using the least-significant differences method.
- (d) A higher score is associated with a harsher moral judgment. Using the results from parts (a) and (b), write a short summary of your conclusions.

### 12.25 Organic foods and friendly behavior?

Refer to Exercise 12.23 for the design of the experiment. After rating the moral transgressions, the participants were told “that another professor from another department is also conducting research and really needs volunteers.” They were told that they would not receive compensation or course credit for their help and then were asked to write down the number of minutes (out of 30) that they would be willing to volunteer. This sort of question is often used to measure a person’s prosocial behavior.

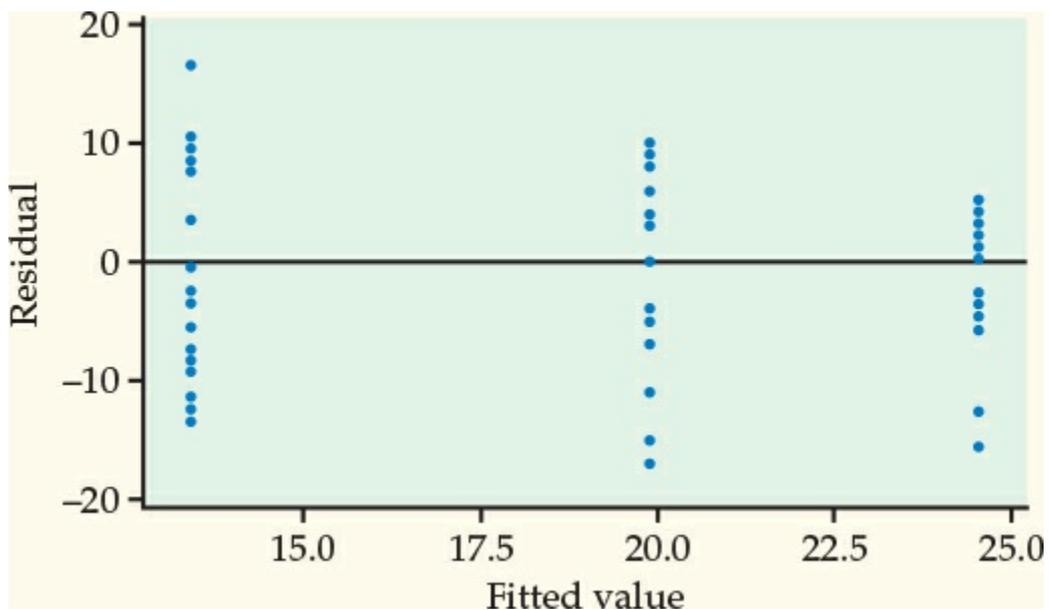
- (a) Figure 12.12 contains the Minitab output for the analysis of this response variable. Write a one-paragraph summary of your conclusions.

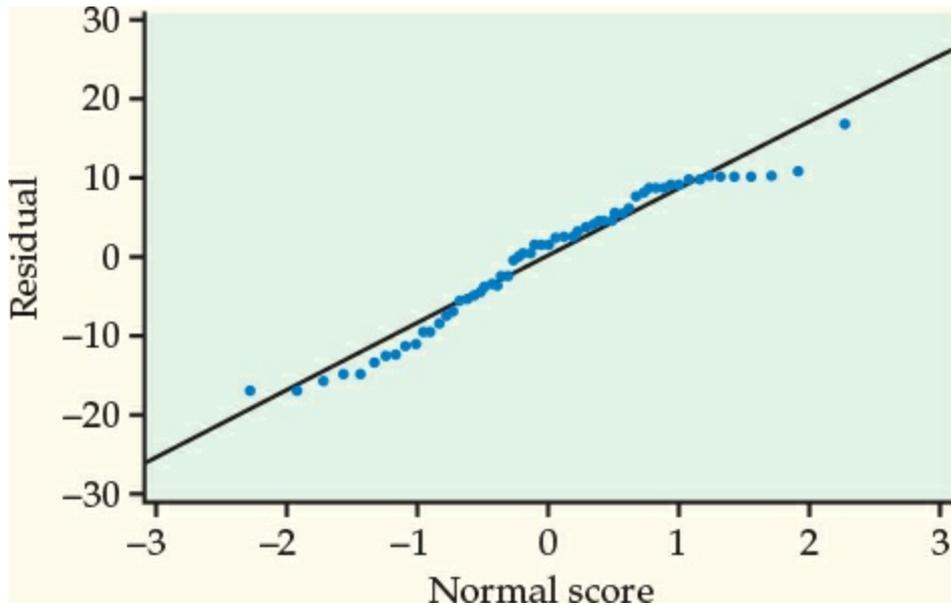
(b) Figure 12.13 contains a residual plot and a Normal quantile plot of the residuals. Are there any concerns regarding the assumptions necessary for inference? Explain your answer.



**FIGURE 12.12**

Minitab output for Exercise 12.25.





**FIGURE 12.13**  
Residual plot and Normal quantile plot for Exercise 12.25.

### 12.26 Massage therapy for osteoarthritis of the knee.

Various studies have shown the benefits of massage to manage pain. In one study, 125 adults suffering from osteoarthritis of the knees were randomly assigned to one of five 8-week regimens.<sup>9</sup> The primary outcome was the change in the Western Ontario and McMaster Universities Arthritis Index (WOMAC-Global). This index is used extensively to assess pain and functioning in those suffering from arthritis. Negative values indicate improvement. The following table summarizes the results of those completing the study.

Regimen	$n$	$\bar{x}$	$s$
30 min massage 1 $\times$ /wk	22	-17.4	17.9
30 min massage 2 $\times$ /wk	24	-18.4	20.7
60 min massage 1 $\times$ /wk	24	-24.0	18.4
60 min massage 2 $\times$ /wk	25	-24.0	19.8
Usual care, no massage	24	-6.3	14.6

- What proportion of adults dropped out of the study before completion?
- Is it reasonable to use the assumption of equal standard deviations when we analyze these data? Give a reason for your answer.
- Find the pooled standard deviation.
- The  $SS(\text{Regimen}) = 5060.346$ . Test the null hypothesis that the mean change in WOMAC-Global score is the same for all regimens.
- There are 10 pairs of means to compare. For the Bonferroni multiple-comparisons method, the critical  $t$ -value is 2.863. Which pairs of means are found to be significantly different? Write a short summary of your analysis.

### 12.27 Shopping and bargaining in Mexico.

Price haggling and other bargaining behaviors among consumers have been observed for a long time. However, research addressing these behaviors, especially in a real-life setting, remains relatively sparse. A group of researchers performed a small study to determine whether gender or nationality of the bargainer has an effect in the final price obtained.<sup>10</sup> The study took place in Mexico because of the prevalence of price haggling in informal markets. Salespersons working at various informal shops were approached by one of three bargainers looking for a specific product. After an initial price was stated by the vendor, bargaining took place. The response was the difference between the initial and the final price of the product. The bargainers were a Spanish-speaking Hispanic male, a Spanish-speaking Hispanic female, and an Anglo non-Spanish-speaking male. The following table summarizes the results:

Bargainer	<i>n</i>	Average reduction
Hispanic male	40	1.055
Anglo male	40	1.050
Hispanic female	40	2.310

- (a) To compare the mean reductions in price, what are the degrees of freedom for the ANOVA  $F$  statistic?
- (b) The reported test statistic is  $F = 8.708$ . Give an approximate (from a table) or exact (from software)  $P$ -value. What do you conclude?
- (c) To what extent do you think the results of this study can be generalized? Give reasons for your answer.



### 12.28 The effect of increased variation within groups.

The *One-Way ANOVA* applet lets you see how the  $F$  statistic and the  $P$ -value depend on the variability of the data within groups, the sample size, and the differences among the means.

- (a) The black dots are at the means of the three groups. Move these up and down until you get a configuration that gives a  $P$ -value of about 0.01. What is the value of the  $F$  statistic?
- (b) Now increase the variation within the groups by sliding the standard deviation bar to the right. Describe what happens to the  $F$  statistic and the  $P$ -value.
- (c) Using between- and within-group variation, explain why the  $F$  statistic and  $P$ -value change in this way.



### 12.29 The effect of increased variation between groups.

Set the pooled standard error for the *One-Way ANOVA* applet at a middle value. Drag the black dots so that they are approximately equal.

- (a) What is the  $F$  statistic? Give its  $P$ -value.
- (b) Drag the mean of the second group up and the mean of the third group down. Describe the effect on the  $F$  statistic and its  $P$ -value. Explain why they change in this way.



### 12.30 The effect of increased sample size.

Set the pooled standard error for the *One-Way ANOVA* applet at a middle value and drag the black dots so that the means are roughly 5.00, 4.50, and 5.25, respectively.

- (a) What are the  $F$  statistic, its degrees of freedom, and the  $P$ -value?
- (b) Slide the sample size bar to the right so  $n = 80$ . Also drag the black dots back to the values of 5.00, 4.50, and 5.25, respectively. What are the  $F$  statistic, its degrees of freedom, and the  $P$ -value?
- (c) Explain why the  $F$  statistic and  $P$ -value change in this way as  $n$  increases.



### 12.31 Financial incentives for weight loss.

The use of financial incentives has shown promise in promoting weight loss and healthy behaviors. In one study, 104 employees of the Children's Hospital of Philadelphia, with BMIs of 30 to 40 kilograms per square meter ( $\text{kg}/\text{m}^2$ ), were each randomly assigned to one of three weight-loss programs.<sup>11</sup> Participants in the control program were provided a link to weight-control information. Participants in the individual-incentive program received this link but were also told that \$100 would be given to them each time they met or exceeded their target monthly weight loss. Finally, participants in the group-incentive program received similar information and financial incentives as the individual-incentive program but were also told that they were placed in secret groups of 5 and at the end of each 4-week period, those in their group who met their goals throughout the period would equally split an additional \$500. The study ran for 24 weeks and the total change in weight (in pounds) was recorded.  LOSS

- (a) Make a table giving the sample size, mean, and standard deviation for each group.
- (b) Is it reasonable to pool the variances? Explain your answer.
- (c) Generate a histogram for each of the programs. Can we feel confident that the sample means are approximately Normal? Defend your answer.

### 12.32 Financial incentives for weight loss, continued.

Refer to the previous exercise.  LOSS

- (a) Analyze the change in weight using analysis of variance. Report the test statistic, degrees of freedom,  $P$ -value, and your conclusions.
- (b) Even though you assessed the model assumptions in the previous exercise, let's check the assumptions again by examining the residuals. Summarize your findings.
- (c) Compare the groups using the least-significant differences method.
- (d) Using the results from parts (a), (b), and (c), write a short summary of your conclusions.

### 12.33 Changing the response variable.

Refer to the previous two exercises, where we compared three weight-loss programs using change in weight measured in pounds. Suppose that you decide to instead make the comparison using change in weight measured in kilograms.  LOSS

- (a) Convert the weight loss from pounds to kilograms by dividing each response by 2.2.

(b) Analyze these new weight changes using analysis of variance. Compare the test statistic, degrees of freedom, and  $P$ -value you obtain here with those reported in part (a) of the previous exercise. Summarize what you find.

### 12.34 Do labels matter?

A study was performed to examine the self-identification of college students of Asian descent with various identity categories and assess whether there are attitudinal differences across these categories. Undergraduates at a large midwestern university who had identified themselves as being of Asian descent on their admission application were asked to participate in the study.<sup>12</sup> A total of 620 undergraduates filled out the survey. One question classified the participants into groups by asking them to indicate the option with which they primarily identify: (a) Asian American, (b) specific ethnicity (for example, Chinese), (c) ethnicity American (for example, Chinese American), and (d) other. The responses to the remaining survey items were then compared across these four groups. One item was “The campus is supportive of Asian American students.” Responses were on a four-point scale (1 = strongly disagree, 4 = strongly agree). A summary of the results follows:

Label	$n$	$\bar{x}$
Asian American	130	2.93
Specific ethnicity	248	3.00
Ethnicity American	174	3.01
Other	68	3.39

- What are the numerator and denominator degrees of freedom for the  $F$  test?
- Using the formula on page 661 and the preceding results, calculate SSG.
- Given  $SSE = 797.25$ , use your result from part (b) to compute the  $F$  statistic.
- Compute the  $P$ -value and state your conclusions.
- Without doing any additional analysis, describe the pattern in the means that is likely responsible for your conclusions in part (d).

### 12.35 The multiple-play strategy.

Multiple play is a bundling strategy through which multiple services are provided over a single network. A common triple-play service these days is Internet, television, and telephone. The market for this service has become a key battleground among telecommunication, cable, and broadband service providers. A study compared the pricing (average monthly cost in U.S. dollars) among triple-play providers using DSL, cable, or fiber platforms.<sup>13</sup> The following table summarizes the results for 47 providers.

Group	$n$	$\bar{x}$	$s$
DSL	19	104.49	26.09
Cable	20	119.98	40.39
Fiber	8	83.87	31.78

- Plot the means versus the platform type. Does there appear to be a difference in pricing?
- Is it reasonable to assume that the variances are equal? Explain.
- The  $F$  statistic is 3.39. Give the degrees of freedom and either an approximate (from a table) or

an exact (from software)  $P$ -value. What do you conclude?

### 12.36 The two-sample $t$ test and one-way ANOVA.

Refer to the diet and mood data in Exercise 7.74 (page 469). Find the two-sample pooled  $t$  statistic for comparing the two energy-restricted diets. Then formulate the problem as an ANOVA and report the results of this analysis. Verify that  $F = t^2$ .

### 12.37 Do we experience emotions differently?

Do people from different cultures experience emotions differently? One study designed to examine this question collected data from 410 college students from five different cultures.<sup>14</sup> The participants were asked to record, on a 1 (never) to 7 (always) scale, how much of the time they typically felt eight specific emotions. These were averaged to produce the global emotion score for each participant. Here is a summary of this measure:

Culture	$n$	Mean ( $s$ )
European American	46	4.39 (1.03)
Asian American	33	4.35 (1.18)
Japanese	91	4.72 (1.13)
Indian	160	4.34 (1.26)
Hispanic American	80	5.04 (1.16)

Note that the convention of giving the standard deviations in parentheses after the means saves a great deal of space in a table such as this.

(a) From the information given, do you think that we need to be concerned that a possible lack of Normality in the data will invalidate the conclusions that we might draw using ANOVA to analyze the data? Give reasons for your answer.

(b) Is it reasonable to use a pooled standard deviation for these data? Why or why not?

(c) The ANOVA  $F$  statistic was reported as 5.69. Give the degrees of freedom and either an approximate (from a table) or an exact (from software)  $P$ -value. Sketch a picture of the  $F$  distribution that illustrates the  $P$ -value. What do you conclude?

(d) Without doing any additional formal analysis, describe the pattern in the means that appears to be responsible for your conclusion in part (c). Are there pairs of means that are quite similar?

### 12.38 The emotions study, continued.

Refer to the previous exercise. The experimenters also measured emotions in some different ways. For a period of a week, each participant carried a device that sounded an alarm at random times during a 3-hour interval 5 times a day. When the alarm sounded, participants recorded several mood ratings indicating their emotions for the time immediately preceding the alarm. These responses were combined to form two variables: frequency, the number of emotions recorded, expressed as a percent; and intensity, an average of the intensity scores measured on a scale of 0 to 6. At the end of the 1-week experimental period, the subjects were asked to recall the percent of time that they experienced different emotions. This variable was called “recall.” Here is a summary of the results:

Culture	$n$	Frequency mean ( $s$ )	Intensity mean ( $s$ )	Recall mean ( $s$ )
European American	46	82.87 (18.26)	2.79 (0.72)	49.12 (22.33)

Asian American	33	72.68 (25.15)	2.37 (0.60)	39.77 (23.24)
Japanese	91	73.36 (22.78)	2.53 (0.64)	43.98 (22.02)
Indian	160	82.71 (17.97)	2.87 (0.74)	49.86 (21.60)
Hispanic American	80	92.25 (8.85)	3.21 (0.64)	59.99 (24.64)
<i>F</i> statistic		11.89	13.10	7.06

(a) For each response variable state whether or not it is reasonable to use a pooled standard deviation to analyze these data. Give reasons for your answer.

(b) Give the degrees of freedom for the *F* statistics and find the associated *P*-values. Summarize what you can conclude from these ANOVA analyses.

(c) Summarize the means, paying particular attention to similarities and differences across cultures and across variables. Include the means from the previous exercise in your summary.

(d) The European American and Asian American subjects were from the University of Illinois, the Japanese subjects were from two universities in Tokyo, the Indian subjects were from eight universities in or near Kolkata, and the Hispanic American subjects were from California State University at Fresno. Participants were paid \$25 or an equivalent monetary incentive for the Japanese and Indians. Ads were posted on or near the campuses to recruit volunteers for the study. Discuss how these facts influence your conclusions and the extent to which you would generalize the results.

(e) The percents of female students in the samples were as follows: European American, 83%; Asian American, 67%; Japanese, 63%; Indian, 64%; and Hispanic American, 79%. Use a chi-square test to compare these proportions (see Section 9.2, page 551) and discuss how this information influences your interpretation of the results that you have found in this exercise.

### 12.39 The effects of two stimulant drugs.

An experimenter was interested in investigating the effects of two stimulant drugs (labeled A and B). She divided 25 rats equally into 5 groups (placebo, Drug A low, Drug A high, Drug B low, and Drug B high) and, 20 minutes after injection of the drug, recorded each rat's activity level (higher score is more active). The following table summarizes the results:

Treatment	$\bar{x}$	$s^2$
Placebo	11.80	17.20
Low A	15.25	13.10
High A	18.55	10.25
Low B	16.15	7.75
High B	17.10	12.50

(a) Plot the means versus the type of treatment. Does there appear to be a difference in the activity level? Explain.

(b) Is it reasonable to assume that the variances are equal? Explain your answer, and if reasonable, compute  $s_p$ .

(c) Give the degrees of freedom for the *F* statistic.

(d) The *F* statistic is 2.64. Find the associated *P*-value and state your conclusions.

### 12.40 Restaurant ambiance and consumer behavior.

There have been numerous studies investigating the effects of restaurant ambiance on consumer behavior. One study investigated the effects of musical genre on consumer spending.<sup>15</sup> At a single high-end restaurant in England over a 3-week period, there were a total of 141 participants; 49 of them were subjected to background pop music (for example, Britney Spears, Culture Club, and Ricky Martin) while dining, 44 to background classical music (for example, Vivaldi, Handel, and Strauss), and 48 to no background music. For each participant, the total food bill, adjusted for time spent dining, was recorded. The following table summarizes the means and standard deviations (in British pounds):

Background music	Mean bill	<i>n</i>	<i>s</i>
Classical	24.130	44	2.243
Pop	21.912	49	2.627
None	21.697	48	3.332
Total	22.531	141	2.969

- Plot the means versus the type of background music. Does there appear to be a difference in spending?
- Is it reasonable to assume that the variances are equal? Explain.
- The  $F$  statistic is 10.62. Give the degrees of freedom and either an approximate (from a table) or an exact (from software)  $P$ -value. What do you conclude?
- Refer back to part (a). Without doing any formal analysis, describe the pattern in the means that is likely responsible for your conclusion in part (c).
- To what extent do you think the results of this study can be generalized to other settings? Give reasons for your answer.

### 12.41 Writing contrasts.

Return to the eye study described in Example 12.25 (page 673). Let  $\mu_1$ ,  $\mu_2$ ,  $\mu_3$ , and  $\mu_4$  represent the mean scores for blue, brown, gaze down, and green eyes.

- Because a majority of the population in this study are Hispanic (eye color predominantly brown), we want to compare the average score of the brown eyes with the average of the other two eye colors. Write a contrast that expresses this comparison.
- Write a contrast to compare the average score when the model is looking at you versus the score when looking down.

### 12.42 Analyzing contrasts.

Answer the following questions for the two contrasts that you defined in Exercise 12.41.  EYES

- For each contrast give  $H_0$  and an appropriate  $H_a$ . In choosing the alternatives you should use information given in the description of the problem, but you may not consider any impressions obtained by inspection of the sample means.
- Find the values of the corresponding sample contrasts  $c_1$  and  $c_2$ .
- Calculate the standard errors  $SE_{c_1}$  and  $SE_{c_2}$ .
- Give the test statistics and approximate  $P$ -values for the two significance tests. What do you

conclude?

(e) Compute 95% confidence intervals for the two contrasts.

### 12.43 College dining facilities.

University and college food service operations have been trying to keep up with the growing expectations of consumers in regard to the overall campus dining experience. Since customer satisfaction has been shown to be associated with repeat patronage and new customers through word-of-mouth, a public university in the Midwest took a sample of patrons from their eating establishments and asked them about their overall dining satisfaction.<sup>16</sup> The following table summarizes the results for three groups of patrons:

Category	$\bar{x}$	$n$	$s$
Student—meal plan	3.44	489	0.804
Faculty—meal plan	4.04	69	0.824
Student—no meal plan	3.47	212	0.657

(a) Is it reasonable to use a pooled standard deviation for these data? Why or why not? If yes, compute it.

(b) The ANOVA  $F$  statistic was reported as 17.66. Give the degrees of freedom and either an approximate (from a table) or an exact (from software)  $P$ -value. Sketch a picture of the  $F$  distribution that illustrates the  $P$ -value. What do you conclude?

(c) Prior to performing this survey, food service operations thought that satisfaction among faculty would be higher than satisfaction among students. Use the results in the table to test this contrast. Make sure to specify the null and alternative hypotheses, test statistic, and  $P$ -value.

### 12.44 Animals on product labels?

Recall Exercise 7.72 (page 469). This experiment actually involved comparing product preference for a group of consumers who were “primed” and two groups of consumers who served as controls. A bottle of MagicCoat pet shampoo was the product, and participants indicated their attitude toward this product on a seven-point scale (from 1 = dislike very much to 7 = like very much). The bottle of shampoo had either a picture of a collie on the label or just the wording. Also, prior to giving this score, participants were asked to do a word find where four of the words were shown to all groups (pet, grooming, bottle, label) and four were either related to the image (dog, collie, puppy, woof) or image conflicting (cat, feline, kitten, meow). A summary of the groups follows:  **BPREF1**

Group	Label with dog	Image words	$n$
1	Y	Y	22
2	Y	N	20
3	N	Y	10

(a) Use graphical and numerical methods to describe the data.

(b) Run the ANOVA and report the results.

(c) Examine the assumptions necessary for inference using your results in part (a) and an examination of the residuals. Summarize your findings.

(d) Use a multiple-comparisons method to compare the three groups. State your conclusions.

## 12.45 Do isoflavones increase bone mineral density?

Kudzu is a plant that was imported to the United States from Japan and now covers over seven million acres in the South. The plant contains chemicals called isoflavones that have been shown to have beneficial effects on bones. One study used three groups of rats to compare a control group with rats that were fed either a low dose or a high dose of isoflavones from kudzu.<sup>17</sup> One of the outcomes examined was the bone mineral density in the femur (in grams per square centimeter).

Here are the data:  **BMD**

Treatment	Bone mineral density (g/cm <sup>2</sup> )							
Control	0.228	0.207	0.234	0.220	0.217	0.228	0.209	0.221
	0.204	0.220	0.203	0.219	0.218	0.245	0.210	
Low dose	0.211	0.220	0.211	0.233	0.219	0.233	0.226	0.228
	0.216	0.225	0.200	0.208	0.198	0.208	0.203	
High dose	0.250	0.237	0.217	0.206	0.247	0.228	0.245	0.232
	0.267	0.261	0.221	0.219	0.232	0.209	0.255	

- Use graphical and numerical methods to describe the data.
- Examine the assumptions necessary for ANOVA. Summarize your findings.
- Run the ANOVA and report the results.
- Use a multiple-comparisons method to compare the three groups.
- Write a short report explaining the effect of kudzu isoflavones on the femur of the rat.

## 12.46 Do poets die young?

According to William Butler Yeats, “She is the Gaelic muse, for she gives inspiration to those she persecutes. The Gaelic poets die young, for she is restless, and will not let them remain long on earth.” One study designed to investigate this issue examined the age at death for writers from different cultures and genders.<sup>18</sup> Three categories of writers examined were novelists, poets, and nonfiction writers. The ages at death for female writers in these categories from North America are given in Table 12.1. Most of the writers are from the United States, but Canadian and Mexican writers are also included.  **POETS**

**TABLE 12.1** Age at Death for North American Women Writers

Type	Age at death (years)														
Novels ( $n = 67$ )	57	90	67	56	90	72	56	90	80	74	73	86	53	72	86
	82	74	60	79	80	79	77	64	72	88	75	79	74	85	71
	78	57	54	50	59	72	60	77	50	49	73	39	73	61	90
	77	57	72	82	54	62	74	65	83	86	73	79	63	72	85
	91	77	66	75	90	35	86								
Poems ( $n = 32$ )	88	69	78	68	72	60	50	47	74	36	87	55	68	75	78
	85	69	38	58	51	72	58	84	30	79	90	66	45	70	48
	31	43													

Nonfiction ( $n = 24$ )	74	86	87	68	76	73	63	78	83	86	40	75	90	47	91
	94	61	83	75	89	77	86	66	97						

- Use graphical and numerical methods to describe the data.
- Examine the assumptions necessary for ANOVA. Summarize your findings.
- Run the ANOVA and report the results.
- Use a contrast to compare the poets with the two other types of writers. Do you think that the quotation from Yeats justifies the use of a one-sided alternative for examining this contrast? Explain your answer.
- Use another contrast to compare the novelists with the nonfiction writers. Explain your choice for an alternative hypothesis for this contrast.
- Use a multiple-comparisons procedure to compare the three means. How do the conclusions from this approach compare with those using the contrasts?

### 12.47 Exercise and healthy bones.

Many studies have suggested that there is a link between exercise and healthy bones. Exercise stresses the bones and this causes them to get stronger. One study examined the effect of jumping on the bone density of growing rats.<sup>19</sup> There were three treatments: a control with no jumping, a low-jump condition (the jump height was 30 centimeters), and a high-jump condition (60 centimeters). After 8 weeks of 10 jumps per day, 5 days per week, the bone density of the rats (expressed in milligrams per cubic centimeter) was measured. Here are the data:  **JUMP**

Group	Bone density (mg/cm <sup>3</sup> )												
Control	611	621	614	593	593	653	600	554	603	569			
Low jump	635	605	638	594	599	632	631	588	607	596			
High jump	650	622	626	626	631	622	643	674	643	650			

- Make a table giving the sample size, mean, and standard deviation for each group of rats. Is it reasonable to pool the variances?
- Run the analysis of variance. Report the  $F$  statistic with its degrees of freedom and  $P$ -value. What do you conclude?

### 12.48 Exercise and healthy bones, continued.

Refer to the previous exercise.  **JUMP**

- Examine the residuals. Is the Normality assumption reasonable for these data?
- Use the Bonferroni or another multiple-comparisons procedure to determine which pairs of means differ significantly. Summarize your results in a short report. Be sure to include a graph.

### 12.49 Two contrasts of interest for the stimulant study.

Refer to Exercise 12.39 (page 684). There are two comparisons of interest to the experimenter. They are (1) placebo versus the average of the two low-dose treatments; and (2) the difference between

High A and Low A versus the difference between High B and Low B.

- (a) Express each contrast in terms of the means ( $\mu$ 's) of the treatments.
- (b) Give estimates with standard errors for each of the contrasts.
- (c) Perform the significance tests for the contrasts. Summarize the results of your tests and your conclusions.

### 12.50 Orthogonal polynomial contrasts.

Recall the Facebook friends study. In Example 12.16 (page 664) we used a contrast to assess the linear trend between the social attractiveness score and number of Facebook friends. With orthogonal polynomial contrasts, we can assess the contributions of different polynomial trends to the overall pattern. Given the 5 equally spaced levels of the factor in this problem, we can investigate up to a quartic ( $x^4$ ) trend. The derivation of the coefficients is beyond the scope of this book, so we will just investigate the trends here. The coefficients for the linear, quadratic, and cubic trends follow:  FRIENDS

Trend	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$
Linear	-2	-1	0	1	2
Quadratic	2	-1	-2	-1	2
Cubic	-1	2	0	-2	1

- (a) Plot the  $a_i$  versus  $i$  for the linear trend. Describe the pattern. Suppose that all the  $\mu_i$  were constant. What would the value of  $\psi$  equal?
- (b) Plot the  $a_i$  versus  $i$  for the quadratic trend. Describe the pattern. Suppose that all the  $\mu_i$  were constant. What would the value of  $\psi$  equal? Suppose that  $\mu_i = 5i$  (that is, a linear trend). What would the value of  $\psi$  equal?
- (c) Construct the sample contrasts for the quadratic and cubic trends using the Facebook data.
- (d) Test the hypotheses that there is a quadratic trend and that there is a cubic trend. Combine these results with the earlier linear trend results. What do you conclude?

### 12.51 A comparison of different types of scaffold material.

One way to repair serious wounds is to insert some material as a scaffold for the body's repair cells to use as a template for new tissue. Scaffolds made from extracellular material (ECM) are particularly promising for this purpose. Because they are made from biological material, they serve as an effective scaffold and are then resorbed. Unlike biological material that includes cells, however, they do not trigger tissue rejection reactions in the body. One study compared six types of scaffold material.<sup>20</sup> Three of these were ECMs and the other three were made of inert materials (MAT). There were three mice used per scaffold type. The response measure was the percent of glucose phosphated isomerase (Gpi) cells in the region of the wound. A large value is good, indicating that there are many bone marrow cells sent by the body to repair the tissue.  ECM

Material	Gpi (%)		
ECM1	55	70	70
ECM2	60	65	65
ECM3	75	70	75

MAT1	20	25	25
MAT2	5	10	5
MAT3	10	15	10

(a) Make a table giving the sample size, mean, and standard deviation for each of the six types of material. Is it reasonable to pool the variances? Note that the sample sizes are small and the data are rounded.

(b) Run the analysis of variance. Report the  $F$  statistic with its degrees of freedom and  $P$ -value. What do you conclude?

### 12.52 A comparison of different types of scaffold material, continued.

Refer to the previous exercise.  ECM

(a) Examine the residuals. Is the Normality assumption reasonable for these data?

(b) Use the Bonferroni or another multiple-comparisons procedure to determine which pairs of means differ significantly. Summarize your results in a short report. Be sure to include a graph.

(c) Use a contrast to compare the three ECM materials with the three other materials. Summarize your conclusions. How do these results compare with those that you obtained from the multiple-comparisons procedure in part (b)?

### 12.53 Contrasts for the massage study.

Refer to Exercise 12.26 (page 681). There are several comparisons of interest in this study. They are (1) usual care versus the average of the massage groups; (2) the average of the two 30-minute massage groups versus the average of the two 60-minute massage groups; and (3) the difference between a 30-minute massage once a week and twice a week versus the difference between a 60-minute massage once a week and twice a week.

(a) Express each contrast in terms of the means ( $\mu$ 's) of the treatments.

(b) Give estimates with standard errors for each of the contrasts.

(c) Perform the significance tests for the contrasts. Summarize the results of your tests and your conclusions.

### 12.54 A dandruff study.

Analysis of variance methods are often used in clinical trials where the goal is to assess the effectiveness of one or more treatments for a particular medical condition. One such study compared three treatments for dandruff and a placebo. The treatments were 1% pyrithione zinc shampoo (PyrI), the same shampoo but with instructions to shampoo two times (PyrII), 2% ketoconazole shampoo (Keto), and a placebo shampoo (Placebo). After six weeks of treatment, eight sections of the scalp were examined and given a score that measured the amount of scalp flaking on a 0 to 10 scale. The response variable was the sum of these eight scores. An analysis of the baseline flaking measure indicated that randomization of patients to treatments was successful in that no differences were found between the groups. At baseline there were 112 subjects in each of the three treatment groups and 28 subjects in the Placebo group. During the clinical trial, 3 dropped out from the PyrII group and 6 from the Keto group. No patients dropped out of the other two groups. 

**DANDRUFF**

(a) Find the mean, standard deviation, and standard error for the subjects in each group. Summarize these, along with the sample sizes, in a table and make a graph of the means.

(b) Run the analysis of variance on these data. Write a short summary of the results and your conclusion. Be sure to include the hypotheses tested, the test statistic with degrees of freedom, and the  $P$ -value.

### 12.55 The dandruff study, continued.

Refer to the previous exercise.  **DANDRUFF**

(a) Plot the residuals versus case number (the first variable in the data set). Describe the plot. Is there any pattern that would cause you to question the assumption that the data are independent?

(b) Examine the standard deviations for the four treatment groups. Is there a problem with the assumption of equal standard deviations for ANOVA in this data set? Explain your answer.

(c) Create Normal quantile plots for each treatment group. What do you conclude from these plots?

(d) Obtain the residuals from the analysis of variance and create a Normal quantile plot of these. What do you conclude?

### 12.56 Comparing each pair of dandruff treatments.

Refer to Exercise 12.54. Use the Bonferroni or another multiple-comparisons procedure that your software provides to compare the individual group means in the dandruff study. Write a short summary of your conclusions.  **DANDRUFF**

### 12.57 Testing several contrasts from the dandruff study.

Refer to Exercise 12.54. There are several natural contrasts in this experiment that describe comparisons of interest to the experimenters. They are (1) Placebo versus the average of the other three treatments; (2) Keto versus the average of the two Pyr treatments; and (3) PyrI versus PyrII.

 **DANDRUFF**

(a) Express each of these three contrasts in terms of the means ( $\mu$ 's) of the treatments.

(b) Give estimates with standard errors for each of the contrasts.

(c) Perform the significance tests for the contrasts. Summarize the results of your tests and your conclusions.

### 12.58 Changing the response variable.

Refer to Exercise 12.51 (page 687), where we compared six types of scaffold material to repair wounds. The data are given as percents ranging from 5 to 75.  **ECM**

(a) Convert these percents into their decimal form by dividing by 100. Calculate the transformed means, standard deviations, and standard errors and summarize them, along with the sample sizes, in a table.

(b) Explain how you could have calculated the table entries directly from the table you gave in part

(a) of Exercise 12.51.

(c) Analyze the decimal forms of the percents using analysis of variance. Compare the test statistic, degrees of freedom,  $P$ -value, and conclusion you obtain here with the corresponding values that you found in Exercise 12.51.

### 12.59 More on changing the response variable.

Refer to the previous exercise and Exercise 12.51 (page 687). A calibration error was found with the device that measured  $G_{pi}$ , which resulted in a shifted response. Add 5% to each response and redo the calculations. Summarize the effects of transforming the data by adding a constant to all responses.  ECM

### 12.60 Linear transformation of the response variable.

Refer to the previous two exercises. Can you suggest a general conclusion regarding what happens to the test statistic, degrees of freedom,  $P$ -value, and conclusion when you perform analysis of variance on data that have been transformed by multiplying the raw data by a constant and then adding another constant? (That is, if  $y$  is the original data, we analyze  $y^*$ , where  $y^* = a + by$  and  $a$  and  $b \neq 0$  are constants.)

### 12.61 Comparing three levels of reading comprehension instruction.

A study of reading comprehension in children compared three methods of instruction.<sup>21</sup> The three methods of instruction are called Basal, DRTA, and Strategies. As is common in such studies, several pretest variables were measured before any instruction was given. One purpose of the pretest was to see if the three groups of children were similar in their comprehension skills. The READING data file gives two pretest measures that were used in this study. Use one-way ANOVA to analyze these data and write a summary of your results.  READING

### 12.62 More on the reading comprehension study.

In the study described in the previous exercise, Basal is the traditional method of teaching, while DRTA and Strategies are two innovative methods based on similar theoretical considerations. The READING data file includes three response variables that the new methods were designed to improve. Analyze these variables using ANOVA methods. Be sure to include multiple comparisons or contrasts as needed. Write a report summarizing your findings.  READING

### 12.63 More on the Facebook friends study.

Refer to the Facebook friends study that we began to examine in Example 12.3 (page 648). The explanatory variable in this study is the number of Facebook friends, with possible values of 102, 302, 502, 702, and 902. When using analysis of variance we treat the explanatory variable as categorical. An alternative analysis is to use simple linear regression. Perform this analysis and summarize the results. Plot the residuals from the regression model versus the number of Facebook friends. What do you conclude?  FRIENDS

### 12.64 Overall standard deviation versus the pooled standard deviation.

The last line of an ANOVA table usually reports the total degrees of freedom and the total sums of squares. The ratio of these two (SS/df) provides an estimate of the variance when you combine all the data into one population. Explain why you would expect this variance to be larger than the pooled variance (MSE) of an ANOVA table.

### 12.65 Search the Internet.

Search the Internet or your library to find a study that is interesting to you and that used one-way ANOVA to analyze the data. First describe the question or questions of interest and then give the details of how ANOVA was used to provide answers. Be sure to include how the study authors examined the assumptions for the analysis. Evaluate how well the authors used ANOVA in this study. If your evaluation finds the analysis deficient, make suggestions for how it could be improved.

### 12.66 A power calculation exercise.

In Example 12.26 (page 676) the power calculation indicated that there was a fairly small chance of detecting the alternative given. Redo the calculations for the alternative  $\mu_1 = 39$ ,  $\mu_2 = 47$ , and  $\mu_3 = 42$ . Do you think that the choice of 10 students per treatment is adequate for this alternative?

### 12.67 Planning another emotions study.

Scores on an emotional scale were compared for five different cultures in Exercise 12.37 (page 684). Suppose that you are planning a new study using the same outcome variable. Your study will use European American, Asian American, and Hispanic American students from a large university.

- (a) Explain how you would select the students to participate in your study.
- (b) Use the data from Exercise 12.37 to perform power calculations to determine sample sizes for your study.
- (c) Write a report that could be understood by someone with limited background in statistics and that describes your proposed study and why you think it is likely that you will obtain interesting results.

### 12.68 Planning another isoflavone study.

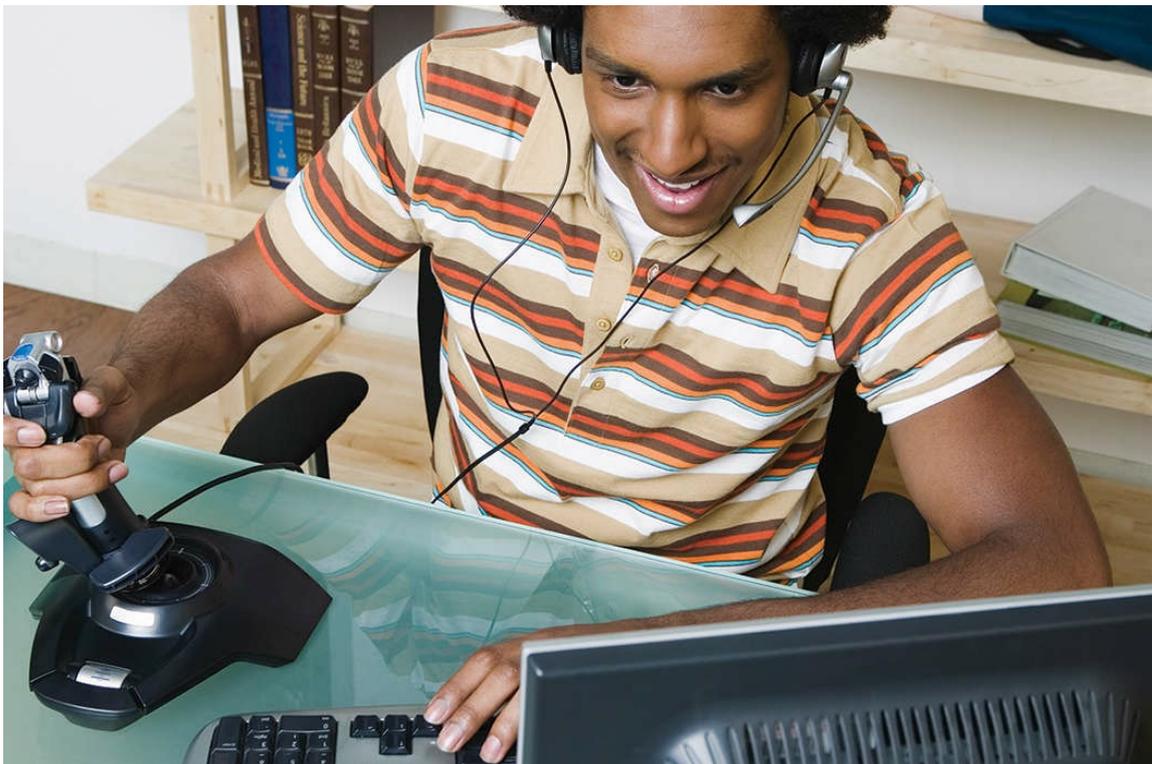
Exercise 12.45 (page 686) gave data for a bone health study that examined the effect of isoflavones on rat bone mineral density. In this study there were three groups. Controls received a placebo, and the other two groups received either a low or a high dose of isoflavones from kudzu. You are planning a similar study of a new kind of isoflavone. Use the results of the study described in Exercise 12.45 to plan your study. Write a proposal explaining why your study should be funded.

### 12.69 Planning another restaurant ambiance study.

Exercise 12.40 (page 685) gave data for a study that examined the effect of background music on total food spending at a high-end restaurant. You are planning a similar study but intend to look at total food spending at a more casual restaurant. Use the results of the study described in Exercise 12.40 to plan your study.

# 13 Two-Way Analysis of Variance

CHAPTER



**13.1 The Two-Way ANOVA Model**

**13.2 Inference for Two-Way ANOVA**

## Introduction

In this chapter, we move from one-way ANOVA, which compares the means of several populations, to two-way ANOVA. Two-way ANOVA compares the means of populations that can be classified in two ways or the mean responses in two-factor experiments.

Many of the key concepts are similar to those of one-way ANOVA, but the presence of more than one classification factor also introduces some new ideas. We once more assume that the data are approximately Normal and that although groups may have different means, they have the same standard deviation; we again pool to estimate the variance; and we again use  $F$  statistics for significance tests.

*The major difference between one-way and two-way ANOVA is in the FIT part of the model.* We will carefully study this term, and we will find much that is both new and useful. This will allow us to address comparisons such as the following:

- Can zinc supplementation reduce the occurrence and severity of malaria in both nutrient-sufficient and nutrient-deficient African children?
- What effects do the shapes of the flowers of male and female jack-in-the-pulpit plants have on the degree to which insects eat them?
- Do calcium supplements prevent bone loss in elderly people? Does this depend on whether the person is receiving adequate vitamin D?

## 13.1 The Two-Way ANOVA Model

When you complete this section, you will be able to

- Discuss the advantages of a two-way ANOVA design.
- Describe the two-way ANOVA model and when it is used for inference.
- Interpret the relationship between two factors in terms of main effects and interaction.
- Construct an interaction plot and determine whether it shows that there is interaction among the factors.

We begin with a discussion of the advantages of the two-way ANOVA design and illustrate these with some examples. Then we discuss the model and the assumptions.

### Advantages of two-way ANOVA

In one-way ANOVA, we classify populations according to one categorical variable, or factor. In the two-way ANOVA model, there are two factors, each with its own number of levels. When we are interested in the effects of two factors, a two-way design offers great advantages over several single-factor studies. Several examples will illustrate these advantages.

### Example

#### 13.1 Design 1: Does haptic feedback improve performance?

In Example 12.1 (page 645), a group of technology students wanted to see if haptic feedback (forces and vibrations applied through a joystick) is helpful in navigating a simulated game environment. To do this, they plan to randomly assign each of 60 students to one of the three joystick types and record the time it takes to complete a navigation mission.

It turns out that their simulated game has several different difficulty levels. Suppose that a second experiment is planned to compare these levels. A similar experimental design will be used, with the four difficulty levels

randomly assigned to 60 students. All students will use the standard joystick.

Here is a picture of the design of the first experiment with the sample sizes:

Joystick	<i>n</i>
1	20
2	20
3	20
Total	60

And this represents the second experiment:

Difficulty	<i>n</i>
1	15
2	15
3	15
4	15
Total	60

In the first experiment 20 students were assigned to each level of the factor for a total of 60 students. In the second experiment 15 students were assigned to each level of the factor for a total of 60 students. If each experiment takes one week, the total amount of time for the two experiments is two weeks.

Each experiment will be analyzed using one-way ANOVA. The factor in the first experiment is joystick type with three levels, and the factor in the second experiment is game difficulty level with four levels. Let's now consider combining the two experiments into one.

## Example

### 13.2 Design 2: Does haptic feedback improve performance regardless of difficulty level?

Suppose that we use a two-way approach for the simulated game problem. There are two factors, joystick type and difficulty level. Since joystick type has three levels and difficulty level has four levels, this is a  $3 \times 4$  design. This gives a total of 12 possible combinations of type and difficulty level. With a total of 60 students, we could assign each combination of type and difficulty level to 5 students. The time it takes to complete a navigation mission is the outcome variable.

Here is a picture of the two-way design with the sample sizes:

Joystick	Difficulty				Total
	1	2	3	4	
1	5	5	5	5	20
2	5	5	5	5	20
3	5	5	5	5	20
Total	15	15	15	15	60

Each combination of the factors in a two-way design corresponds to a *cell*. The  $3 \times 4$  ANOVA for the haptic feedback experiment has 12 cells, each corresponding to a particular combination of joystick type and difficulty level.

### cell

With the two-way design, notice that we have 20 students assigned to each joystick type, the same as we had for the one-way experiment for type alone. Similarly, there are 15 students assigned to each level of difficulty. Thus, the two-way design gives us the same amount of information for estimating the completion time for each level of each factor as we had with the two one-way designs. The difference is that we can collect all the information in only one experiment (in one week instead of two). By combining the two factors into one experiment, we have increased our efficiency by reducing the amount of data to be collected by half.

## Example

### 13.3 Can dietary supplementation with zinc prevent malaria?

Malaria is a serious health problem causing an estimated one million deaths per year, mostly among African children.<sup>1</sup> Several studies, run in Asia, Latin America, and developed countries, have shown zinc supplementation to be an effective control of common infections in children. Can this supplementation program also be effective in Africa, where the primary threat to a child's health is malaria? A group of researchers have set out to answer this question.<sup>2</sup>

To design a study to answer this question the researchers first need to determine an appropriate target group. Since malaria is a serious problem for young children, they will concentrate on children who are 6 months to 5 years of age. A supplement will be prepared that contains either no zinc or 10 milligrams (mg) of zinc. Because the response to zinc may be different in children who lack other

important nutrients, they decide to also take this factor into account. Specifically, their supplement will contain daily doses of essential vitamins and minerals or it will not.

## Example

### 13.4 Implementing the two-way ANOVA.

The factors for the two-way ANOVA are zinc supplementation with two levels and vitamin supplementation with two levels. There are  $2 \times 2 = 4$  cells in their study. They plan to enroll 600 children and randomly assign 150 to each of the cells. One outcome variable will be a measure of the child's T cell immune response.

Here is a table that summarizes the design:

Zinc	Vitamins		Total
	No	Yes	
No	150	150	300
Yes	150	150	300
Total	300	300	600

This example illustrates another advantage of two-way designs. Although the researchers are primarily interested in the possible benefit of zinc supplementation, they also included vitamin supplementation in the design because they suspected that the zinc effect might be different in children who are nutritionally deficient.

Consider an alternative one-way design where we assign 300 children to the two levels of zinc and ignore nutritional status. With this design we will have the same number of children at each of the zinc levels, so in this way it is similar to our two-way design.

However, suppose that there are, in fact, differences due to nutritional status. In this case, the one-way ANOVA would assign this variation to the RESIDUAL (within groups) part of the model. In the two-way ANOVA, vitamin supplementation is included as a factor, and therefore this variation is included in the FIT part of the model. *Whenever we can move variation from RESIDUAL to FIT, we reduce the  $\sigma$  of our model and increase the power of our tests.*

## Example

### 13.5 Vitamin D and osteoporosis.

Osteoporosis is a disease primarily of the elderly. People with osteoporosis have low bone mass and an increased risk of bone fractures. Over 10 million people in the United States, 1.4 million Canadians, and many millions throughout the world have this disease. Adequate calcium in the diet is necessary for strong bones, but vitamin D is also needed for the body to efficiently use calcium. High doses of calcium in the diet will not prevent osteoporosis unless there is adequate vitamin D. Exposure of the skin to the ultraviolet rays in sunlight enables our bodies to make vitamin D. However, elderly people often don't go outside as much as younger people do, and in northern areas such as Canada, there is not sufficient ultraviolet light for the body to make vitamin D, particularly in the winter months.



Suppose that we wanted to see if calcium supplements will increase bone mass (or prevent a decrease in bone mass) in an elderly Canadian population. Because of the vitamin D complication, we will make this a factor in our design.

## Example

### 13.6 Designing the osteoporosis study.

We will use a  $2 \times 2$  design for our osteoporosis study. The two factors are calcium and vitamin D. The levels of each factor will be zero (placebo) and an amount that is expected to be adequate, 800 milligrams per day (mg/d) for calcium and 300 international units per day (IU/d) for vitamin D. Women between the ages of 70 and 80 will be recruited as subjects. Bone mineral

density (BMD) will be measured at the beginning of the study, and supplements will be taken for one year. The change in BMD over the one-year period is the outcome variable. We expect a dropout rate of 20% and we would like to have about 20 subjects providing data in each group at the end of the study. We will therefore recruit 100 subjects and randomly assign 25 to each treatment combination.

Here is a table that summarizes the design with the sample sizes at the start of the study:

Calcium	Vitamin D		Total
	Placebo	300 IU/d	
Placebo	25	25	50
800 mg/d	25	25	50
Total	50	50	100

This example illustrates a third reason for using two-way designs. The effectiveness of the calcium supplement on BMD depends on having adequate vitamin D. We call this an *interaction*. In contrast, the average values for the calcium effect and the vitamin D effect are represented as *main effects*. The two-way model represents FIT as the sum of a main effect for each of the two factors *and* an interaction. One-way designs that vary a single factor and hold other factors fixed cannot discover interactions. We will discuss interactions more fully later.

*interaction*

*main effects*

These examples illustrate several reasons why two-way designs are preferable to one-way designs.

## ADVANTAGES OF TWO-WAY ANOVA

1. It is more efficient to study two factors simultaneously rather than separately.
2. We can reduce the residual variation in a model by including a second factor thought to influence the response.
3. We can investigate interactions between factors.

These considerations also apply to study designs with more than two factors. We will be content to explore only the two-way case. The choice of sampling or experimental design is fundamental to any statistical study. *Factors and levels must*

be carefully selected by an individual or team who understands both the statistical models and the issues that the study will address.



## The two-way ANOVA model

When discussing two-way models in general, we will use the labels A and B for the two factors. For particular examples and when using statistical software, it is better to use meaningful names for these categorical variables. Thus, in Example 13.2 we would say that the factors are joystick type and difficulty level and in Example 13.4 we would say that the factors are the zinc and vitamin supplementation.

The numbers of levels of the factors are often used to describe the model. Again using our earlier examples, we would say that Example 13.2 represents a  $3 \times 4$  ANOVA and Example 13.4 illustrates a  $2 \times 2$  ANOVA. In general, Factor A will have  $I$  levels and Factor B will have  $J$  levels. Therefore, we call the general two-way problem an  $I \times J$  ANOVA.

In a two-way design every level of A appears in combination with every level of B, so that  $I \times J$  groups are compared. The sample size for level  $i$  of Factor A and level  $j$  of Factor B is  $n_{ij}$ . In our examples so far, the  $n_{ij}$  have been equal but this is not required.<sup>3</sup> The total number of observations is

$$N = \sum_{i,j} n_{ij}$$

### ASSUMPTIONS FOR TWO-WAY ANOVA

We have independent SRSs of size  $n_{ij}$  from each of  $I \times J$  Normal populations. The population means  $\sigma_{ij}$  may differ, but all populations have the same standard deviation  $\sigma$ . The  $\sigma_{ij}$  and  $\sigma$  are unknown parameters.

Let  $x_{ijk}$  represent the  $k$ th observation from the population having Factor A at level  $i$  and Factor B at level  $j$ . The statistical model is

$$x_{ijk} = \sigma_{ij} + \epsilon_{ijk}$$

for  $i = 1, \dots, I$  and  $j = 1, \dots, J$  and  $k = 1, \dots, n_{ij}$ , where the deviations  $\epsilon_{ijk}$  are from an  $N(0, \sigma)$  distribution.

Similar to the one-way model, the FIT part is the group means  $\mu_{ij}$  and the RESIDUAL part is the deviations  $\epsilon_{ijk}$  of the individual observations from their

group means. To estimate a group mean  $\mu_{ij}$  we use the sample mean of the observations in the samples from this group:

$$\bar{x}_{ij} = \frac{1}{n_{ij}} \sum_k x_{ijk}$$

The  $k$  below the  $\Sigma$  means that we sum the  $n_{ij}$  observations that belong to the  $(i,j)$ th sample.

The RESIDUAL part of the model contains the unknown  $\sigma$ . We calculate the sample variances for each SRS and pool these to estimate  $\sigma^2$ :

$$s_p^2 = \frac{\sum (n_{ij} - 1) s_{ij}^2}{\sum (n_{ij} - 1)}$$

Just as in one-way ANOVA, the numerator in this fraction is SSE and the denominator is DFE. Also, DFE is the total number of observations minus the number of groups. That is,  $DFE = N - IJ$ . The estimator of  $\sigma$  is  $s_p$ .



one-way model, p. 652

## Main effects and interactions

In this section we will further explore the FIT part of the two-way ANOVA, which is represented in the model by the population means  $\mu_{ij}$ . The two-way design gives some structure to the set of means  $\mu_{ij}$ .

So far, because we have independent samples from each of  $I \times J$  groups, we have presented the problem as a one-way ANOVA with  $IJ$  groups. Each population mean  $\mu_{ij}$  is estimated by the corresponding sample mean  $\bar{x}_{ij}$ , and we can calculate sums of squares and degrees of freedom as in one-way ANOVA. In accordance with the conventions used by many computer software packages, we use the term *model* when discussing the sums of squares and degrees of freedom calculated as in one-way ANOVA with  $IJ$  groups. Thus, SSM is a model sum of squares constructed from deviations of the form  $\bar{x}_{ij} - \bar{x}$  where  $\bar{x}$  is the average of all the observations and  $\bar{x}_{ij}$  is the mean of the  $(i,j)$ th group. Similarly, DFM is simply  $IJ - 1$ .

In two-way ANOVA, the terms SSM and DFM can be further broken down into terms corresponding to a main effect for A, a main effect for B, and an AB interaction. Each of SSM and DFM is then a sum of terms:

$$SSM = SSA + SSB + SSAB$$

and

$$DFM = DFA + DFB + DFAB$$

The term SSA represents variation among the means for the different levels of Factor A. Because there are  $I$  such means,  $DFA = I - 1$  degrees of freedom.

Similarly, SSB represents variation among the means for the different levels of Factor B, with  $DFB = J - 1$ .

Interactions are a bit more involved. We can see that SSAB, which is  $SSM - SSA - SSB$ , represents the variation in the model that is not accounted for by the main effects. By subtraction we see that its degrees of freedom are

$$DFAB = (IJ - 1) - (I - 1) - (J - 1) = (I - 1)(J - 1)$$

There are many kinds of interactions. The easiest way to study them is through examples.

## Example

### 13.7 Investigating differences in sugar-sweetened beverage consumption.



Consumption of sugar-sweetened beverages (SSBs) has been linked to Type 2 diabetes and obesity. One study used data from the National Health and Nutrition Examination Survey (NHANES) to estimate SSB consumption among adults. More than 20,000 individuals provided data for this study. Individuals were divided into 3 age categories: adolescents (12 to 19 years old), young adults (20 to 34 years old), and adults ( $\geq 35$  years old).<sup>4</sup> Here are the means for the number of calories in SSBs consumed per day during 2003 to 2004 and 2007 to 2008:

Group	Year		Mean
	2004	2008	
Adolescents	336	286	311
Young adults	391	338	365

Adults	236	236	236
Mean	321	287	304

The table in Example 13.7 includes averages of the means in the rows and columns. For example, in 2004 the mean of calories in SSBs consumed per day is

$$336+391+2363=321$$

Similarly, the corresponding value for 2008 is

$$286+338+2363=286.7$$

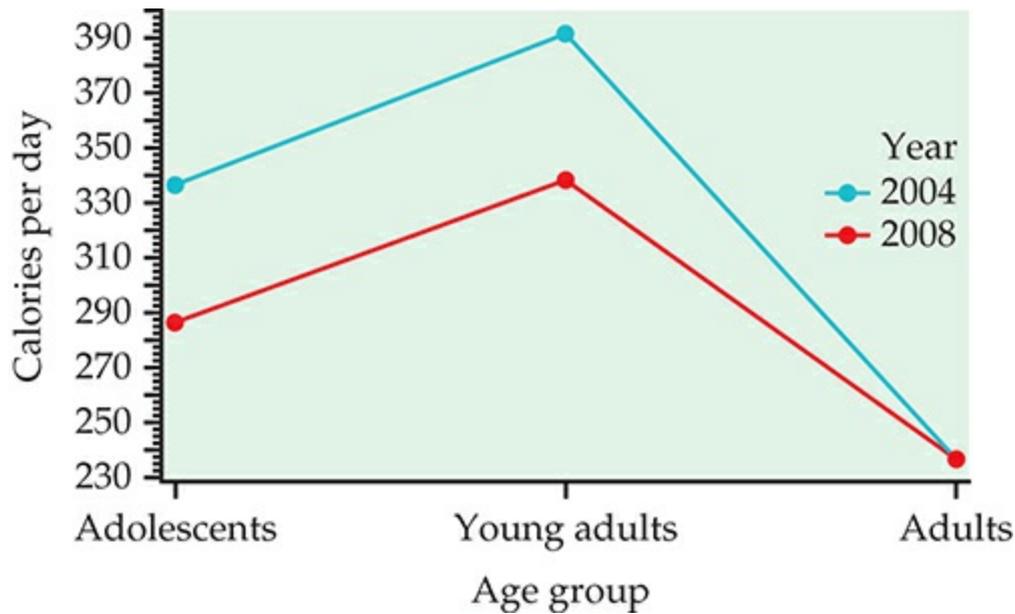
which is rounded to 287 in the table. These averages are called *marginal means* (because of their location at the *margins* of such tabulations). The grand mean (304 in this case) can be obtained by averaging either set of marginal means.

### marginal means

Figure 13.1 is a plot of the group means. From the plot we see that the calories in SSBs consumed by each group in 2008 are less than or equal to those consumed in 2004. In statistical language, there is a main effect for year. We also see that the means are different across age categories. This means that there is a main effect for age. These main effects can be described by differences between the marginal means. For example, the mean for 2004 is 321 calories and then decreases 34 calories to 287 calories in 2008. Similarly, the mean for adolescents is 311, it increases 54 calories to 365 for young adults, and then drops 129 calories to 236 for adults.

*To examine two-way ANOVA data for a possible interaction, always construct a plot similar to Figure 13.1.* In this case, it is debatable whether the two profiles (the collections of marginal means for a given year) should be considered parallel. Profiles that are roughly parallel imply that there is *no* clear interaction between the two factors. When no interaction is present, the marginal means provide a reasonable description of the two-way table of means.





**FIGURE 13.1**

Plot of the mean calories in sugar-sweetened beverages consumed per day in 2003 to 2004 and 2007 to 2008 for different age groups, for Example 13.7.

*When there is an interaction, the marginal means do not tell the whole story.* For example, with these data, the marginal mean difference between years is 34 calories. This is smaller than the difference in calories for the adolescent and young adult age classes and larger than the zero change in the adult age class. If differences of 20 to 30 calories per day are scientifically meaningful, then we would say that there is evidence for an interaction.



Interactions come in many shapes and forms. *When we find an interaction, a careful examination of the means is needed to properly interpret the data.* Simply stating that interactions are significant tells us very little. Plots of the group means, called interaction plots, are essential. Here is another example.




---

## Example

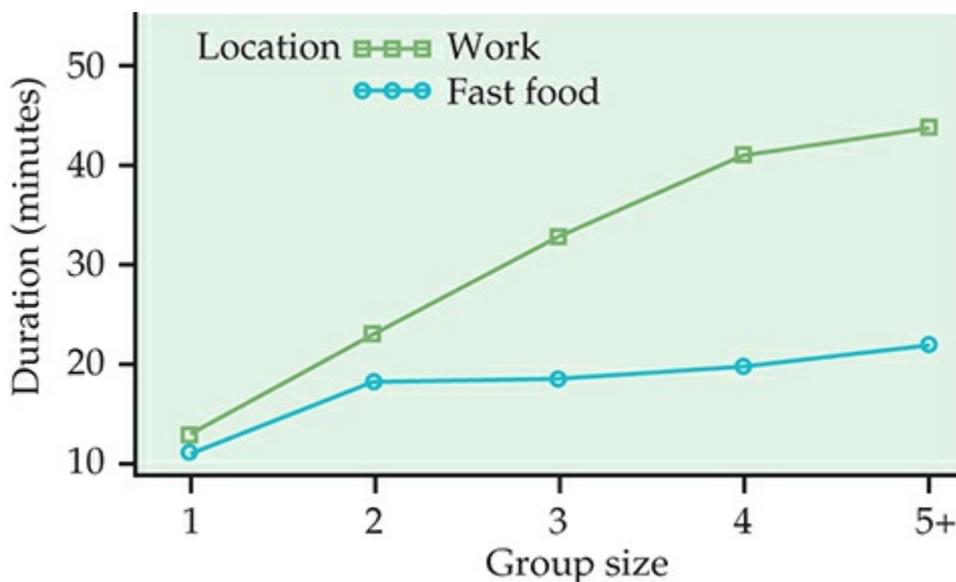
### 13.8 Eating in groups.



Some research has shown that people eat more when they eat in groups. One possible mechanism for this phenomenon is that they may spend more time eating when in a larger group. A study designed to examine this idea measured the length of time spent (in minutes) eating lunch in different settings.<sup>5</sup> Here are some data from this study:

Lunch setting	Number of People Eating					Mean
	1	2	3	4	5 or more	
Workplace	12.6	23.0	33.0	41.1	44.0	30.7
Fast-food restaurant	10.7	18.2	18.4	19.7	21.9	17.8
Mean	11.6	20.6	25.7	30.4	32.9	24.2

Figure 13.2 gives the plot of the means for this example. The patterns are not parallel, so it appears that we have an interaction. Meals take longer when there are more people present, but this phenomenon is much greater for the meals consumed at work. For fast-food eating, the meal durations are fairly similar when there is more than one person present.



## FIGURE 13.2

Plot of mean meal duration versus lunch setting and group size, for Example 13.8.

A different kind of interaction is present in the next example. Here, we must be very cautious in our interpretation of the main effects since either one of them leads to a distorted conclusion.

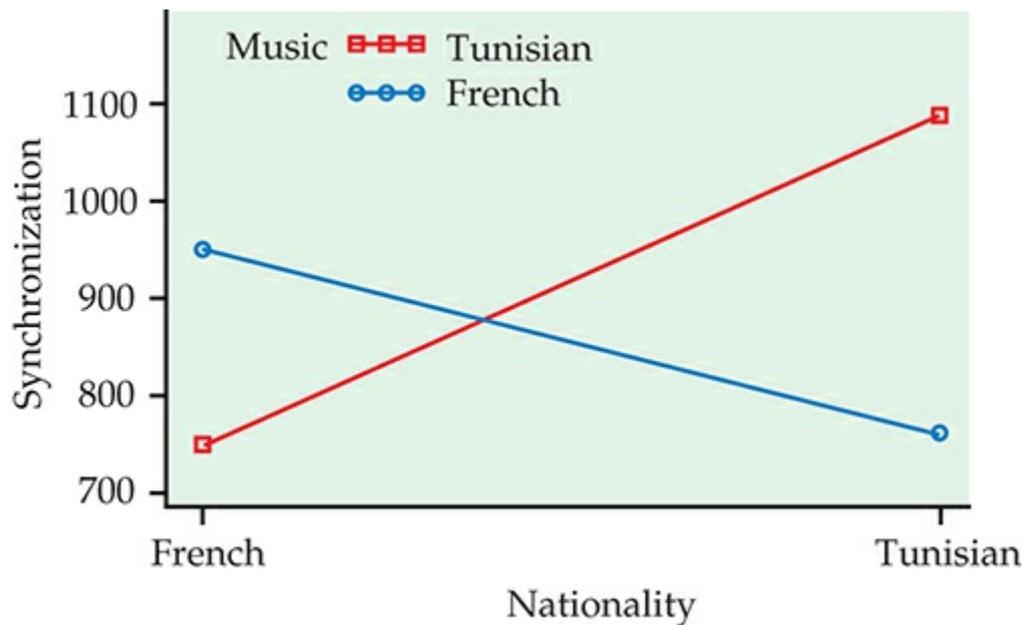
### Example

#### 13.9 We got the beat?

When we hear music that is familiar to us, we can quickly pick up the beat and our mind synchronizes with the music. However, if the music is unfamiliar, it takes us longer to synchronize. In a study that investigated the theoretical framework for this phenomenon, French and Tunisian nationals listened to French and Tunisian music.<sup>6</sup> Each subject was asked to tap in time with the music being played. A synchronization score, recorded in milliseconds, measured how well the subjects synchronized with the music. A higher score indicates better synchronization. Six songs of each music type were used. Here are the means:

Nationality	Music		Mean
	French	Tunisian	
French	950	750	850
Tunisian	760	1090	925
Mean	855	920	887

The means are plotted in Figure 13.3. In the study the researchers were not interested in main effects. Their theory predicted the interaction that we see in the figure. Subjects synchronize better with music from their own culture. The main effects, on the other hand, suggest that Tunisians synchronize better than the French (regardless of music type) and that it is easier to synchronize to Tunisian music (regardless of nationality).



**FIGURE 13.3**

Plot of mean synchronization score versus type of music for French and Tunisian nationals, for Example 13.9.

The interaction in Figure 13.3 is very different from those that we saw in Figures 13.1 and 13.2. These examples illustrate the point that it is necessary to plot the means and carefully describe the patterns when interpreting an interaction.

The design of the study in Example 13.9 allows us to examine two main effects and an interaction. However, this setting does not meet all the assumptions needed for statistical inference using the two-way ANOVA framework of this chapter. *As with one-way ANOVA, we require that observations be independent.*



In this study, we have a design that has each subject contributing data for two types of music, so these two scores will be dependent. The framework is similar to the matched pairs setting. The design is called a *repeated-measures design*. More advanced texts on statistical methods cover this important design.

repeated-measures design



matched pairs *t* test, p. 429

---

## USE YOUR KNOWLEDGE

### 13.1 What's wrong?

For each of the following, explain what is wrong and why.

- (a) A two-way ANOVA is used when the outcome variable can take only two possible values.
- (b) In a  $2 \times 3$  ANOVA each level of Factor A appears with two levels of Factor B.
- (c) The FIT part of the model in a two-way ANOVA represents the variation that is sometimes called error or residual.
- (d) In an  $I \times J$  ANOVA,  $DFAB = IJ - 1$ .

### 13.2 What's wrong?

For each of the following, explain what is wrong and why.

- (a) Parallel profiles of cell means imply that a strong interaction is present.
- (b) You can perform a two-way ANOVA only when the sample sizes are the same in all cells.
- (c) The estimate  $sp^2$  is obtained by pooling the marginal sample variances.
- (d) When interaction is present, the marginal means are always uninformative.

## 13.2 Inference for Two-Way ANOVA

When you complete this section, you will be able to

- Construct the two-way ANOVA table in terms of sources and degrees of freedom. Summarize what the  $F$  tests can tell you about main effects and interactions and what they cannot without further analysis.
- Interpret statistical software ANOVA output for a two-way ANOVA.
- Use diagnostic plots and sample statistics to check the assumptions of the two-way ANOVA model.

Inference for two-way ANOVA involves  $F$  statistics for each of the two main effects and an additional  $F$  statistic for the interaction. As with one-way ANOVA, the calculations are organized in an ANOVA table.

### The ANOVA table for two-way ANOVA

Two-way ANOVA is the statistical analysis for a two-way design with a quantitative response variable. The results of a two-way ANOVA are summarized in an ANOVA table based on splitting the total variation SST and the total degrees of freedom DFT among the two main effects and the interaction. Both the sums of squares (which measure variation) and the degrees of freedom add:

$$SST = SSA + SSB + SSAB + SSE$$

$$DFT = DFA + DFB + DFAB + DFE$$

The sums of squares are always calculated in practice by statistical software. *When the  $n_{ij}$  are not all equal, some methods of analysis can give sums of squares that do not add.*



From each sum of squares and its degrees of freedom we find the mean square in the usual way:

$$\text{mean square} = \frac{\text{sum of squares}}{\text{degrees of freedom}}$$

The significance of each of the main effects and the interaction is assessed by an  $F$  statistic that compares the variation due to the effect of interest with the

within-group variation. Each  $F$  statistic is the mean square for the source of interest divided by MSE. Here is the general form of the two-way ANOVA table:

Source	Degrees of freedom	Sum of squares	Mean square	$F$
A	$I - 1$	SSA	SSA/DFA	MSA/MSE
B	$J - 1$	SSB	SSB/DFB	MSB/MSE
AB	$(I - 1)(J - 1)$	SSAB	SSAB/DFAB	MSAB/MSE
Error	$N - IJ$	SSE	SSE/DFE	
Total	$N - 1$	SST		

There are three null hypotheses in two-way ANOVA, with an  $F$  test for each. We can test for significance of the main effect of A, the main effect of B, and the AB interaction. *It is generally good practice to examine the test for interaction first, since the presence of a strong interaction may influence the interpretation of the main effects.* Be sure to plot the means as an aid to interpreting the results of the significance tests.



## SIGNIFICANCE TESTS IN TWO-WAY ANOVA

To test the main effect of A, use the  $F$  statistic

$$F_A = \frac{MS_A}{MSE}$$

To test the main effect of B, use the  $F$  statistic

$$F_B = \frac{MS_B}{MSE}$$

To test the interaction of A and B, use the  $F$  statistic

$$F_{AB} = \frac{MS_{AB}}{MSE}$$

The  $P$ -value is the probability that a random variable having an  $F$  distribution with numerator degrees of freedom corresponding to the effect and denominator degrees of freedom equal to DFE is greater than or equal to the calculated  $F$  statistic.

The following example illustrates how to do a two-way ANOVA. As with the one-way ANOVA, we focus our attention on interpretation of the computer output.

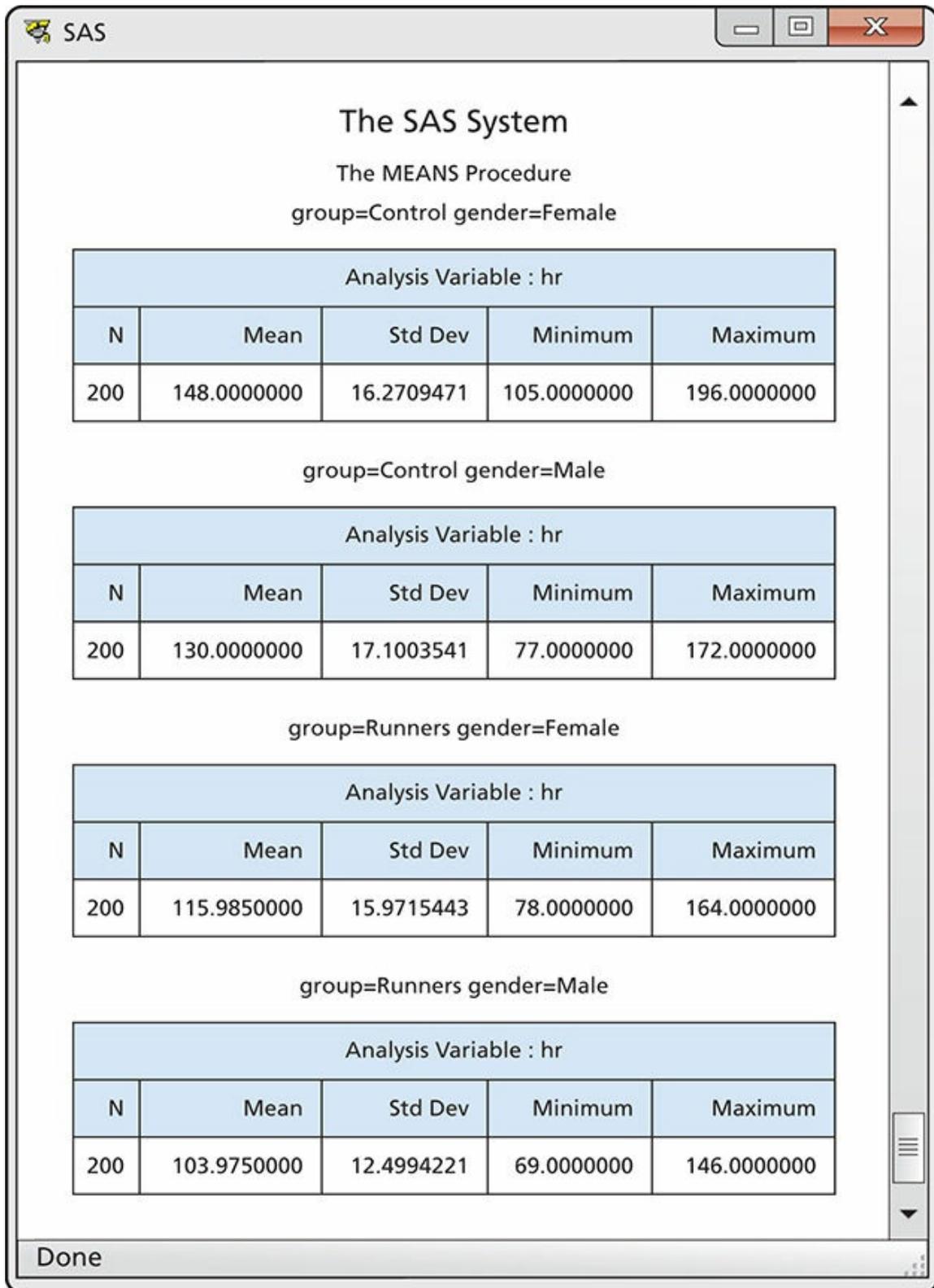
## Example

### 13.10 A study of cardiovascular risk factors.

A study of cardiovascular risk factors compared runners who averaged at least 15 miles per week with a control group described as “generally sedentary.” Both men and women were included in the study.<sup>7</sup> The design is a  $2 \times 2$  ANOVA with the factors group and gender. There were 200 subjects in each of the four combinations. One of the variables measured was the heart rate after 6 minutes of exercise on a treadmill. SAS computer analysis produced the outputs in Figure 13.4 and Figure 13.5.

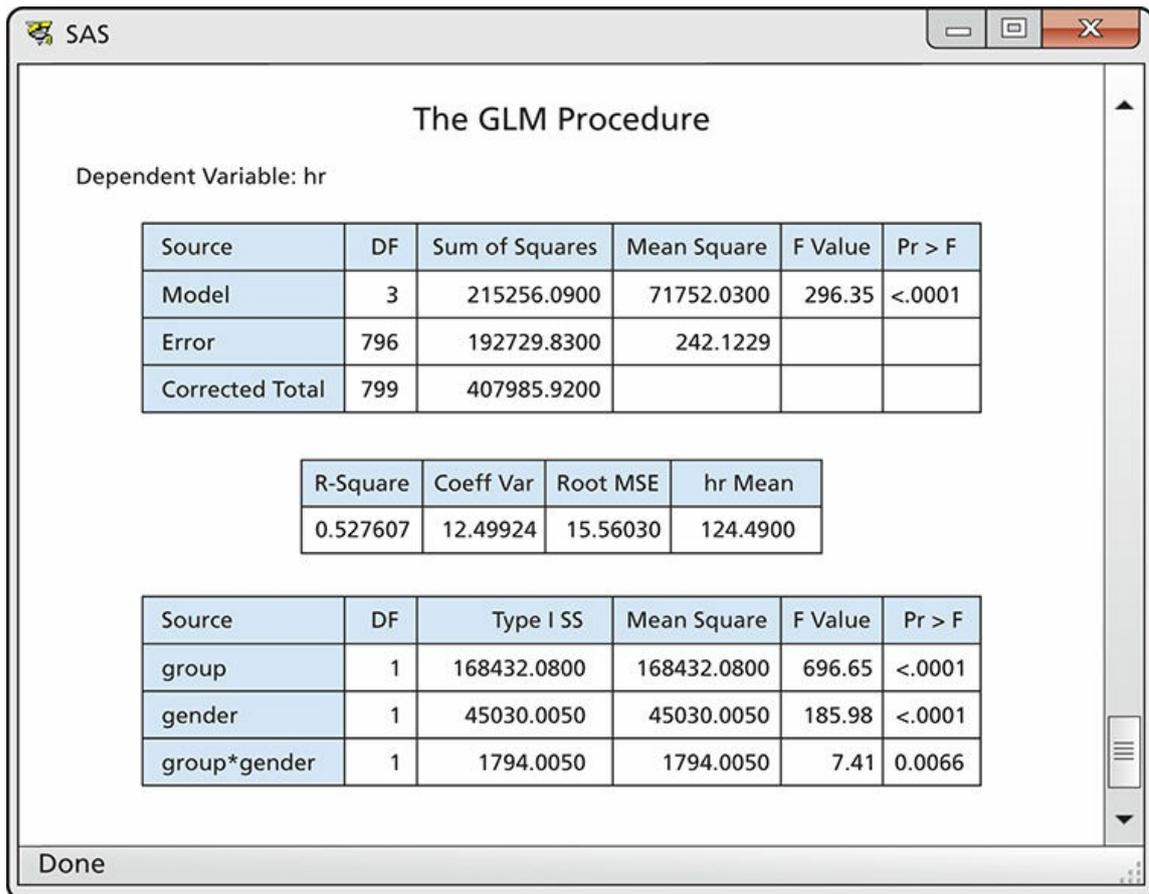


HRTRATE



**FIGURE 13.4**

Summary statistics for heart rates in the four groups of a  $2 \times 2$  ANOVA, for Example 13.10.



**FIGURE 13.5**

Two-way ANOVA output for heart rates, for Example 13.10.

We begin with the usual preliminary examination. From Figure 13.4 we see that the ratio of the largest to the smallest standard deviation is less than 2. Therefore, we are not concerned about violating the assumption of equal population standard deviations. Normal quantile plots (not shown) do not reveal any outliers, and the data appear to be reasonably Normal.

The ANOVA table at the top of the output in Figure 13.5 is in effect a one-way ANOVA with four groups: female control, female runner, male control, and male runner. In this analysis Model has 3 degrees of freedom, and Error has 796 degrees of freedom. *Since we will be relying on software to do all these calculations, it is always a good idea to do some quick arithmetic checks like degrees of freedom to make sure things make sense.* The  $F$  test and its associated  $P$ -value for this analysis refer to the hypothesis that all four groups have the same population mean. We are interested in the main effects and interaction, so we ignore this test.



The sums of squares for the group and gender main effects and the group-by-gender interaction appear at the bottom of Figure 13.5 under the heading “Type I

SS.” These sum to the sum of squares for Model. Similarly, the degrees of freedom for these sums of squares sum to the degrees of freedom for Model. Two-way ANOVA splits the variation among the means (expressed by the Model sum of squares) into three parts that reflect the two-way layout.

Because the degrees of freedom are all 1 for the main effects and the interaction, the mean squares are the same as the sums of squares. The  $F$  statistics for the three effects appear in the column labeled “F Value,” and the  $P$ -values are under the heading “Pr > F.” For the group main effect, we verify the calculation of  $F$  as follows:

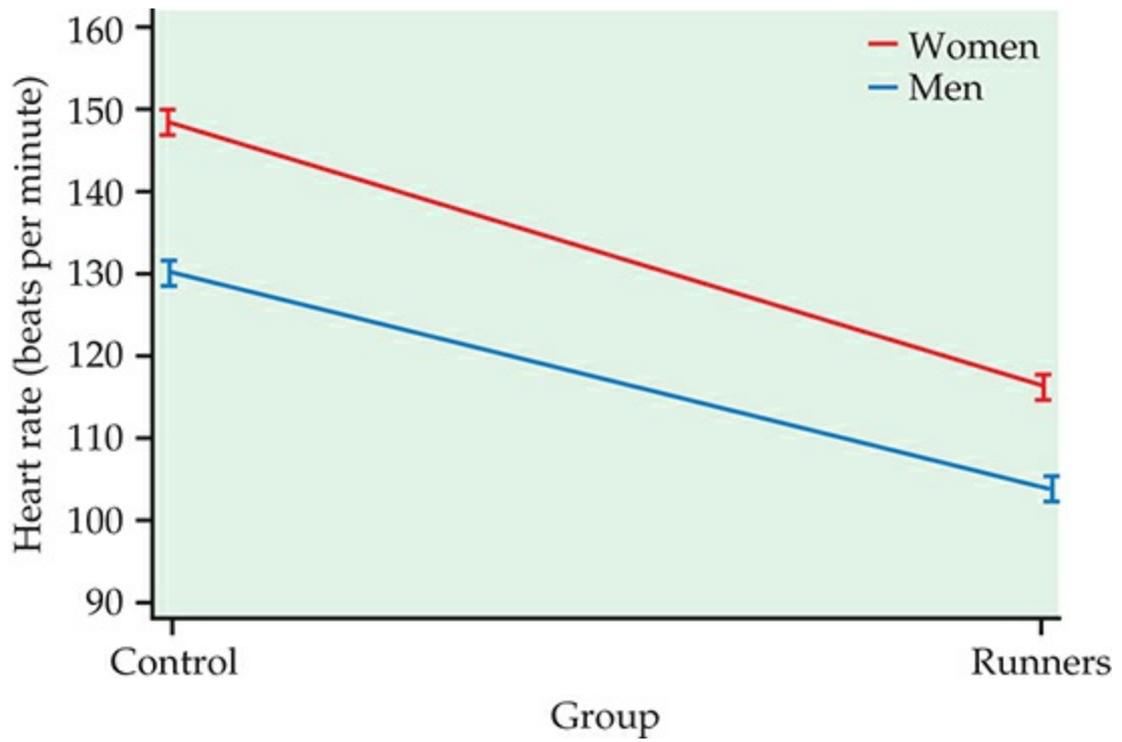
$$F = \text{MSGMSE} = 168,432,242.12 = 695.65$$

All three effects are statistically significant. The group effect has the largest  $F$ , followed by the gender effect and then the group-by-gender interaction. To interpret these results, we examine the plot of means, with bars indicating one standard error, in Figure 13.6. Note that the standard errors are quite small due to the large sample sizes. The significance of the main effect for group is due to the fact that the controls have higher average heart rates than the runners for both genders. This is the largest effect evident in the plot.

The significance of the main effect for gender is due to the fact that the females have higher heart rates than the men in both groups. The differences are not as large as those for the group effect, and this is reflected in the smaller value of the  $F$  statistic.

The analysis indicates that a complete description of the average heart rates requires consideration of the interaction in addition to the main effects. The two lines in the plot are not parallel. This interaction can be described in two ways. The female-male difference in average heart rates is greater for the controls than for the runners. Alternatively, the difference in average heart rates between controls and runners is greater for women than for men. As the plot suggests, the interaction is not large. It is statistically significant because there were 800 subjects in the study.

Two-way ANOVA output for other software is similar to that given by SAS. Figure 13.7 gives the analysis of the heart rate data using Excel and Minitab.



**FIGURE 13.6**

Plot of the group means, with standard errors indicated, for heart rates in the  $2 \times 2$  ANOVA, for Example 13.10.

Excel

	A	B	C	D	E	F	G
1	Anova: Two-Factor With Replication						
2							
3	SUMMARY	Control	Runners	Total			
4	<i>Female</i>						
5	Count	200	200	400			
6	Sum	29600	23197	52797			
7	Average	148	115.985	131.9925			
8	Variance	264.7437	255.0902	516.1478			
9							
10	<i>Male</i>						
11	Count	200	200	400			
12	Sum	26000	20795	46795			
13	Average	130	103.975	116.9875			
14	Variance	292.4221	156.2356	393.5161			
15							
16	<i>Total</i>						
17	Count	400	400				
18	Sum	55600	43992				
19	Average	139	109.98				
20	Variance	359.0877	241.2978				
21							
22							
23	ANOVA						
24	<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
25	Sample	45030	1	45030	185.9799	3.29E-38	3.853168
26	Columns	168432.1	1	168432.1	695.647	1.1E-110	3.853168
27	Interaction	1794.005	1	1794.005	7.409481	0.00663	3.853168
28	Within	192729.8	796	242.1229			
29							
30	Total	407985.9	799				

Minitab

**Two-way ANOVA: hr versus group, gender**

Source	DF	SS	MS	F	P
group	1	168432	168432	695.65	0.000
gender	1	45030	45030	185.98	0.000
Interaction	1	1794	1794	7.41	0.007
Error	796	192730	242		
Total	799	407986			

S = 15.56 R-Sq = 52.76% R-Sq(adj) = 52.58%

**FIGURE 13.7**

Excel and Minitab two-way ANOVA output for the heart rate study, for Example 13.10.

## CHAPTER 13 Summary

**Two-way analysis of variance (ANOVA)** is used to compare population means when populations are classified according to two factors.

We assume that independent SRSs are drawn from each population and that the responses from each population are Normal with possibly different means but the same standard deviation.

As with one-way ANOVA, these assumptions should be assessed. Preliminary analysis includes examination of means, standard deviations, and Normal quantile plots.

**Marginal means** are calculated by taking averages of the cell means, either across rows or down columns. These means can be used in an interaction plot to aid in the interpretation of results.

Similar to one-way ANOVA, the total variation is separated into parts for the **model** and **error**. Pooling is also used to estimate the error, or within-group variance. However, given that there are now two factors, the model variation is separated into parts for each of the **main effects** and the **interaction**.

The calculations are organized into an **ANOVA table**.  $F$  statistics and  $P$ -values are used to test hypotheses about the main effects and the interaction.

Careful inspection of the means is necessary to interpret significant main effects and interactions. Plots are a useful aid.

## CHAPTER 13 Exercises

For Exercises 13.1 and 13.2, see page 702.

### 13.3 What's wrong?

For each of the following, explain what is wrong and why.

- (a) You should reject the null hypothesis that there is no interaction in a two-way ANOVA when the AB  $F$  statistic is small.
- (b) Sums of squares are equal to mean squares divided by degrees of freedom.
- (c) The test statistics for the main effects in a two-way ANOVA have a chi-square distribution when the null hypothesis is true.
- (d) The sums of squares always add in two-way ANOVA.

### 13.4 Is there an interaction?

Each of the following tables gives means for a two-way ANOVA. Make a plot of the means with the levels of Factor A on the  $x$  axis. State whether or not there is an interaction, and if there is, describe it.

(a)

Factor B	Factor A		
	1	2	3
1	11	18	21
2	6	13	16

(b)

Factor B	Factor A		
	1	2	3
1	10	25	15
2	20	35	25

(c)

Factor B	Factor A		
	1	2	3
1	10	15	20
2	15	25	35

(d)

Factor B	Factor A		
	1	2	3
1	10	15	12
2	50	52	55

### 13.5 Describing a two-way ANOVA model.

A  $3 \times 2$  ANOVA was run with 6 observations per cell.

- Give the degrees of freedom for the  $F$  statistic that is used to test for interaction in this analysis and the entries from Table E that correspond to this distribution.
- Sketch a picture of this distribution with the information from the table included.
- The calculated value of this  $F$  statistic is 2.23. Report the  $P$ -value and state your conclusion.
- Based on your answer to part (c), would you expect an interaction plot to have mean profiles that look parallel? Explain your answer.

### 13.6 Determining the critical value of $F$ .

For each of the following situations, state how large the  $F$  statistic needs to be for rejection of the null hypothesis at the 5% level. Sketch each distribution and indicate the region where you would reject.

- The main effect for the first factor in a  $3 \times 4$  ANOVA with 3 observations per cell
- The interaction in a  $3 \times 4$  ANOVA with 6 observations per cell
- The interaction in a  $2 \times 2$  ANOVA with 26 observations per cell

### 13.7 Identifying the factors of a two-way ANOVA model.

For each of the following situations, identify both factors and the response variable. Also, state the number of levels for each factor ( $I$  and  $J$ ) and the total number of observations ( $N$ ).

- A child psychologist is interested in studying how a child's percent of pretend play differs with gender and age (4, 8, and 12 months). There are 11 infants assigned to each cell of the experiment.
- Brewers malt is produced from germinating barley. A homebrewer wants to determine the best conditions for germinating barley. Thirty lots of barley seed were equally and randomly assigned to 10 germination conditions. The conditions are combinations of the week after harvest (1, 3, 6, 9, or 12 weeks) and the amount of water used in the process (4 or 8 milliliters). The percent of seeds germinating is the outcome variable.
- The strength of concrete depends upon the formula used to prepare it. An experiment compares six different mixtures. Nine specimens of concrete are poured from each mixture. Three of these specimens are subjected to 0 cycles of freezing and thawing, three are subjected to 100 cycles, and three are subjected to 500 cycles. The strength of each specimen is then measured.
- A company wants to compare four different training programs for its new employees. Each of these programs takes 6 hours to complete. The training can be given for 6 hours on one day or for 3 hours on two consecutive days. The next 80 employees hired by the company will be the subjects for this study.

### 13.8 Determining the degrees of freedom.

For each part in Exercise 13.7, outline the ANOVA table, giving the sources of variation and the degrees of freedom.

### 13.9 Writing about testing worries and exam performance.

For many students, self-induced worries and pressure to perform well on exams cause them to perform below their ability. This is because these worries compete with the working memory available for performance. Expressive writing has been shown to be an effective technique to overcome traumatic or emotional experiences. Thus, a group of researchers decided to investigate whether expressive writing prior to test-taking may help performance.<sup>8</sup>

The small study involved 20 subjects. Half the subjects were assigned to the expressive-writing group and the others to a control group. Each subject took two short mathematics exams. Prior to the first exam, students were told just to perform their best. Prior to the second exam, students were told that they each had been paired with another student and if the members of a pair both performed well on the exam, the pair would receive a monetary reward. Each student was then told privately that his or her partner had already scored well. This was done to create a high-stakes testing environment for the second exam. Those in the control group sat quietly for 10 minutes prior to taking the second exam. Those in the expressive-writing group had 10 minutes to write about their thoughts and feelings regarding the exam. The following table summarizes the test results (% correct).

Group	First Exam		Second Exam	
	$\bar{x}$	$s$	$\bar{x}$	$s$
Control	83.4	11.5	70.1	14.3
Expressive-writing	86.2	6.3	90.1	5.8

- Explain why this is a repeated-measures design and not a standard two-way ANOVA design.
- Generate a plot to look at changes in score across time and across group. Describe what you see in terms of the main effects and interaction.
- Because exam scores can run only between 0% and 100%, variances for populations with means near 0% or 100% may be smaller and the distribution of scores may be skewed. Does it appear reasonable here to pool variances? Explain your answer.

### 13.10 Influence of age and gender on motor performance.

The slowing of motor performance as humans age is well established. Differences in gender, however, are less so. A recent study assessed the motor performance of 246 healthy adults.<sup>9</sup> One task was to tap the thumb and forefinger of the right hand together 20 times as quickly as possible. The following table summarizes the results (in seconds) for 7 age classes and 2 genders.

Age class (years)	Males			Females		
	$n$	$\bar{x}$	$s$	$n$	$\bar{x}$	$s$
41–50	19	4.72	1.31	19	5.88	0.82
41–50	19	4.72	1.31	19	5.88	0.82
51–55	12	4.10	1.62	12	5.93	1.13
56–60	12	4.80	1.04	12	5.85	0.87
61–65	24	5.08	0.98	24	5.81	0.94

66–70	17	5.47	0.85	17	6.50	1.23
71–75	23	5.84	1.44	23	6.12	1.04
>75	16	5.86	1.00	16	6.19	0.91

Generate a plot to look at changes in the time across age class and across gender. Describe what you see in terms of the main effects for age and gender as well as their interaction.

### 13.11 Influence of age and gender on motor performance, continued.

Refer to the previous exercise.

(a) In their article, the researchers state that each of their response variables was assessed for Normality prior to performing a two-way ANOVA. Is it necessary for the 246 time measurements to be Normally distributed? Explain your answer.

(b) Is it reasonable to pool the variances?

(c) Suppose for these data that  $SS(\text{gender}) = 44.66$ ,  $SS(\text{age}) = 31.97$ ,  $SS(\text{interaction}) = 13.22$ , and  $SSE = 280.95$ . Construct an ANOVA table and state your conclusions.

### 13.12 Fuzzy fish?

Drugs used to treat anxiety persist in wastewater effluent, resulting in relatively high concentrations of these drugs in our rivers and streams. To better understand the effects of these drugs on fish, researchers commonly expose fish to various levels of an anxiety drug in a laboratory setting and observe their behavior. In one study, researchers considered the effects of three doses of oxazepam on the behavior of the European perch.<sup>10</sup> Twenty-five perch were each assigned to doses of 0, 1.8, or 910 micrograms per liter of water ( $\mu\text{g/l}$ ). Each fish was first observed prior to treatment and then observed 7 days after treatment. The following table summarizes the results for activity (number of swimming bouts greater than 0.25 cm during 10 minutes).

Dose ( $\mu\text{g/l}$ )	Number of Movements			
	Pretreatment		Posttreatment	
	$\bar{x}$	$s$	$\bar{x}$	$s$
0	3.92	2.38	3.68	1.80
1.8	3.76	1.94	6.32	2.01
910	4.08	1.58	8.68	3.05

(a) The response is the number of movements in 10 minutes, so this variable takes only integer values. Should we be concerned about violating the assumption of Normality? Explain your answer.

(b) Often with this type of count, one considers taking the square root of the count and performing ANOVA on the transformed response. Explain why a transformation might be used here.

(c) Construct an interaction plot and comment on the main effects of dose and time and their interaction.

### 13.13 The influences of transaction history and a thank-you statement.

A service failure is defined as any service-related problem (real or perceived) that transpires during a customer's experience with a firm. In the hotel industry, there is a high human component, so these sorts of failures commonly occur regardless of extensive training and established policies. As a

result, hotel firms must learn to effectively react to these failures. A recent study investigated the relationship between a consumer's transaction history (levels: long and short) and an employee thank-you statement (levels: yes and no) on a consumer's repurchase intent.<sup>11</sup> Each subject was randomly assigned to one of the four treatment groups and asked to read some service failure/resolution scenarios and respond accordingly. Repurchase intent was measured using a 9-point scale. Here is a summary of the means:

History	Thank-you	
	No	Yes
Short	5.69	6.80
Long	7.53	7.37

(a) Plot the means. Do you think there is an interaction? If yes, describe the interaction in terms of the two factors.

(b) Find the marginal means. Are they useful for understanding the results of this study? Explain your answer.

### 13.14 Transaction history and a thank-you statement, continued.

Refer to the previous exercise. The numbers of subjects in the cells were not equal so the researchers used linear regression to analyze the data. This was done by creating an indicator variable for each factor and the interaction. Below is a partial ANOVA table. Complete it and state your conclusions regarding the main effects and interaction described in the previous exercise.

Source	DF	SS	MS	F	P-value
Transaction history		61.445			
Thank-you statement		21.810			
Interaction		15.404			
Error	160	759.904			

### 13.15 The effects of proximity and visibility on food intake.

A study investigated the influence that proximity and visibility of food have on food intake.<sup>12</sup> A total of 40 secretaries from the University of Illinois participated in the study. A candy dish full of individually wrapped chocolates was placed either at the desk of the participant or at a location 2 meters from the participant. The candy dish was either a clear (candy visible) or opaque (candy not visible) covered bowl. After a week, the researchers noted not only the number of candies consumed per day but also the self-reported number of candies consumed by each participant. The following table summarizes the mean difference between these two values (reported minus actual).

Proximity	Visibility	
	Clear	Opaque
Proximate	-1.2	-0.8
Less proximate	0.5	0.4

(a) Make a plot of the means and describe the patterns that you see. Does the plot suggest an interaction between visibility and proximity?

(b) This study actually took four weeks, with each participant being observed at each treatment combination in a random order. Explain why a repeated-measures design like this may be beneficial.

### 13.16 Bilingualism.

Not only does speaking two languages have many practical benefits in this globalized world, but there is also growing evidence that it appears to help with brain functioning as we age. In one study, 80 participants were divided equally among 4 groups: younger adult bilinguals, older adult bilinguals, younger adult monolinguals, and older adult monolinguals.<sup>13</sup> Each participant was asked to complete a series of color-shape task-switching tests. For our analysis, we'll focus on the total reaction time (in microseconds) for these experiments. The shorter the reaction time, the better. 

#### BILING

- (a) Make a table giving the sample size, mean, and standard deviation for each group. Is it reasonable to pool the variances?
- (b) Generate a histogram for each of the groups. Can we feel confident that the sample means are approximately Normal? Explain your answer.

### 13.17 Bilingualism, continued.

Refer to the previous exercise.  BILING

- (a) If bilingualism helps with brain functioning as we age, explain why we'd expect to find an interaction between age and bilingualism. Also, create an interaction plot of what sort of pattern we'd expect.
- (b) Analyze the reaction times using analysis of variance. Report the test statistics, degrees of freedom, and  $P$ -values.
- (c) Based on part (b), write a short paragraph summarizing your findings.

### 13.18 Hypotension and endurance exercise.

In sedentary individuals, low blood pressure (hypotension) often occurs after a single bout of aerobic exercise and lasts nearly two hours. This can cause dizziness, light-headedness, and possibly fainting upon standing. It is thought that endurance exercise training can reduce the degree of postexercise hypotension. To test this, researchers studied 16 endurance-trained and 16 sedentary men and women.<sup>14</sup> The following table summarizes the postexercise systolic arterial pressure (mm Hg) after 60 minutes of upright cycling.

Group	$n$	$\bar{x}$	SE
Women, sedentary	8	100.7	3.4
Women, endurance	8	105.3	3.6
Men, sedentary	8	114.2	3.8
Men, endurance	8	110.2	2.3

- (a) Make a plot similar to Figure 13.3 (page 701) with the systolic blood pressure on the  $y$  axis and training level on the  $x$  axis. Describe the pattern you see.
- (b) From the table, one can show that  $SSA = 677.12$ ,  $SSB = 0.72$ ,  $SSAB = 147.92$ , and  $SSE = 2478$  where A is the gender effect and B is the training level. Construct the ANOVA table with  $F$  statistics and degrees of freedom, and state your conclusions regarding main effects and interaction.
- (c) The researchers also measured the before-exercise systolic blood pressure of the participants and looked at a model that incorporated both the pre- and postexercise values. Explain why it is likely to

be beneficial to incorporate both measurements in the study.

### 13.19 The effect of humor.

In advertising, humor is often used to overcome sales resistance and stimulate customer purchase behavior. One experiment looked at the use of humor to offset the negative feelings often associated with website encounters.<sup>15</sup> The setting of the experiment was an online travel agency, and the researchers used a three-factor design, each factor with two levels. The factors were humor (used, not used), process (favorable, unfavorable), and outcome (favorable, unfavorable). For the humor condition, cartoons and jokes of the day about skiing were presented on the site. For the no humor condition, standard pictures of ski sites were used. Two hundred and forty-one business students from a large Dutch university participated in the experiment. Each was randomly assigned to one of the eight treatment conditions. The students were asked to book a skiing holiday and then rate their perceived enjoyment and satisfaction with the process. All responses were measured on a 7-point Likert scale. A summary of the results for satisfaction follows.

Treatment	<i>n</i>	$\bar{x}$	<i>s</i>
No humor—favorable process—unfavorable outcome	27	3.04	0.79
No humor—favorable process—favorable outcome	29	5.36	0.47
No humor—unfavorable process—unfavorable outcome	26	2.84	0.59
No humor—unfavorable process—favorable outcome	31	3.08	0.59
Humor—favorable process—unfavorable outcome	32	5.06	0.59
Humor—favorable process—favorable outcome	30	5.55	0.65
Humor—unfavorable process—unfavorable outcome	36	1.95	0.52
Humor—unfavorable process—favorable outcome	30	3.27	0.71

(a) Plot the means of the four treatments without humor. Do you think there is an interaction? If yes, describe the interaction in terms of the process and outcome factors.

(b) Plot the means of the four treatments that used humor. Do you think there is an interaction? If yes, describe the interaction in terms of the process and outcome factors.

(c) The three-factor interaction can be assessed by looking at the two interaction plots created in parts (a) and (b). If the relationship between process and outcome is different across the two humor conditions, there is evidence of an interaction among all three factors. Do you think there is a three-factor interaction? Explain your answer.

### 13.20 Pooling the standard deviations.

Refer to the previous exercise. Find the pooled estimate of the standard deviation for these data. What are its degrees of freedom? Using the rule from Chapter 12 (page 654), is it reasonable to use a pooled standard deviation for the analysis? Explain your answer.

### 13.21 Describing the effects.

Refer to Exercise 13.19. The *P*-values for all main effects and two-factor interactions are significant at the 0.05 level. Using the table, find the marginal means (that is, the mean for the no humor treatment, the mean for the no humor and unfavorable process treatment combination, etc.) and use them to describe these effects.

### 13.22 Acceptance of functional foods.

Functional foods are foods that are fortified with health-promoting supplements, like calcium-enriched orange juice or vitamin-enriched cereal. Although the number of functional foods is growing in the marketplace, very little is known about how the next generation of consumers views these foods. Because of this, a questionnaire was given to college students from the United States, Canada, and France.<sup>16</sup> This questionnaire measured the students' attitudes and beliefs about general food and functional food. One of the response variables collected concerned cooking enjoyment. This variable was the average of numerous items, each measured on a 10-point scale, where 1 = most negative value and 10 = most positive value. Here are the means:

Gender	Culture		
	Canada	United States	France
Female	7.70	7.36	6.38
Male	6.39	6.43	5.69

- (a) Make a plot of the means and describe the patterns that you see.
- (b) Does the plot suggest that there is an interaction between culture and gender? If your answer is Yes, describe the interaction.

### 13.23 Estimating the within-group variance.

Refer to the previous exercise. Here are the cell standard deviations and sample sizes for cooking enjoyment:

Gender	Culture					
	Canada		United States		France	
	<i>s</i>	<i>n</i>	<i>s</i>	<i>n</i>	<i>s</i>	<i>n</i>
Female	1.668	238	1.736	178	2.024	82
Male	1.909	125	1.601	101	1.875	87

Find the pooled estimate of the standard deviation for these data. Use the rule for examining standard deviations in ANOVA from Chapter 12 (page 654) to determine if it is reasonable to use a pooled standard deviation for the analysis of these data.

### 13.24 Comparing the groups.

Refer to Exercises 13.22 and 13.23. The researchers presented a table of means with different superscripts indicating pairs of means that differed at the 0.05 significance level, using the Bonferroni method.

- (a) What denominator degrees of freedom would be used here?
- (b) How many pairwise comparisons are there for this problem?
- (c) Perform these comparisons using  $t^{**} = 2.94$  and summarize your results.

### 13.25 More on acceptance of functional foods.

Refer to Exercise 13.22. The means for four of the response variables associated with functional foods are as follows.

General Attitude	Product Benefits
------------------	------------------

Gender	Credibility of Information			Purchase Intention		
	Canada	United States	France	Canada	United States	France
Female	4.93	4.69	4.10	4.59	4.37	3.91
Male	4.50	4.43	4.02	4.20	4.09	3.87

Gender	Credibility of Information			Purchase Intention		
	Canada	United States	France	Canada	United States	France
Female	4.54	4.50	3.76	4.29	4.39	3.30
Male	4.23	3.99	3.83	4.11	3.86	3.41

For each of the four response variables, give a graphical summary of the means. Use this summary to discuss any interactions that are evident. Write a short report summarizing any differences in culture and gender with respect to the response variables measured.

### 13.26 Interpreting the results.

The goal of the study in the previous exercise was to understand cultural and gender differences in functional food attitudes and behaviors among young adults, the next generation of food consumers. The researchers used a sample of undergraduate students and had each participant fill out the survey during class time. How reasonable is it to generalize these results to the young adult population in these countries? Explain your answer.

### 13.27 Evaluation of an intervention program.

The National Crime Victimization Survey estimates that there were over 400,000 violent crimes committed against women by their intimate partner that resulted in physical injury. An intervention study designed to increase safety behaviors of abused women compared the effectiveness of six telephone intervention sessions with a control group of abused women who received standard care. Fifteen different safety behaviors were examined.<sup>17</sup> One of the variables analyzed was the total number of behaviors (out of 15) that each woman performed. Here is a summary of the means of this variable at baseline (just before the first telephone call) and at follow-up three and six months later:

Group	Time		
	Baseline	3 months	6 months
Intervention	10.4	12.5	11.9
Control	9.6	9.9	10.4

- (a) Find the marginal means. Are they useful for understanding the results of this study?
- (b) Plot the means. Do you think there is an interaction? Describe the meaning of an interaction for this study.  
*(Note: This exercise is from a repeated-measures design, and the data are not particularly Normal because they are counts with values from 1 to 15. Although we cannot use the methods in this chapter for statistical inference in this setting, the example does illustrate ideas about interactions.)*

### 13.28 More on the evaluation of an intervention program.

Refer to the previous exercise. Table 13.1 gives the percents of women who responded that they performed each of the 15 safety behaviors studied.

TABLE 13.1

Safety Behaviors of Abused Women

Behavior	Intervention Group (%)			Control Group (%)		
	Baseline	3 months	6 months	Baseline	3 months	6 months
Hide money	68.0	60.0	62.7	60.0	37.8	35.1
Hide extra keys	52.7	76.0	68.9	53.3	33.8	39.2
Abuse code to alert family	30.7	74.7	60.0	22.7	27.0	43.2
Hide extra clothing	37.3	73.6	52.7	42.7	32.9	27.0
Ask neighbors to call police	49.3	73.0	66.2	32.0	45.9	40.5
Know Social Security number	93.2	93.2	100.0	89.3	93.2	98.6
Keep rent, utility receipts	75.3	95.5	89.4	70.3	84.7	80.9
Keep birth certificates	84.0	90.7	93.3	77.3	90.4	93.2
Keep driver's license	93.3	93.3	97.3	94.7	95.9	98.6
Keep telephone numbers	96.0	98.7	100.0	90.7	97.3	100.0
Remove weapons	50.0	70.6	38.5	40.7	23.8	5.9
Keep bank account numbers	81.0	94.3	96.2	76.2	85.5	94.4
Keep insurance policy number	70.9	90.4	89.7	68.3	84.2	94.8
Keep marriage license	71.1	92.3	84.6	63.3	73.2	80.0
Hide valuable jewelry	78.7	84.5	83.9	74.0	75.0	80.3

(a) Summarize these data graphically. Do you think that your graphical display is more effective than Table 13.1 for describing the results of this study? Explain why or why not.

(b) Note any particular patterns in the data that would be important to someone who wants to use these results to design future intervention programs for abused women.

(c) The study was conducted “at a family violence unit of a large urban District Attorney’s Office that serves an ethnically diverse population of three million citizens.” To what extent do you think that this fact limits the conclusions that can be drawn?

### 13.29 What can you conclude?

Analysis of data for a  $3 \times 2$  ANOVA with 6 observations per cell gave the  $F$  statistics in the following table.

Effect	$F$
A	3.45
B	2.49
AB	1.14

What can you conclude from the information given?

### 13.30 What can you conclude?

A study reported the following results for data analyzed using the methods that we studied in this chapter.

Effect	<i>F</i>	<i>P</i> -value
A	0.50	0.609
B	10.06	0.001
AB	4.48	0.003

- (a) What can you conclude from the information given?  
 (b) What additional information would you need to write a summary of the results for this study?

### 13.31 Conspicuous consumption and men’s testosterone levels.

It is argued that conspicuous consumption is a means by which men communicate their social status to prospective mates. One study looked at changes in a male’s testosterone level in response to fluctuations in his status created by the consumption of a product.<sup>18</sup> The products considered were a new and luxurious sports car and an old family sedan. Participants were asked to drive on either an isolated highway or a busy city street. A table of cell means and standard deviations for the change (post – pre) in testosterone level follows.

Car	Location			
	Highway		City	
	$\bar{x}$	<i>s</i>	$\bar{x}$	<i>s</i>
Old sedan	0.03	0.12	–0.03	0.12
New sports car	0.15	0.14	0.13	0.13

- (a) Make a plot of the means and describe the patterns that you see. Does the plot suggest an interaction between location and type of car?  
 (b) Compute the pooled standard error  $s_p$ , assuming equal sample sizes.  
 (c) The researchers wanted to test the following hypotheses:
- (1) Testosterone levels will increase more in men who drive the new car.
  - (2) For men driving the new car, testosterone levels will increase more in men who drive in the city.
  - (3) For men driving the old car, testosterone levels will decrease less in men who drive the old car on the highway.

Write out the contrasts for each of these hypotheses.

- (d) This study actually involved each male participating in all four combinations. Half of them drove the sedan first and the other half drove the sports car first. Explain why a repeated-measures design like this may be beneficial.

### 13.32 The effects of peer pressure on mathematics achievement.

Researchers were interested in comparing the relationship between high achievement in mathematics and peer pressure across several countries.<sup>19</sup> They hypothesized that in countries where high achievement is not valued highly, considerable peer pressure may exist. A questionnaire was

distributed to 14-year-olds from three countries (Germany, Canada, and Israel). One of the questions asked students to rate how often they fear being called a nerd or teacher's pet on a 4-point scale (1 = never, 4 = frequently). The following table summarizes the response.

Country	Gender	$n$	$\bar{x}$
Germany	Female	336	1.62
Germany	Male	305	1.39
Israel	Female	205	1.87
Israel	Male	214	1.63
Canada	Female	301	1.91
Canada	Male	304	1.88

- (a) The  $P$ -values for the interaction and the main effects for country and for gender are 0.016, 0.068, and 0.108, respectively. Using the table and  $P$ -values, summarize the results both graphically and numerically.
- (b) The researchers contend that Germany does not value achievement as highly as Canada and Israel. Do the results from part (a) allow you to address their primary hypothesis? Explain.
- (c) The students were also asked to indicate their current grade in mathematics on a 6-point scale (1 = excellent, 6 = insufficient). How might both responses be used to address the researchers' primary hypothesis?

### 13.33 The effect of chromium on insulin metabolism.

The amount of chromium in the diet has an effect on the way the body processes insulin. In an experiment designed to study this phenomenon, four diets were fed to male rats. There were two factors. Chromium had two levels: low (L) and normal (N). The rats were allowed to eat as much as they wanted (M), or the total amount that they could eat was restricted (R). We call the second factor Eat. One of the variables measured was the amount of an enzyme called GITH.<sup>20</sup> The means for this response variable appear in the following table.

Chromium	Eat	
	M	R
L	4.545	5.175
N	4.425	5.317

- (a) Make a plot of the mean GITH for these diets, with the factor Chromium on the  $x$  axis and GITH on the  $y$  axis. For each Eat group, connect the points for the two Chromium means.
- (b) Describe the patterns you see. Does the amount of chromium in the diet appear to affect the GITH mean? Does restricting the diet rather than letting the rats eat as much as they want appear to have an effect? Is there an interaction?
- (c) Compute the marginal means. Compute the differences between the M and R diets for each level of Chromium. Use this information to summarize numerically the patterns in the plot.

### 13.34 Use of animated agents in a multimedia environment.

Multimedia learning environments are designed to enhance learning by providing a more hands-on and exploratory investigation of a topic. Often animated agents (human-like characters) are used with the hope of enhancing social interaction with the software and thus improving learning. One group of researchers decided to investigate whether the presence of an agent and the type of verbal

feedback provided were actually helpful.<sup>21</sup> To do this, they recruited 135 college students and randomly divided them between 4 groups: agent/simple feedback, agent/elaborate feedback, no agent/simple feedback, and no agent/elaborate feedback. The topic of the software was thermodynamics. The change in score on a 20-question test taken before and after using the software was the response.  **AGENT**

- Make a table giving the sample size, mean, and standard deviation for each group.
- Use these means to construct an interaction plot. Describe the main effects for agent presence and for feedback type as well as their interaction.
- Analyze the change in score using analysis of variance. Report the test statistics, degrees of freedom, and  $P$ -values.
- Use the residuals to check model assumptions. Are there any concerns? Explain your answer.
- Based on parts (b) and (c), write a short paragraph summarizing your findings.

### 13.35 Trust of individuals and groups.

Trust is an essential element in any exchange of goods or services. The following trust game is often used to study trust experimentally:

*A sender starts with  $\$X$  and can transfer any amount  $x \leq X$  to a responder. The responder then gets  $\$3x$  and can transfer any amount  $y \leq 3x$  back to the sender. The game ends with final amounts  $X - x + y$  and  $3x - y$  for the sender and responder, respectively.*

The value  $x$  is taken as a measure of the sender's trust, and the value  $y/3x$  indicates the responder's trustworthiness. A recent study used this game to study the dynamics between individuals and groups of three.<sup>22</sup> The following table summarizes the average amount  $x$  sent by senders starting with  $\$100$ .

Sender	Responder	$n$	$\bar{x}$	$s$
Individual	Individual	32	65.5	36.4
Individual	Group	25	76.3	31.2
Group	Individual	25	54.0	41.6
Group	Group	27	43.7	42.4

- Find the pooled estimate of the standard deviation for this study and its degrees of freedom.
- Is it reasonable to use a pooled standard deviation for the analysis? Explain your answer.
- Compute the marginal means.
- Plot the means. Do you think there is an interaction? If yes, describe it.
- The  $F$  statistics for sender, responder, and interaction are 9.05, 0.001, and 2.08, respectively. Compute the  $P$ -values and state your conclusions.

### 13.36 Does the type of cooking pot affect iron content?

Iron-deficiency anemia is the most common form of malnutrition in developing countries, affecting

about 50% of children and women and 25% of men. Iron pots for cooking foods had traditionally been used in many of these countries, but they have been largely replaced by aluminum pots, which are cheaper and lighter. Some research has suggested that food cooked in iron pots will contain more iron than food cooked in other types of pots. One study designed to investigate this issue compared the iron content of some Ethiopian foods cooked in aluminum, clay, and iron pots.<sup>23</sup> Foods considered were *yesiga wet*, beef cut into small pieces and prepared with several Ethiopian spices; *shiro wet*, a legume-based mixture of chickpea flour and Ethiopian spiced pepper; and *ye-atkilt allych'a*, a lightly spiced vegetable casserole. Four samples of each food were cooked in each type of pot. The iron in the food is measured in milligrams of iron per 100 grams of cooked food. The data are shown in Table 13.2. 

**TABLE 13.2** Iron Content (mg/100 g) of Food Cooked in Different Pots

Type of pot	Meat				Legumes				Vegetables			
Aluminum	1.77	2.36	1.96	2.14	2.40	2.17	2.41	2.34	1.03	1.53	1.07	1.30
Clay	2.27	1.28	2.48	2.68	2.41	2.43	2.57	2.48	1.55	0.79	1.68	1.82
Iron	5.27	5.17	4.06	4.22	3.69	3.43	3.84	3.72	2.45	2.99	2.80	2.92

- (a) Make a table giving the sample size, mean, and standard deviation for each type of pot. Is it reasonable to pool the variances? Although the standard deviations vary more than we would like, this is partially due to the small sample sizes, and we will proceed with the analysis of variance.
- (b) Plot the means. Give a short summary of how the iron content of foods depends upon the cooking pot.
- (c) Run the analysis of variance. Give the ANOVA table, the  $F$  statistics with degrees of freedom and  $P$ -values, and your conclusions regarding the hypotheses about main effects and interactions.

### 13.37 Interpreting the results.

Refer to the previous exercise. Although there is a statistically significant interaction, do you think that these data support the conclusion that foods cooked in iron pots contain more iron than foods cooked in aluminum or clay pots? Discuss.

### 13.38 Analysis using a one-way ANOVA.

Refer to Exercise 13.36. Rerun the analysis as a one-way ANOVA with 9 groups and 4 observations per group. Report the results of the  $F$  test. Examine differences in means using a multiple-comparisons procedure. Summarize your results and compare them with those you obtained in Exercise 13.36.

### 13.39 Examination of a drilling process.

One step in the manufacture of large engines requires that holes of very precise dimensions be drilled. The tools that do the drilling are regularly examined and are adjusted to ensure that the holes meet the required specifications. Part of the examination involves measurement of the diameter of the drilling tool. A team studying the variation in the sizes of the drilled holes selected this measurement procedure as a possible cause of variation in the drilled holes. They decided to use a designed experiment as one part of this examination. Some of the data are given in Table 13.3. The

diameters in millimeters (mm) of five tools were measured by the same operator at three times (8:00 A.M., 11:00 A.M., and 3:00 P.M.). Three measurements were taken on each tool at each time. The person taking the measurements could not tell which tool was being measured, and the measurements were taken in random order.<sup>24</sup>  **DRILL**

- (a) Make a table of means and standard deviations for each of the  $5 \times 3$  combinations of the two factors.
- (b) Plot the means and describe how the means vary with tool and time. Note that we expect the tools to have slightly different diameters. These will be adjusted as needed. It is the process of measuring the diameters that is important.
- (c) Use a two-way ANOVA to analyze these data. Report the test statistics, degrees of freedom, and *P*-values for the significance tests.
- (d) Write a short report summarizing your results.

Tool	Time	Diameter (mm)		
1	1	25.030	25.030	25.032
1	2	25.028	25.028	25.028
1	3	25.026	25.026	25.026
2	1	25.016	25.018	25.016
2	2	25.022	25.020	25.018
2	3	25.016	25.016	25.016
3	1	25.005	25.008	25.006
3	2	25.012	25.012	25.014
3	3	25.010	25.010	25.008
4	1	25.012	25.012	25.012
4	2	25.018	25.020	25.020
4	3	25.010	25.014	25.018
5	1	24.996	24.998	24.998
5	2	25.006	25.006	25.006
5	3	25.000	25.002	24.999

### 13.40 Examination of a drilling process, continued.

Refer to the previous exercise. Multiply each measurement by 0.04 to convert from millimeters to inches. Redo the plots and rerun the ANOVA using the transformed measurements. Summarize what parts of the analysis have changed and what parts have remained the same.  **DRILL**

### 13.41 Do left-handed people live shorter lives than right-handed people?

A study of this question examined a sample of 949 death records and contacted next of kin to determine handedness.<sup>25</sup> Note that there are many possible definitions of “left-handed.” The researchers examined the effects of different definitions on the results of their analysis and found that their conclusions were not sensitive to the exact definition used. For the results presented here,

people were defined to be right-handed if they wrote, drew, and threw a ball with the right hand. All others were defined to be left-handed. People were classified by gender (female or male) and handedness (left or right), and a  $2 \times 2$  ANOVA was run with the age at death as the response variable. The  $F$  statistics were 22.36 (handedness), 37.44 (gender), and 2.10 (interaction). The following marginal mean ages at death (in years) were reported: 77.39 (females), 71.32 (males), 75.00 (right-handed), and 66.03 (left-handed).

(a) For each of the  $F$  statistics given, find the degrees of freedom and an approximate  $P$ -value. Summarize the results of these tests.

(b) Using the information given, write a short summary of the results of the study.

### 13.42 A radon exposure study.

Scientists believe that exposure to the radioactive gas radon is associated with some types of cancers in the respiratory system. Radon from natural sources is present in many homes in the United States. A group of researchers decided to study the problem in dogs because dogs get similar types of cancers and are exposed to environments similar to those of their owners. Radon detectors are available for home monitoring, but the researchers wanted to obtain actual measures of the exposure of a sample of dogs. To do this, they placed the detectors in holders and attached them to the collars of the dogs. One problem was that the holders might in some way affect the radon readings. The researchers therefore devised a laboratory experiment to study the effects of the holders. Detectors from four series of production were available, so they used a two-way ANOVA design (series with 4 levels and holder with 2, representing the presence or absence of a holder). All detectors were exposed to the same radon source and the radon measure in picocuries per liter was recorded.<sup>26</sup> The  $F$  statistic for the effect of series is 7.02, for holder it is 1.96, for the interaction it is 1.24, and  $N = 69$ .

(a) Using Table E or statistical software, find approximate  $P$ -values for the three test statistics. Summarize the results of these three significance tests.

(b) The mean radon readings for the four series were 330, 303, 302, and 295. The results of the significance test for series were of great concern to the researchers. Explain why.

### 13.43 A comparison of plant species under low water conditions.

The PLANTS1 data file gives the percent of nitrogen in four different species of plants grown in a laboratory. The species are *Leucaena leucocephala*, *Acacia saligna*, *Prosopis juliflora*, and *Eucalyptus citriodora*. The researchers who collected these data were interested in commercially growing these plants in parts of the country of Jordan where there is very little rainfall. To examine the effect of water, they varied the amount per day from 50 millimeters (mm) to 650 mm in 100 mm increments. There were 9 plants per species-by-water combination. Because the plants are to be used primarily for animal food, with some parts that can be consumed by people, a high nitrogen content is very desirable.  **PLANTS1**

(a) Find the means for each species-by-water combination. Plot these means versus water for the four species, connecting the means for each species by lines. Describe the overall pattern.

(b) Find the standard deviations for each species-by-water combination. Is it reasonable to pool the standard deviations for this problem? Note that with sample sizes of size 9, we expect these standard deviations to be quite variable.

(c) Run the two-way analysis of variance. Give the results of the hypothesis tests for the main effects and the interaction.

### 13.44 Examination of the residuals.

Refer to the previous exercise. Examine the residuals. Are there any unusual patterns or outliers? If you think that there are one or more points that are somewhat extreme, rerun the two-way analysis without these observations. Does this change the results in any substantial way?  PLANTS1

### 13.45 Analysis using multiple one-way ANOVAs.

Refer to Exercise 13.43. Run a separate one-way analysis of variance for each water level. If there is evidence that the species are not all the same, use a multiple-comparisons procedure to determine which pairs of species are significantly different. In what way, if any, do the differences appear to vary by water level? Write a short summary of your conclusions.  PLANTS1

### 13.46 More on the analysis using multiple one-way ANOVAs.

Refer to Exercise 13.43. Run a separate oneway analysis of variance for each species and summarize the results. Since the amount of water is a quantitative factor, we can also analyze these data using regression. Run simple linear regressions separately for each species to predict nitrogen percent from water. Use plots to determine whether or not a line is a good way to approximate this relationship.

Summarize the regression results and compare them with the one-way ANOVA results.  PLANTS1

### 13.47 Another comparison of plant species under low water conditions.

Refer to Exercise 13.43. Additional data collected by the same researchers according to a similar design are given in the PLANTS2 data file. Here, there are two response variables. They are fresh biomass and dry biomass. High values for both of these variables are desirable. The same four species and seven levels of water are used for this experiment. Here, however, there are four plants per species-by-water combination. Analyze each of the response variables in the PLANTS2 data file using the outline from Exercise 13.43.  PLANTS2

### 13.48 Examination of the residuals.

Perform the tasks described in Exercise 13.44 for the two response variables in the PLANTS2 data file.  PLANTS2

### 13.49 Analysis using multiple one-way ANOVAs.

Perform the tasks described in Exercise 13.45 for the two response variables in the PLANTS2 data file.  PLANTS2

### 13.50 More on the analysis using multiple one-way ANOVAs.

Perform the tasks described in Exercise 13.46 for the two response variables in the PLANTS2 data file.  PLANTS2

### 13.51 Are insects more attracted to male plants?

Some scientists wanted to determine if there are gender-related differences in the level of herbivory for the jack-in-the-pulpit, a spring-blooming perennial plant common in deciduous forests. A study was conducted in southern Maryland in forests associated with the Smithsonian Environmental Research Center (SERC).<sup>27</sup> To determine the effects of flowering and floral characteristics on herbivory, the researchers altered the floral morphology of male and female plants. The three levels of floral characteristics were (1) the spathes were completely removed; (2) in females, a gap was created in the base of the spathe, and in males, the gap was closed; (3) plants were not altered (control). The percent of leaf area damaged by thrips (an order of insects) between early May and mid-June was recorded for each of 30 plants per combination of sex and floral characteristic. Here is the table of means and standard deviations (in parentheses):

Gender	Floral Characteristic Level		
	1	2	3
Males	0.11 (0.081)	1.28 (0.088)	1.63 (0.382)
Females	0.02 (0.002)	0.58 (0.321)	0.20 (0.035)

- (a) Give the degrees of freedom for the  $F$  statistics that are used to test for gender, floral characteristic, and the interaction.
- (b) Describe the main effects and interaction using appropriate graphs.
- (c) The researchers used the natural logarithm of percent area as the response in their analysis. Using the relationship between the means and standard deviations, explain why this was done.

### 13.52 Change-of-majors study: HSS.

Refer to the data given for the change-of-majors study in the data file MAJORS. Consider gender and whether students changed majors as the two factors. Analyze the data for HSS, the high school science grades. Your analysis should include a table of sample sizes, means, and standard deviations; Normal quantile plots; a plot of the means; and a two-way ANOVA using sex and major as the factors. Write a short summary of your conclusions.  MAJORS

### 13.53 Change-of-majors study: HSE.

Refer to the previous exercise. Analyze the data for HSE, the high school English grades. Your analysis should include a table of sample sizes, means, and standard deviations; Normal quantile plots; a plot of the means; and a two-way ANOVA using sex and major as the factors. Write a short summary of your conclusions.  MAJORS

### 13.54 Change-of-majors study: GPA.

Refer to Exercise 13.52. Analyze the data for GPA, the college grade point average. Your analysis should include a table of sample sizes, means, and standard deviations; Normal quantile plots; a plot of the means; and a two-way ANOVA using sex and major as the factors. Write a short summary of your conclusions.  MAJORS

### 13.55 Change-of-majors study: SATV.

Refer to Exercise 13.52. Analyze the data for SATV, the SAT Verbal score. Your analysis should include a table of sample sizes, means, and standard deviations; Normal quantile plots; a plot of the means; and a two-way ANOVA using sex and major as the factors. Write a short summary of your conclusions.  **MAJORS**

### **13.56 Search the Internet.**

Search the Internet or your library to find a study that is interesting to you and uses a two-way ANOVA to analyze the data. First describe the question or questions of interest and then give the details of how ANOVA was used to provide answers. Be sure to include how the study authors examined the assumptions for the analysis. Evaluate how well the authors used ANOVA in this study. If your evaluation finds the analysis deficient, make suggestions for how it could be improved.

# 14 Logistic Regression

CHAPTER

Girls		Boys
	0	8
6	1	2 8
8 6 5 3	2	1 3 4 4 4 7 7 8 8 8 9 9
8 7 6 5 5 4 3	3	1 1 2 3 7
8 3 3 2 2 1 0	4	
1	5	

14.1 The Logistic Regression Model

14.2 Inference for Logistic Regression

## Introduction

The simple and multiple linear regression methods we studied in Chapters 10 and 11 are used to model the relationship between a quantitative response variable and one or more explanatory variables. In this chapter we describe similar methods that are used when the response variable is a categorical variable with two possible values, such as a student applicant receives or does not receive financial aid, a patient lives or dies during emergency surgery, or your cell phone coverage is acceptable or not.



binomial setting, p. 322

In general, we call the two outcomes of the response variable “success” and “failure” and represent them by 1 (for a success) and 0 (for a failure). The mean is then the proportion of 1s,  $p = P(\text{success})$ . If our data are  $n$  independent observations, we have the *binomial setting*. What is *new* in this chapter is that the data now include at least one *explanatory variable*  $x$  and the probability  $p$  depends on the value of  $x$ . For example, suppose that we are studying whether a student applicant receives ( $y = 1$ ) or is denied ( $y = 0$ ) financial aid. Here,  $p$  is the probability that an applicant receives aid, and possible explanatory variables include (a) the financial support of the parents, (b) the income and savings of the applicant, and (c) whether the applicant has received financial aid before. Just as in multiple linear regression, the explanatory variables can be either categorical or quantitative. Logistic regression is a statistical method for describing these kinds of relationships.<sup>1</sup>

## 14.1 The Logistic Regression Model

When you complete this section, you will be able to

- Find the odds from a single probability.
- Describe the statistical model for logistic regression with a single explanatory variable.
- Find the odds ratio for comparing two proportions.

### Binomial distributions and odds

In Chapter 5 we studied binomial distributions and in Chapter 8 we learned how to do statistical inference for the proportion  $p$  of successes in the binomial setting. We start with a brief review of some of these ideas that we will need in this chapter.

#### Example

##### 14.1 A break from Facebook.

Example 8.1 (page 488) describes a Pew Internet survey of 1006 adults living in the United States. The 525 people who reported that they were Facebook users were asked, “Have you ever voluntarily taken a break from Facebook for a period of several weeks or more?” A total of 320 responded, “Yes, I have done this.”

In the notation of Chapter 5,  $p$  is the proportion of U.S. adult Facebook users who took a break from Facebook. The number of adults who took a break in an SRS of size  $n$  has the binomial distribution with parameters  $n$  and  $p$ . The sample size of Facebook users is  $n = 525$  and the number who took a break is the count  $X = 320$ . The sample proportion is

$$\hat{p} = 320/525 = 0.6095$$



odds, p. 632

Logistic regressions work with odds rather than proportions. The odds are

simply the ratio of the proportions for the two possible outcomes. If  $p^\wedge$  is the proportion for one outcome, then  $1-p^\wedge$  is the proportion for the second outcome:

$$\text{odds} = p^\wedge / (1-p^\wedge)$$

A similar formula for the population odds is obtained by substituting  $p$  for  $p^\wedge$  in this expression.

## Example

### 14.2 Odds of taking a break.

For the Facebook data, the proportion of adults who took a break in the sample of Facebook users is  $p^\wedge = 0.6095$ , so the proportion of adults who did not take a break is

$$1-p^\wedge = 1-0.6095 = 0.3905$$

Therefore, the odds of taking a break are

$$\begin{aligned} \text{odds} &= p^\wedge / (1-p^\wedge) \\ &= 0.6095 / 0.3905 \\ &= 1.561 \end{aligned}$$

When people speak about odds, they often round to integers or fractions. If we round 1.561 to  $1.5 = 3/2$ , we would say that the odds are approximately 3 to 2 that a Facebook user took a break. In a similar way, we could describe the odds that a Facebook user did *not* take a break as 2 to 3.

## USE YOUR KNOWLEDGE

### 14.1 Odds of drawing an ace.

If you deal one card from a standard deck, the probability that the card is an ace is  $4/52 = 1/13$ . Find the odds of drawing an ace.

## 14.2 Given the odds, find the probability.

If you know the odds, you can find the probability by solving the odds equation for the probability. So,  $p^{\wedge} = \text{odds} / (\text{odds} + 1)$  If the odds of an outcome are 4 (or 4 to 1), what is the probability of the outcome?

## Odds for two groups

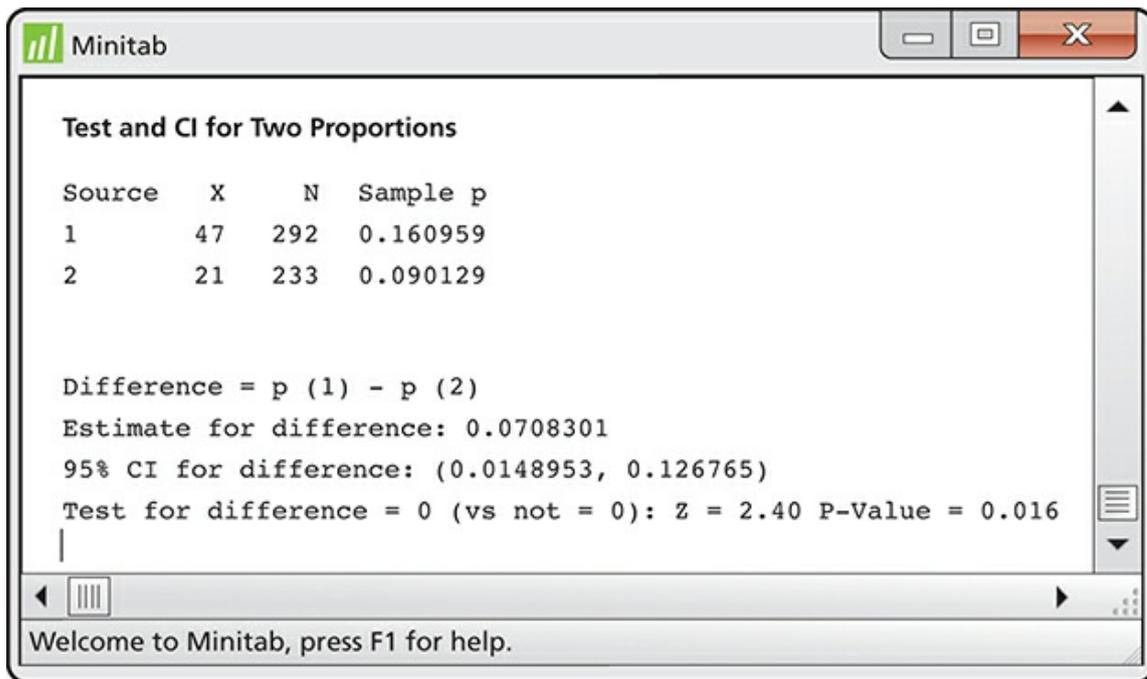
The Facebook users sample of 525 adults contained 292 women and 233 men, with 47 women and 21 men who responded that they increased their use of Facebook over the past year. Using the methods of Chapter 8, we can compare the proportions of male and female Facebook users who increased their use using a confidence interval (page 490) or significance test (page 495).

### Example

#### 14.3 Comparing the proportions of male and female Facebook users who increased their use.



Figure 14.1 contains output for this comparison. The sample proportion for women is 0.160959 (16%), and the sample proportion for men is 0.090129 (9%). The difference is 0.07083, and the 95% confidence interval is (0.015, 0.127). We can summarize this result by saying, “In this sample of Facebook users, the percent who increased their use is 7% higher among women than among men. This difference is statistically significant ( $z = 2.40$ ,  $P = 0.016$ ).”



**FIGURE 14.1**

Minitab output for the comparison of two proportions (female versus male Facebook users who took a break), for Example 14.3.

Another way to analyze these data is to use logistic regression. The explanatory variable is gender, a categorical variable. To use this in a regression (logistic or otherwise), we need to use a numeric code. The usual way to do this is with an **indicator variable**. For our problem we will use an indicator of whether or not the adult is a woman:

indicator variable

$$x = \begin{cases} 1 & \text{if the adult is a woman} \\ 0 & \text{if the adult is a man} \end{cases}$$

The response variable is the proportion of Facebook users who took a break. For use in a logistic regression, we perform two transformations on this variable. First, we convert to odds. For women,

$$\begin{aligned} \text{odds} &= \frac{p}{1-p} \\ &= \frac{0.160959}{1-0.160959} \\ &= 0.19184 \end{aligned}$$

Similarly, for men we have

$$\begin{aligned} \text{odds} &= \frac{p}{1-p} \\ &= \frac{0.090129}{1-0.090129} \\ &= 0.099057 \end{aligned}$$

## USE YOUR KNOWLEDGE

### 14.3 Energy drink commercials.

A study was designed to compare two energy drink commercials. Each participant was shown the commercials, A and B, in random order and asked to select the better one. There were 140 women and 130 men who participated in the study. Commercial A was selected by 65 women and by 67 men. Find the odds of selecting Commercial A for the men. Do the same for the women.

### 14.4 Find the odds.

Refer to the previous exercise. Find the odds of selecting Commercial B for the men. Do the same for the women.

## Model for logistic regression

In simple linear regression we modeled the mean  $\mu_y$  of the response variable  $y$  as a linear function of the explanatory variable:  $\mu_y = \beta_0 + \beta_1 x$ . When  $y$  is just 1 or 0 (success or failure), the mean is the probability  $p$  of a success. Logistic regression models the mean  $p$  in terms of an explanatory variable  $x$ . We might try to relate  $p$  and  $x$  as in simple linear regression:  $p = \beta_0 + \beta_1 x$ . Unfortunately, this is not a good model. Whenever  $\beta_1 \neq 0$ , extreme values of  $x$  will give values of  $\beta_0 + \beta_1 x$  that fall outside the range of possible values of  $p$ ,  $0 \leq p \leq 1$ .

The logistic regression solution to this difficulty is to transform the odds ( $p/(1 - p)$ ) using the natural logarithm. We use the term **log odds** or **logit** for this transformation. As we did with linear regression, we use  $y$  for the response variable. So for women,

log odds

logit

$$y = \log(\text{odds}) = \log(0.19184) = -1.6511$$

and for men,

$$y = \log(\text{odds}) = \log(0.099057) = -2.3121$$

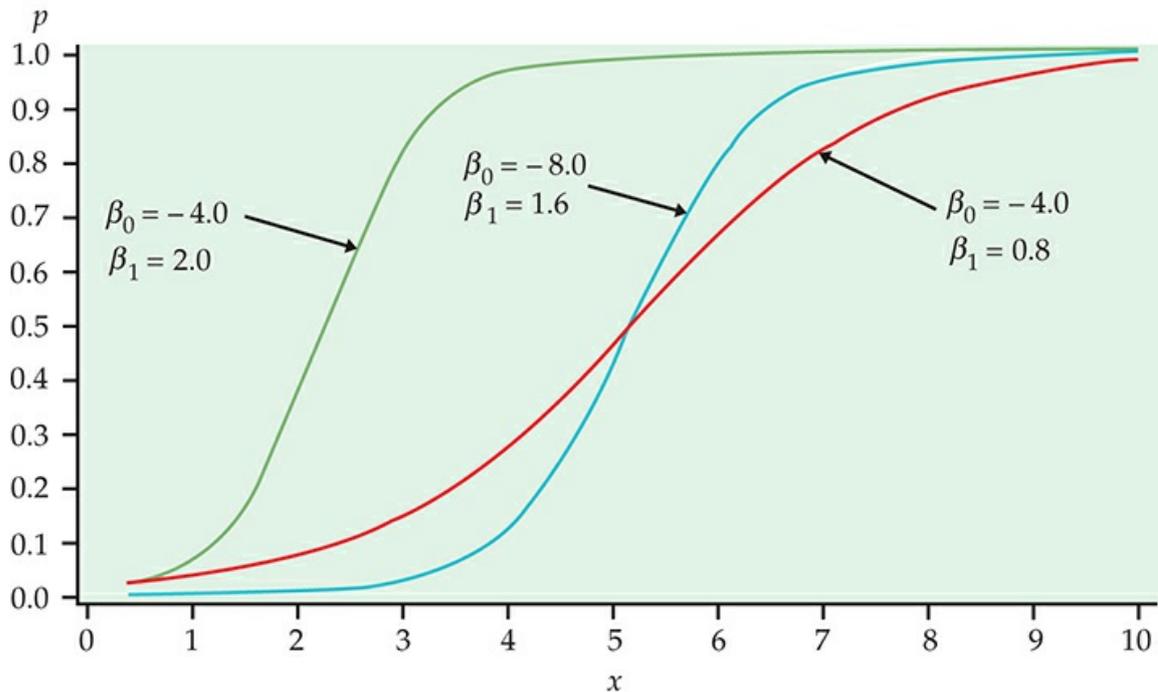
In these expressions for the log odds we use  $y$  as the observed value of the response variable, the log odds of having increased Facebook use. We are now

ready to build the logistic regression model.

We model the log odds as a linear function of the explanatory variable:

$$\log(p/1-p) = \beta_0 + \beta_1 x$$

Figure 14.2 graphs the relationship between  $p$  and  $x$  for some different values of  $\beta_0$  and  $\beta_1$ . For logistic regression we use *natural* logarithms. There are tables of natural logarithms, and many calculators have a built-in function for this transformation.



**FIGURE 14.2**

Plot of  $p$  versus  $x$  for different logistic regression models.

## USE YOUR KNOWLEDGE

### 14.5 Find the odds.

Refer to Exercise 14.3. Find the log odds for the men and the log odds for the women.

### 14.6 Find the odds.

Refer to Exercise 14.4. Find the log odds for the men and the log odds for the women.

## LOGISTIC REGRESSION MODEL

The **statistical model for logistic regression** is

$$\log(p/(1-p)) = \beta_0 + \beta_1 x$$

where  $p$  is a binomial proportion and  $x$  is the explanatory variable. The parameters of the logistic regression model are  $\beta_0$  and  $\beta_1$ .

### Example

#### 14.4 Model for increased use of Facebook.

For our Facebook use example, there are  $n = 525$  Facebook users in the sample. The explanatory variable is gender, which we have coded using an indicator variable with values  $x = 1$  for women and  $x = 0$  for men. The response variable is also an indicator variable. Thus, the Facebook user either increased his or her use of Facebook or did not increase his or her use. Think of the process of randomly selecting a Facebook user and recording the value of  $y$ , and whether or not the Facebook user increased his or her use. The model says that the probability ( $p$ ) that this user increased his or her use can depend upon the user's gender ( $x = 1$  or  $x = 0$ ). So there are two possible values for  $p$ , say  $p_{\text{women}}$  and  $p_{\text{men}}$ .

Logistic regression with an indicator explanatory variable is a very special case. It is important because many multiple logistic regression analyses focus on one or more such variables as the primary explanatory variables of interest. For now, we use this special case to understand a little more about the model.

The logistic regression model specifies the relationship between  $p$  and  $x$ . Since there are only two values for  $x$ , we write both equations. For women,

$$\log(p_{\text{women}}/(1-p_{\text{women}})) = \beta_0 + \beta_1$$

and for men,

$$\log(p_{\text{men}}/(1-p_{\text{men}})) = \beta_0$$

Note that there is a  $\beta_1$  term in the equation for women because  $x = 1$ , but it is missing in the equation for men because  $x = 0$ .

## Fitting and interpreting the logistic regression model

In general, the calculations needed to find estimates  $b_0$  and  $b_1$  for the parameters  $\beta_0$  and  $\beta_1$  are complex and require the use of software. When the explanatory variable has only two possible values, however, we can easily find the estimates. This simple framework also provides a setting where we can learn what the logistic regression parameters mean.

### Example

#### 14.5 Log odds for increasing Facebook use.

In the Facebook example, we found the log odds for women,

$$y = \log(\hat{p}^{\text{women}} / (1 - \hat{p}^{\text{women}})) = -1.6511$$

and for men,

$$y = \log(\hat{p}^{\text{men}} / (1 - \hat{p}^{\text{men}})) = -2.3121$$

The logistic regression model for women is

$$\log(\hat{p}^{\text{women}} / (1 - \hat{p}^{\text{women}})) = \beta_0 + \beta_1$$

and for men it is

$$\log(\hat{p}^{\text{men}} / (1 - \hat{p}^{\text{men}})) = \beta_0$$

To find the estimates  $b_0$  and  $b_1$ , we match the female and male model equations with the corresponding data equations. Thus, we see that the estimate of the intercept  $b_0$  is simply the log odds for the men:

$$b_0 = -2.1321$$

and the estimate of the slope is the difference between the log odds for the women and the log odds for the men:

$$b_1 = -1.6511 - (-2.3121) = 0.6610$$

The fitted logistic regression model is

$$\log(\text{odds}) = -1.6511 + 0.6610x$$

The slope in this logistic regression model is the difference between the log

odds for men and the log odds for women. Most people are not comfortable thinking in the log odds scale, so interpretation of the results in terms of the regression slope is difficult. Usually, we apply a transformation to help us. With a little algebra, it can be shown that

$$\text{odds}_{\text{women}} / \text{odds}_{\text{men}} = e^{0.6610} = 1.94$$

The transformation  $e^{0.6610}$  undoes the logarithm and transforms the logistic regression slope into an **odds ratio**, in this case the ratio of the odds that a woman increases her use of Facebook to the odds that a man increases his use of Facebook. In other words, we can multiply the odds for men by the odds ratio to obtain the odds for women:

**odds ratio**

$$\text{odds}_{\text{women}} = 1.94 \times \text{odds}_{\text{men}}$$

In this case, we would say that the odds for women are about twice the odds for men.

Notice that we have chosen the coding for the indicator variable so that the regression slope is positive. This will give an odds ratio that is greater than 1. Had we coded men as 1 and women as 0, the sign of the slope would be reversed, the fitted equation would be  $\log(\text{odds}) = -1.6511 - 0.6610x$ , and the odds ratio would be  $e^{-0.6610} = 0.5163$ . The odds for women are about half of the odds for men.

Logistic regression with an explanatory variable having two values is a very important special case. Here is an example where the explanatory variable is quantitative.

## Example

### 14.6 Is a movie going to be profitable?



MOVIES

The MOVIES data file (described on page 637) includes both the movie's budget and the total U.S. revenue. For this example, we will classify each movie as "profitable" ( $y = 1$ ) if U.S. revenue is larger than the budget and nonprofitable ( $y = 0$ ) otherwise. This is our response variable. The data file

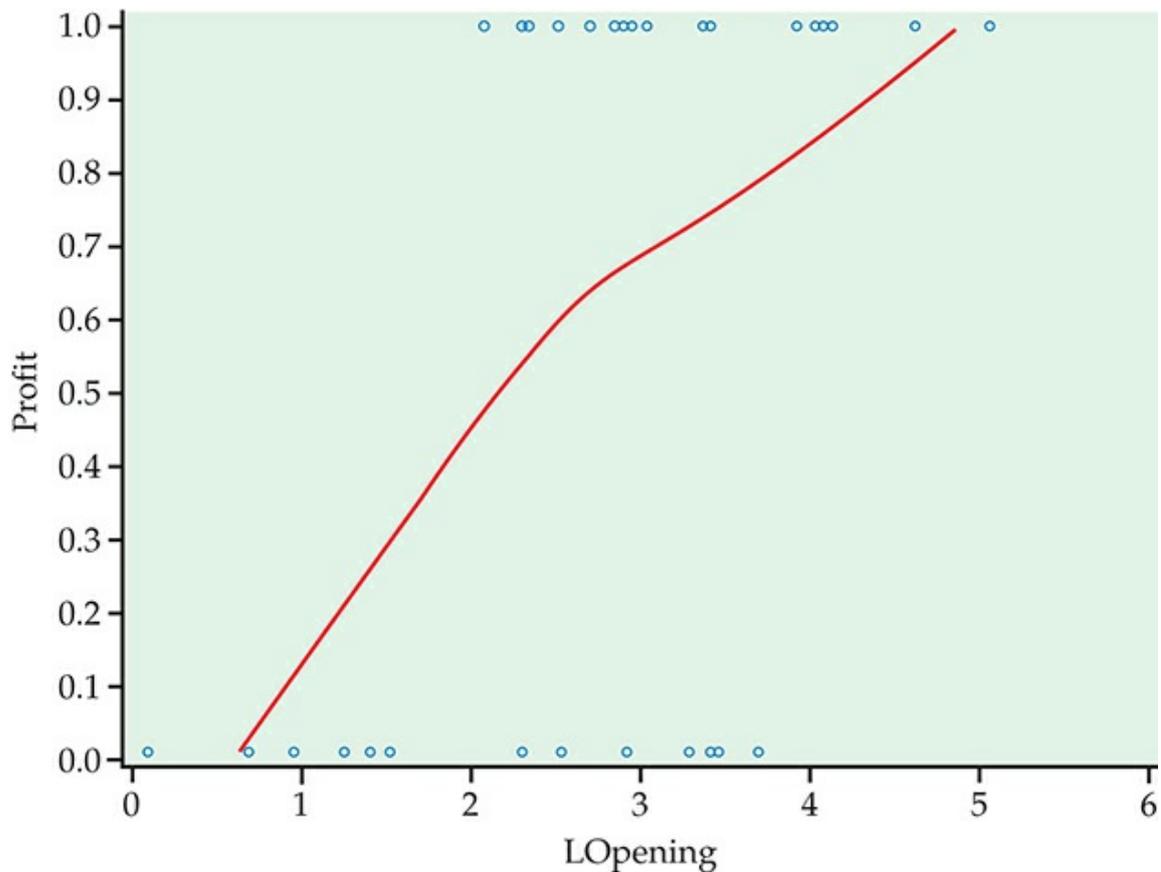
contains several explanatory variables, but we will focus here on the natural logarithm of the opening-weekend revenue. Figure 14.3 is a scatterplot of the data with a scatterplot smoother (page 96). The probability that a movie is profitable increases with the log opening-weekend revenue. Because the curve suggested by the smoother is reasonably close to an S-shaped curve like those in Figure 14.2, we fit the logistic regression model

$$\log(p1-p)=\beta_0+\beta_1x$$

where  $p$  is the probability that the movie is profitable and  $x$  is the log opening-weekend revenue. The model for estimated log odds fitted by software is

$$\log(\text{odds}) = b_0 + b_1x = -3.1658 + 1.3083x$$

The odds ratio is  $e^{b_1} = 3.700$ . This means that if log opening-weekend revenue  $x$  increases by one unit (roughly \$2.71 million), the odds that the movie will be profitable increase by 3.7 times.



**FIGURE 14.3**

Scatterplot of profit (Yes = 1, No = 0) versus the log opening-weekend revenue (LOpening) with a smooth function, for Example 14.6.

## USE YOUR KNOWLEDGE

### **14.7 Find the logistic regression equation and the odds ratio.**

Refer to Exercises 14.3 and 14.5. Find the logistic regression equation and the odds ratio.

### **14.8 Find the logistic regression equation and the odds ratio.**

Refer to Exercises 14.4 and 14.6. Find the logistic regression equation and the odds ratio.

## 14.2 Inference for Logistic Regression

When you complete this section, you will be able to

- For a logistic regression with a single explanatory variable, use software output to identify the estimates of the regression parameters and write the equation for the fitted model.
- For a logistic regression with a single explanatory variable, use software output to identify the 95% confidence interval for the regression slope and the significance test results for the null hypothesis that the slope is zero.
- For a logistic regression with a single explanatory variable, use software output to identify the odds ratio and the 95% confidence interval for the odds ratio. Interpret the odds ratio.
- For a logistic regression with several explanatory variables, use software output to identify the estimates of the regression parameters and write the equation for the fitted model.
- For a logistic regression with several explanatory variables, use software output to identify the significance test results for the null hypothesis that all regression slopes are zero.
- For a logistic regression with several explanatory variables, use software output to identify the 95% confidence intervals for the regression coefficients and the significance test results for the null hypothesis that each of the regression coefficients is zero.
- For a logistic regression with several explanatory variables, use software output to identify the odds ratio and the 95% confidence interval for the odds ratio for each explanatory variable. Interpret the odds ratios.

Statistical inference for logistic regression is very similar to statistical inference for simple linear regression. We calculate estimates of the model parameters and standard errors for these estimates. Confidence intervals are formed in the usual way, but we use standard Normal  $z^*$ -values rather than critical values from the  $t$  distributions. The ratio of the estimate of the slope to the standard error is the basis for hypothesis tests. Often the test statistics are given as the squares of these ratios, and in this case the  $P$ -values are obtained from the chi-square distribution with 1 degree of freedom.

### Confidence Intervals and Significance Tests

## CONFIDENCE INTERVALS AND SIGNIFICANCE TESTS FOR LOGISTIC REGRESSION PARAMETERS

A level  $C$  confidence interval for the slope  $\beta_1$  is

$$b_1 \pm z^* SE_{b_1}$$

The ratio of the odds for a value of the explanatory variable equal to  $x + 1$  to the odds for a value of the explanatory variable equal to  $x$  is the **odds ratio**.

A level  $C$  confidence interval for the odds ratio  $e^{\beta_1}$  is obtained by transforming the confidence interval for the slope:

$$(e^{b_1 - z^* SE_{b_1}}, e^{b_1 + z^* SE_{b_1}})$$

In these expressions  $z^*$  is the value for the standard Normal density curve with area  $C$  between  $-z^*$  and  $z^*$ .

To test the hypothesis  $H_0: \beta_1 = 0$ , compute the **test statistic**

$$z = b_1 / SE_{b_1}$$

The  $P$ -value for the significance test of  $H_0$  against  $H_a: \beta_1 \neq 0$  is computed using the fact that, when the null hypothesis is true,  $z$  has approximately a standard Normal distribution.

The statistic  $z$  is sometimes called a **Wald statistic**. Output from some statistical software reports the significance test result in terms of the square of the  $z$  statistic.

**Wald statistic**

$$X^2 = z^2$$

 **LOOK BACK**

chi-square statistic, p. 538

This statistic is called a chi-square statistic. When the null hypothesis is true, it has a distribution that is approximately a  $\chi^2$  distribution with 1 degree of freedom, and the  $P$ -value is calculated as  $P(\chi^2 \geq X^2)$ . Because the square of a standard Normal random variable has a  $\chi^2$  distribution with 1 degree of freedom, the  $z$  statistic and the chi-square statistic give the same results for statistical inference.

We have expressed the hypothesis-testing framework in terms of the slope  $\beta_1$  because this form closely resembles what we studied in simple linear regression. In many applications, however, the results are expressed in terms of the odds ratio. A slope of 0 is the same as an odds ratio of 1, so we often express the null hypothesis of interest as “the odds ratio is 1.” This means that the two odds are equal and the

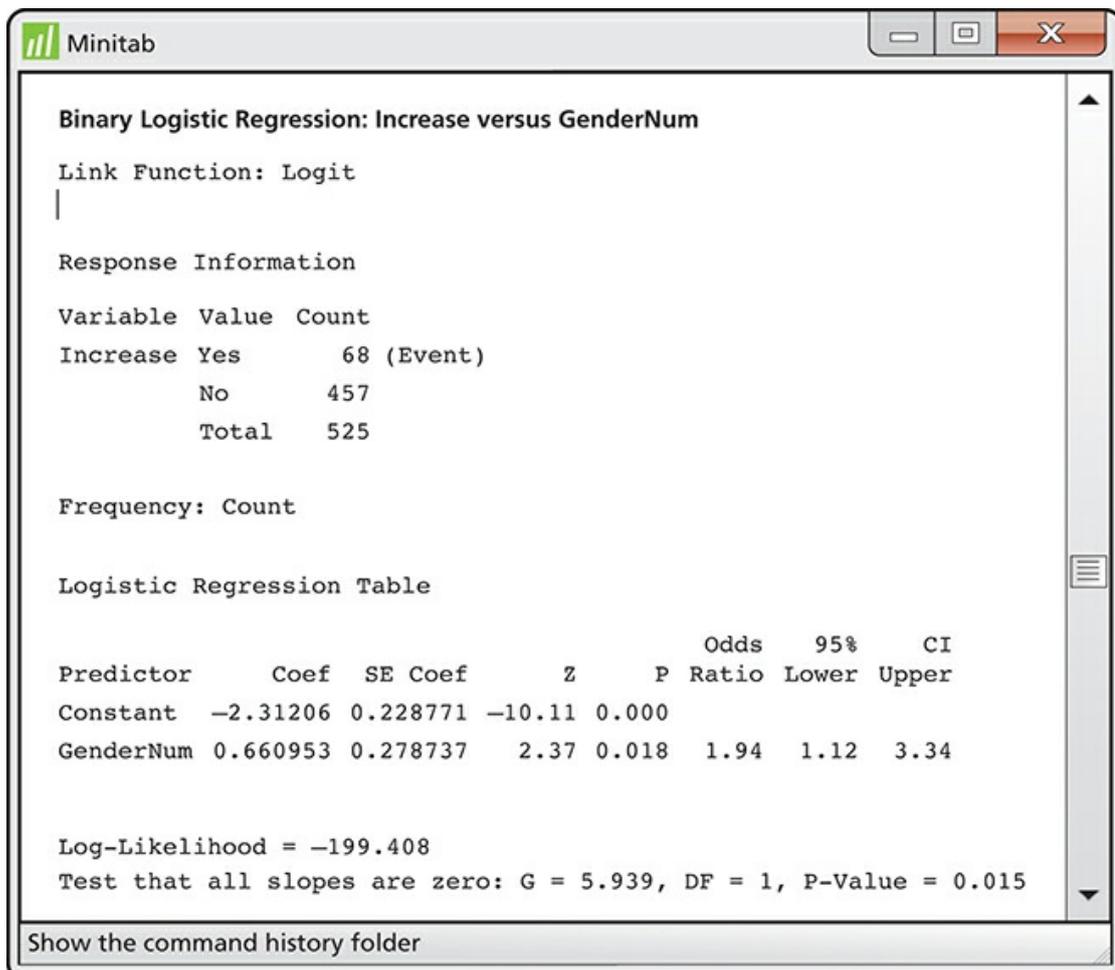
explanatory variable is not useful for predicting the odds.

## Example

### 14.7 Software output.



Figure 14.4 gives the output from Minitab for the Facebook increased use example described in Example 14.5. The parameter estimates are given as  $b_0 = -2.31206$  and  $b_1 = 0.660953$ . The standard errors are 0.228771 and 0.278737, respectively.



## FIGURE 14.4

Logistic regression output from Minitab for the Facebook increased use data, for Example 14.7.

The 95% confidence interval for the slope is

$$\begin{aligned} b_1 \pm z^* SE_{b_1} &= 0.660953 \pm (1.96) (0.278737) \\ &= 0.660953 \pm 0.546325 \end{aligned}$$

We are 95% confident that the slope is between 0.1146 and 1.2073.

The output also provides the odds ratio 1.94 and a 95% confidence interval, 1.12 to 3.34. For this problem we would report, “Female Facebook users are more likely to increase their use of Facebook than male Facebook users (odds ratio = 1.94, 95% CI = 1.12 to 3.34).”

## USE YOUR KNOWLEDGE

### 14.9 Verify the calculation of the odds ratio.



FACETIM

Refer to Example 14.7. Verify that the odds ratio, 1.94, is  $e^{b_1}$ .

### 14.10 Verify the calculation of the confidence interval.



FACETIM

Refer to Example 14.7. Verify that the 95% confidence interval for the odds ratio, 1.12 to 3.34, is

$$(e^{b_1 - z^* SE_{b_1}}, e^{b_1 + z^* SE_{b_1}})$$

where  $z^*$  Explain why we use this value of  $z^*$  in the calculation.

In applications such as these, it is standard to use 95% for the confidence coefficient. With this convention, the confidence interval gives us the result of testing the null hypothesis that the odds ratio is 1 for a significance level of 0.05. If the confidence interval does not include 1, we reject  $H_0$  and conclude that the odds for the two groups are different; if the interval does include 1, the data do not provide enough evidence to distinguish the groups in this way.

The following example is typical of many applications of logistic regression. Here there is a designed experiment with five different values for the explanatory variable.

## Example

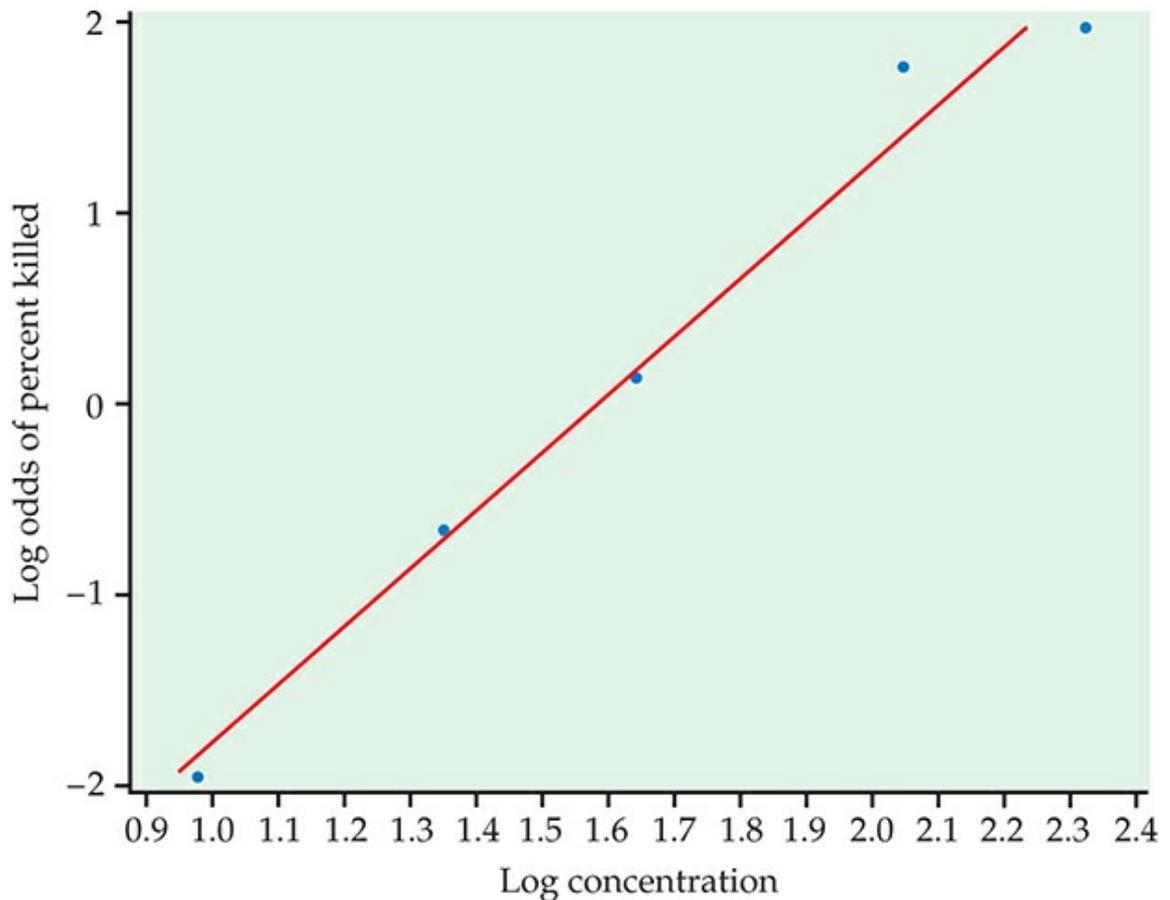
### 14.8 An insecticide for aphids.



An experiment was designed to examine how well the insecticide rotenone kills an aphid, called *Macrosiphoniella sanborni*, that feeds on the chrysanthemum plant.<sup>2</sup> The explanatory variable is the concentration (in log of milligrams per liter) of the insecticide. At each concentration, approximately 50 insects were exposed. Each insect was either killed or not killed. We summarize the data using the number killed. The response variable for logistic regression is the log odds of the proportion killed. Here are the data:

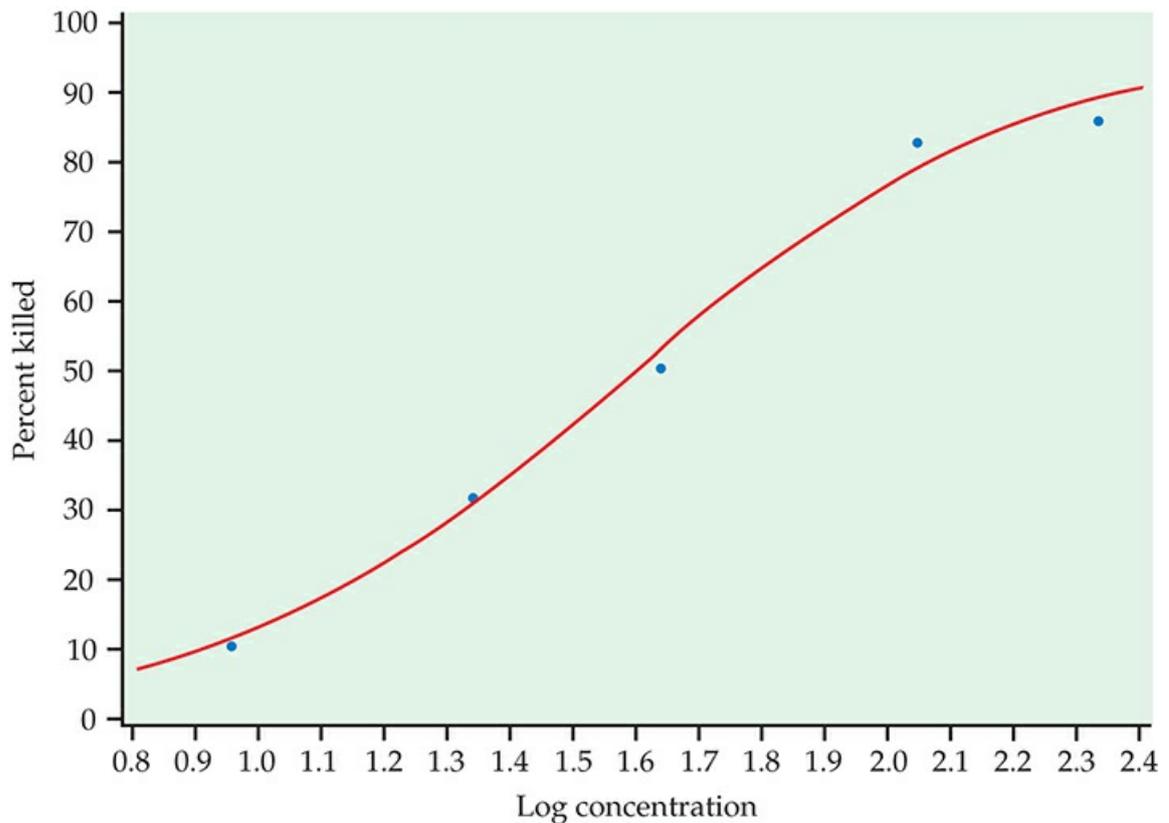
Concentration (log)	Number of insects	Number killed
0.96	50	6
1.33	48	16
1.63	46	24
2.04	49	42
2.32	50	44

If we transform the response variable (by taking log odds) and use least squares, we get the fit illustrated in Figure 14.5. The logistic regression fit is given in Figure 14.6. It is a transformed version of Figure 14.5 with the fit calculated using the logistic model.



**FIGURE 14.5**

Plot of log odds of percent killed versus log concentration for the insecticide data, for Example 14.8.



**FIGURE 14.6**

Plot of the percent killed versus log concentration with the logistic fit for the insecticide data, for Example 14.8.

One of the major themes of this text is that we should present the results of a statistical analysis with a graph. For the insecticide example we have done this with Figure 14.6, and the results appear to be convincing. But suppose that rotenone has no ability to kill *Macrosiphoniella sanborni*. What is the chance that we would observe experimental results at least as convincing as what we observed if this supposition were true? The answer is the  $P$ -value for the test of the null hypothesis that the logistic regression slope is zero. If this  $P$ -value is not small, our graph may be misleading. Statistical inference provides what we need.

## Example

### 14.9 Software output.

Figure 14.7 gives the output from Minitab, SPSS, and JMP for the logistic regression analysis of the insecticide data. The model is

$$\log(p1-p)=\beta_0+\beta_1x$$

where the values of the explanatory variable  $x$  are 0.96, 1.33, 1.63, 2.04, and 2.32. From the output in Minitab and SPSS, we see that the fitted model is

$$\log(\text{odds}) = b_0 + b_1x = -4.89 + 3.11x$$

This is the fit that we plotted in Figure 14.6. The null hypothesis that  $\beta_1 = 0$  is clearly rejected ( $z = 8.01$  in Minitab, Wald  $\chi^2 = 64.233$  in SPSS, and  $\chi^2 = 64.23$  in JMP;  $P < 0.001$  for all). We calculate a 95% confidence interval for  $\beta_1$  using the estimate  $b_1 = 3.1088$  and its standard error  $SE_{b_1} = 0.3879$  given in the output:

$$\begin{aligned} b_1 \pm z^* SE_{b_1} &= 3.1088 \pm (1.96) (0.3879) \\ &= 3.1088 \pm 0.7603 \end{aligned}$$

We are 95% confident that the true value of the slope is between 2.35 and 3.87.

**Binary Logistic Regression: Kill versus Lconc**

Link Function: Logit

Response Information

Variable	Value	Count
Kill	Yes	132 (Event)
	No	111
Total		243

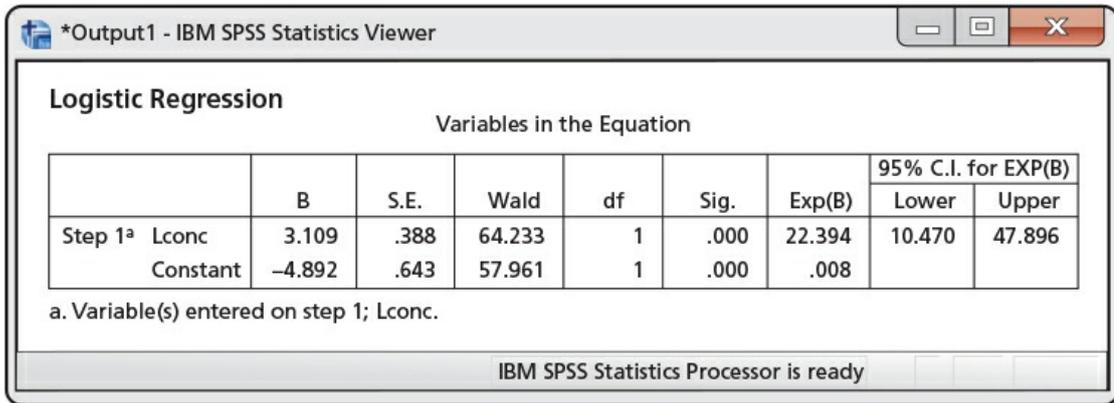
Frequency: NUMBER

Logistic Regression Table

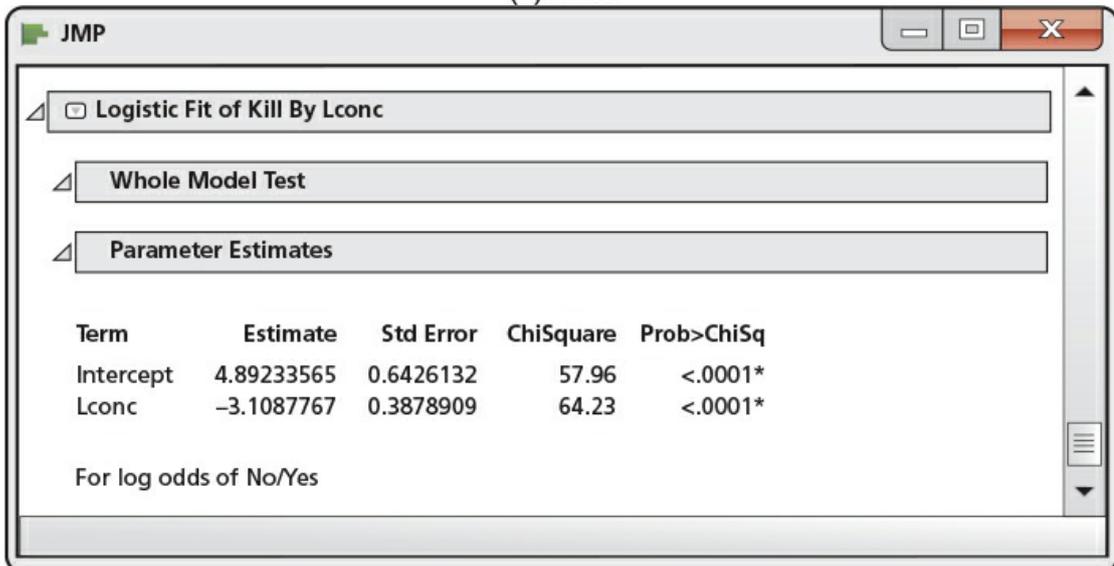
Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% Lower	95% Upper
Constant	-4.89234	0.642613	-7.61	0.000			
Lconc	3.10878	0.387891	8.01	0.000	22.39	10.47	47.90

Welcome to Minitab, press F1 for help.

(a) Minitab



(b) SPSS



(c) JMP

**FIGURE 14.7**

Logistic regression output from (a) Minitab, (b) SPSS, and (c) JMP for the insecticide data, for Example 14.9.

The odds ratio is given on the Minitab output as 22.39. An increase of one unit in the log concentration of insecticide ( $x$ ) is associated with a 22-fold increase in the odds that an insect will be killed. Minitab gives the 95% confidence interval for the odds ratio, 10.47 to 47.90. We could calculate this from the confidence interval for the slope:

$$\begin{aligned}
 (e^{b_1 - z^* SE_{b_1}}, e^{b_1 + z^* SE_{b_1}}) &= (e^{2.3485}, e^{3.8691}) \\
 &= (10.47, 47.90)
 \end{aligned}$$

Note again that the test of the null hypothesis that the slope is 0 is the same as the test of the null hypothesis that the odds are 1. If we were reporting the results in terms of the odds, we could say, “The odds of killing an insect increase by a factor of 22.4 for each unit increase in the log concentration of insecticide ( $X^2 = 64.23$ ,  $P < 0.001$ ; 95% CI = 10.5 to 47.9).”

Note that JMP gives the fitted model as

$$\log(\text{odds}) = 4.89 - 3.11x$$



We see that the regression coefficients  $b_0$  and  $b_1$  are  $-1$  times the coefficients given by Minitab and SPSS. The reason for this is that JMP models the log odds that an insect is *not* killed rather than the log odds that an insect is killed, as shown in the other two outputs. *Always examine software output carefully to be sure that the results you are getting correspond exactly to the analysis that you are trying to perform.* For this analysis, we know from our graph in Figure 14.6 that the slope should be positive.

In Example 14.6 we studied the problem of predicting whether or not a movie was going to make a profit using the log opening weekend revenue as the explanatory variable. We now revisit this example and show how statistical inference is an important part of the conclusion.

## Example

### 14.10 Software output.



#### MOVIES

Figure 14.8 gives the output from Minitab for a logistic regression analysis using log opening-weekend revenue as the explanatory variable. From the Minitab output, we see that the fitted model is

$$\log(\text{odds}) = b_0 + b_1x = -3.166 + 1.3083x$$

From the output, we see that because  $P = 0.007$ , we can reject the null hypothesis that  $\beta_1 = 0$ . The value of the test statistic is  $X^2 = 7.26$  with 1 degree of freedom. We use the estimate  $b_1 = 1.3083$  and its standard error  $SE_{b_1} = 0.4855$  to compute the 95% confidence interval for  $\beta_1$ :

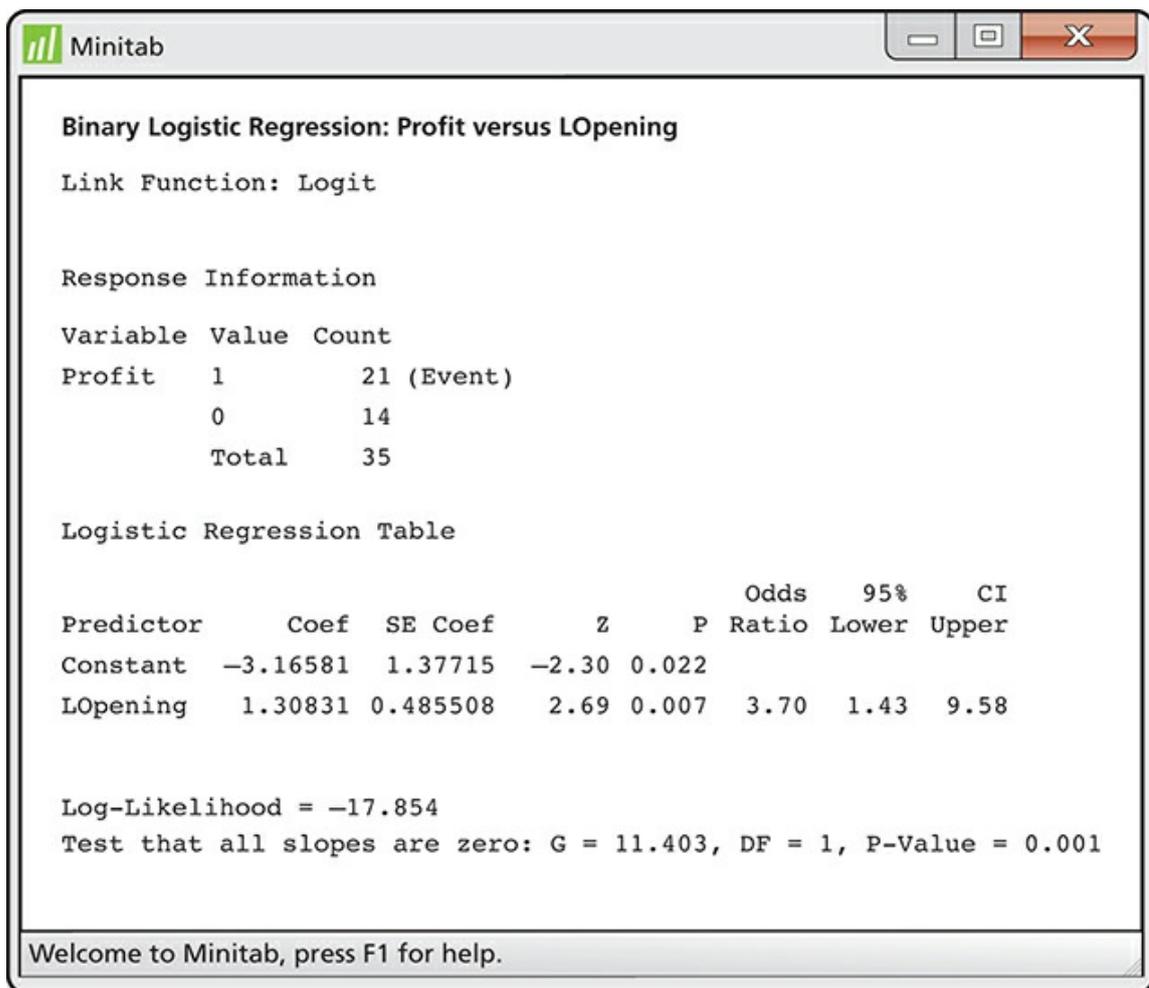
$$b_1 \pm z^*SE_{b_1} = 1.3083 \pm (1.96) (0.4855)$$

$$= 1.3083 \pm 0.9516$$

Our estimate of the slope is 1.3083, and we are 95% confident that the true value is between 0.3567 and 2.2599. For the odds ratio, the estimate on the output is 3.70. The 95% confidence interval is

$$(e^{b_1 - z^*SE_{b_1}}, e^{b_1 + z^*SE_{b_1}}) = (e^{0.3567}, e^{2.2599})$$

$$= (1.43, 9.58)$$



**FIGURE 14.8**

Logistic regression output from Minitab for the movie profitability data with log opening-weekend revenue as the explanatory variable, for Example 14.10.

We estimate that an opening-weekend revenue that is one unit larger (roughly \$2.71 million) will increase the odds that a movie is profitable by about 4 times. The data, however, do not give us a very accurate estimate. The odds ratio could be as small as 1.43 or as large as 9.58 with 95% confidence. We have evidence to conclude that movies with higher opening-weekend revenues are more likely to be profitable, but establishing the relationship accurately would require more data.

Note that the SAS output (not shown), like JMP, gives the same estimates of the regression coefficients but with opposite signs. By default, this software models the odds that the movie is not profitable.

## Multiple logistic regression

The movie example that we just considered naturally leads us to the next topic. The MOVIES data file includes additional explanatory variables. Do these other explanatory variables contain additional information that will give us a better prediction of profitability? We use **multiple logistic regression** to answer this question. Generating the computer output is easy, just as it was when we generalized simple linear regression with one explanatory variable to multiple linear regression with more than one explanatory variable in Chapter 11. The statistical concepts are similar, although the computations are more complex. Here is the example.

multiple logistic regression

### Example

#### 14.11 Software output.



As in Example 14.10, we predict the odds that a movie is profitable. The explanatory variables are log opening-weekend revenue (LOpening), number of theaters (Theaters), and the movie's IMDb rating at the end of the first week (Opinion), which is on a 1 to 10 scale (10 being best). Figure 14.9 gives the outputs from SAS, Minitab, and SPSS. The fitted model is

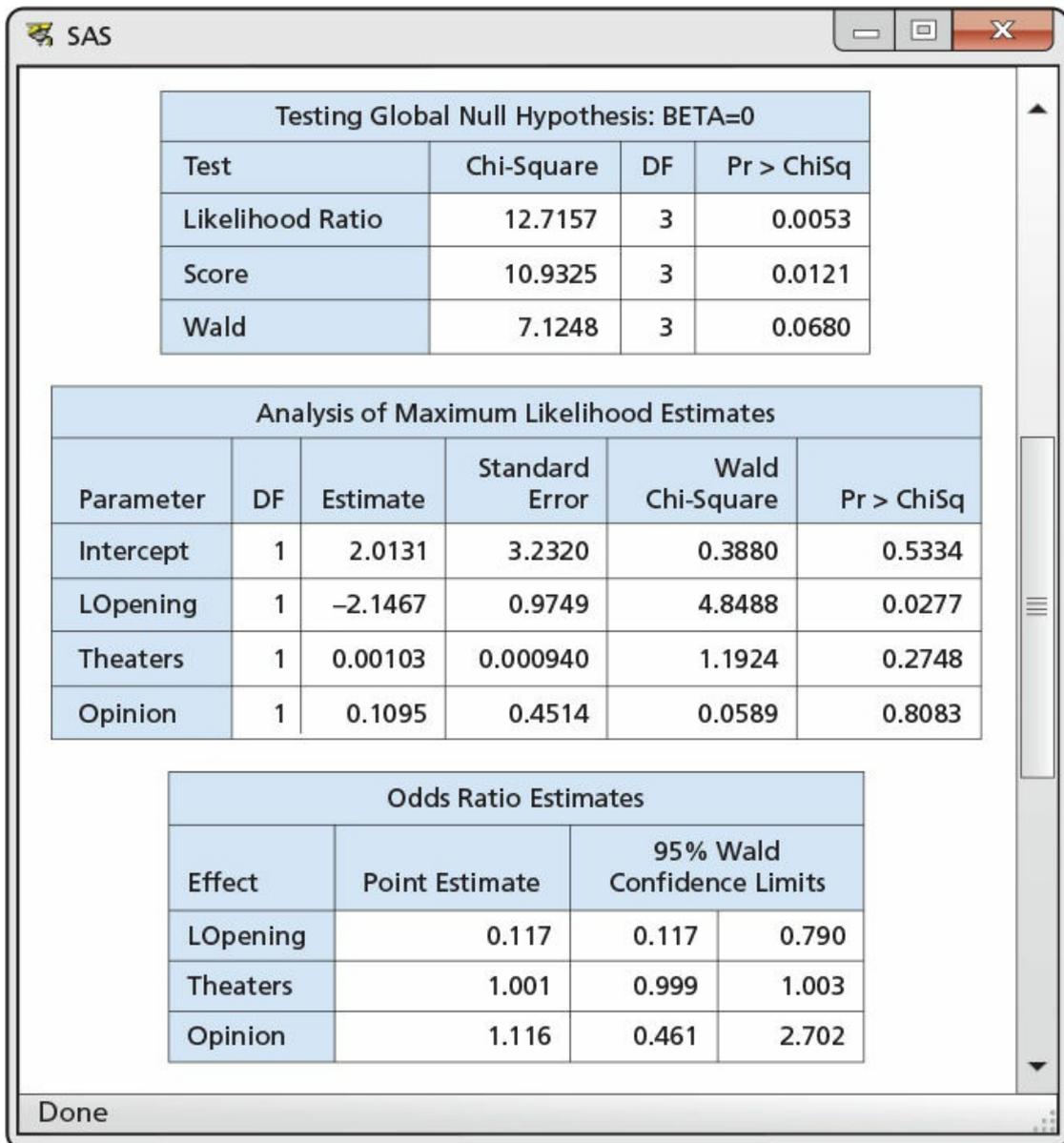


$$\begin{aligned} \log(\text{odds}) &= b_0 + b_1 \text{LOpening} + b_2 \text{Theaters} + b_3 \text{Opinion} \\ &= -2.013 + 2.147 \text{LOpening} - 0.001 \text{Theaters} - 0.109 \text{Opinion} \end{aligned}$$

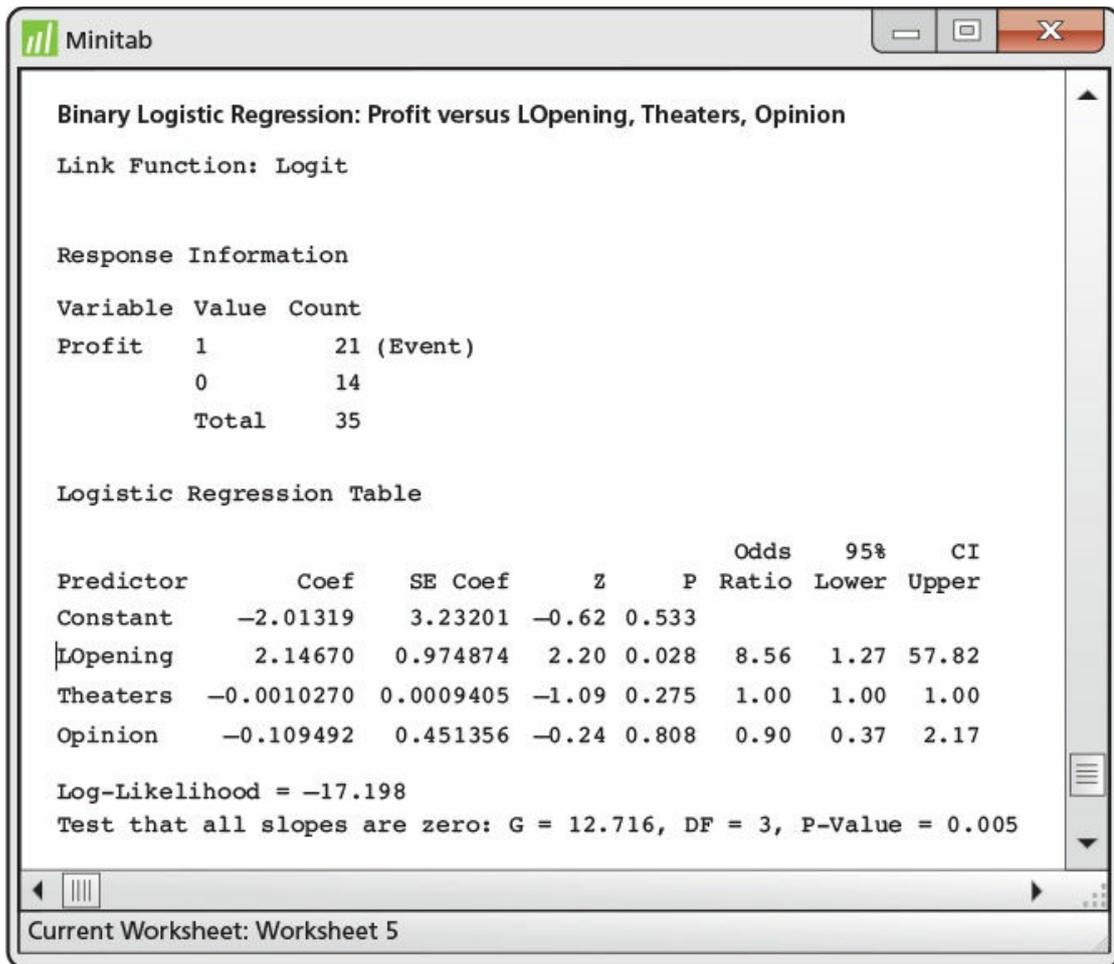
Note that the coefficients given by SAS have the signs reversed because SAS models the odds that the move will not be profitable.

When analyzing data using multiple linear regression, we first examine the hypothesis that all the regression coefficients for the explanatory variables are zero. We do the same for multiple logistic regression. The hypothesis

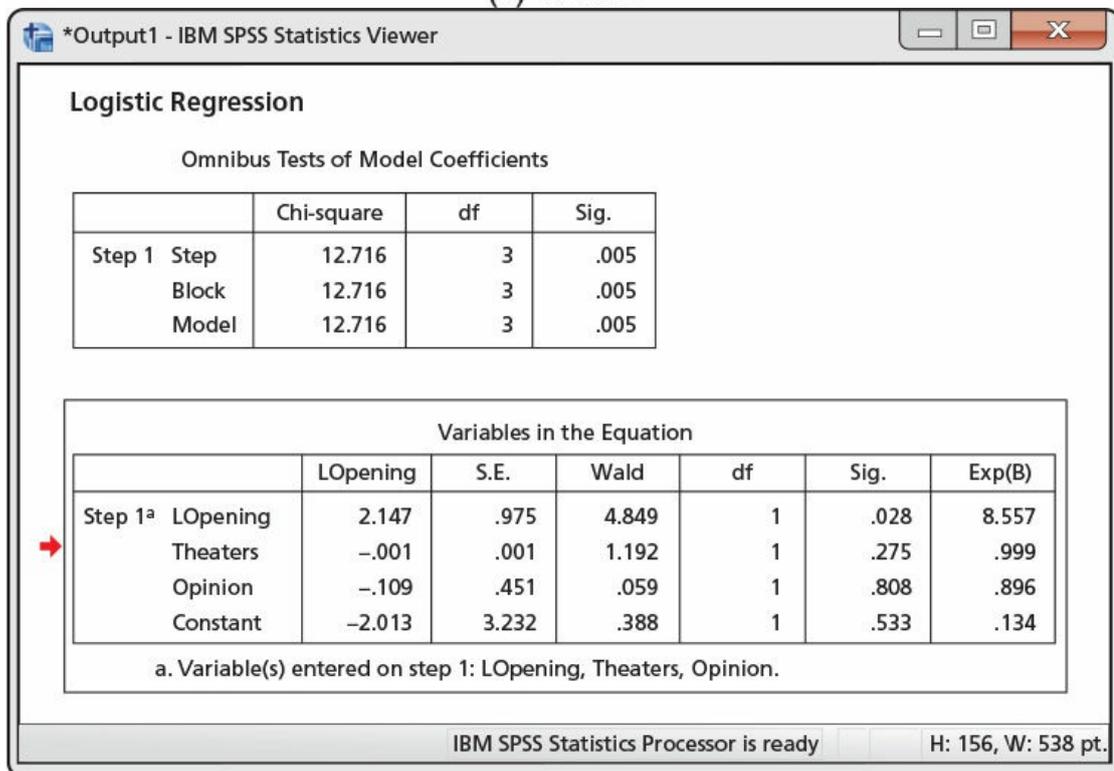
$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$



(a) SAS



(b) Minitab



(c) SPSS

## FIGURE 14.9

Logistic regression output from (a) SAS, (b) Minitab, and (c) SPSS for the movie profitability data with log opening-weekend revenue, number of theaters, and the movie's IMDb rating as the explanatory variables, for Example 14.11.

is tested by a chi-square statistic with 3 degrees of freedom. (The degrees of freedom are 3 because there are 3 coefficients that are set to zero in the null hypothesis.) For Minitab, this is given in the last line of the output, and the statistic is called "G." The value is  $G = 12.716$  and the  $P$ -value is 0.005. We reject  $H_0$  and conclude that one or more of the explanatory variables can be used to predict the odds that a movie is profitable.

We now examine the coefficients for each variable and the tests that each of these is zero *in a model that contains the other two*. The  $P$ -values are 0.028, 0.275, and 0.808. The null hypotheses  $H_0: \beta_2 = 0$  and  $H_0: \beta_3 = 0$  cannot be rejected. That is, log opening-weekend revenue is the only predictor that adds significant predictive ability once the other two are already in the model.

Our initial multiple logistic regression analysis told us that the explanatory variables contain information that is useful for predicting whether or not the movie is profitable. Because the explanatory variables are correlated, however, we cannot clearly distinguish which variables or combinations of variables are important. Further analysis of these data using subsets of the three explanatory variables is needed to clarify the situation. We leave this work for the exercises.

## CHAPTER 14 Summary

If  $\hat{p}$  is the sample proportion, then the **odds** are  $\hat{p}/(1-\hat{p})$ , the ratio of the proportion of times the event happens to the proportion of times the event does not happen.

The **logistic regression model** relates the **log of the odds** to the explanatory variable:

$$\log(\pi/1-\pi)=\beta_0+\beta_1x_i$$

where the response variables for  $i = 1, 2, \dots, n$  are independent binomial random variables with parameters 1 and  $p_i$ ; that is, they are independent with distributions  $B(1, p_i)$ . The explanatory variable is  $x$ .

The **parameters** of the logistic model are  $\beta_0$  and  $\beta_1$ .

The **odds ratio** is  $e^{\beta_1}$  where  $\beta_1$  is the slope in the logistic regression model.

A **level  $C$  confidence interval for the intercept  $\beta_0$**  is

$$b_0 \pm z^* SE_{b_0}$$

A **level  $C$  confidence interval for the slope  $\beta_1$**  is

$$b_1 \pm z^* SE_{b_1}$$

A **level  $C$  confidence interval for the odds ratio  $e^{\beta_1}$**  is obtained by transforming the confidence interval for the slope:

$$(e^{b_1 - z^* SE_{b_1}}, e^{b_1 + z^* SE_{b_1}})$$

In these expressions  $z^*$  is the value for the standard Normal density curve with area  $C$  between  $-z^*$  and  $z^*$ .

To test the hypothesis  $H_0: \beta_1 = 0$ , compute the **test statistic**

$$z = b_1 / SE_{b_1}$$

and use the fact that  $z$  has a distribution that is approximately the standard Normal distribution when the null hypothesis is true. This statistic is sometimes called the **Wald statistic**. An alternative equivalent procedure is to report the square of  $z$ ,

$$X^2 = z^2$$

This statistic has a distribution that is approximately a  $\chi^2$  distribution with 1 degree of freedom, and the  $P$ -value is calculated as  $P(\chi^2 \geq X^2)$ . This is the same as testing the null hypothesis that the odds ratio is 1.

In **multiple logistic regression** the response variable has two possible values, as in logistic regression, but there can be several explanatory variables.

## CHAPTER 14 Exercises

*For Exercises 14.1 and 14.2, see page 14-3; for Exercises 14.3 and 14.4, see page 14-4; for Exercises 14.5 and 14.6, see page 14-6; for Exercises 14.7 and 14.8, see page 14-9; and for Exercises 14.9 and 14.10, see page 14-11.*

### 14.11 How did you use your cell phone?

A Pew Internet Poll asked cell phone owners about how they used their cell phones. One question asked whether or not during the past 30 days they had used their phone while in a store to call a friend or family member for advice about a purchase they were considering. The poll surveyed 1003 adults living in the United States by telephone. Of these, 462 responded that they had used their cell phone while in a store within the last 30 days to call a friend or family member for advice about a purchase they were considering.<sup>3</sup>

- What proportion of those surveyed reported that they used their cell phone while in a store within the last 30 days to call a friend or family member for advice about a purchase they were considering?
- Find the odds for the probability that you found in (a).

### 14.12 Find some odds.

For each of the following probabilities, find the odds: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. Make a plot of the odds versus the probabilities and describe the relationship.

### 14.13 A logistic model for cell phones.

Refer to Exercise 14.11. Suppose that you want to investigate differences in cell phone use among customers of different ages. You create an indicator explanatory variable  $x$  that has the value 1 if the customer is 25 years of age or less and is 0 if the customer is over 25 years of age.

- Describe the statistical model for logistic regression in this setting.
- Explain the relationship between the regression coefficients and the odds ratios for the two groups of customers defined by  $x$ .

### 14.14 Another logistic model for cell phones and age.

Refer to the previous exercise. Suppose that you use the actual value of age in years as the explanatory variable in a logistic regression model.

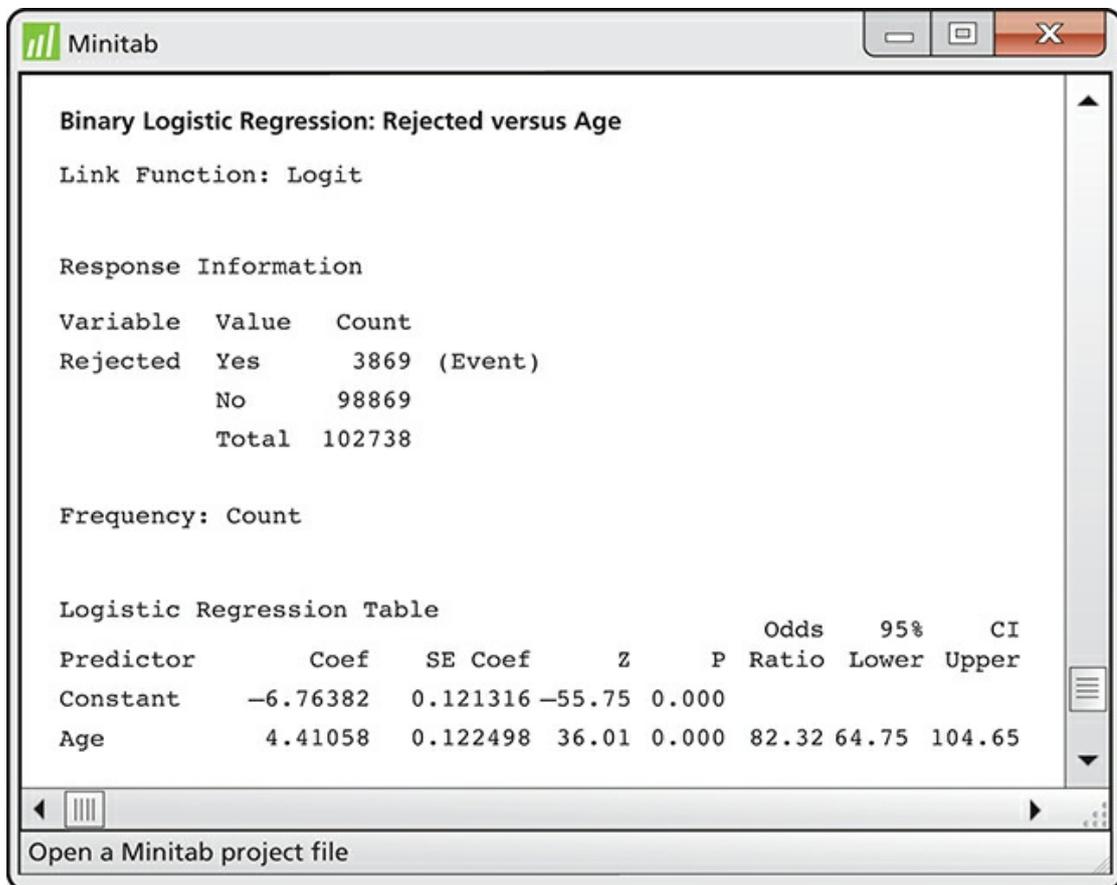
- Describe the statistical model for logistic regression in this setting.
- Interpret the regression slope in terms of an effect based on a difference in age of one year.
- This model requires an assumption that is not needed in the model that you described in the

previous exercise. Explain the assumption and describe a method for examining whether or not it is a reasonable assumption to make for these data. (*Hint*: Refer to Example 14.8 and Figure 14.5, page 14-12.)

### 14.15 A logistic regression for teeth and military service.

Exercise 8.58 (page 523) describes data on the numbers of U.S. recruits who were rejected for service in a war against Spain because they did not have enough teeth. The exercise compared the rejection rate for recruits who were under the age of 20 with the rate for those who were 40 or over. To run a logistic regression for this setting we define an indicator explanatory variable  $x$  with values of 0 for age under 20 and 1 for age 40 or over. Figure 14.10 gives output from Minitab for this analysis.  **TEETH1**

- (a) How many recruits were examined? How many were rejected and how many were not rejected?
- (b) Write the fitted logistic regression model.



**FIGURE 14.10** Logistic regression output from Minitab for predicting recruit rejection using age in two categories, for Exercises 14.15 to 14.17.

### 14.16 Inference for teeth and military service.

Refer to the previous exercise.  **TEETH1**

- (a) Using the information provided in the output in Figure 14.10, calculate and interpret the 95%

confidence interval for the regression slope.

(b) Describe and interpret the results of the significance test for the regression slope. Be sure to give the null and alternative hypotheses, the test statistic, and the  $P$ -value with your conclusion.

### 14.17 Odds ratio for teeth and military service.

Refer to the two previous exercises.  **TEETH1**

- (a) Give the odds ratio for this analysis.
- (b) Give the 95% confidence interval for the odds ratio.
- (c) Give a brief description of the meaning of the odds ratio in this analysis.

### 14.18 Teeth and military service with six age categories.

In Exercises 14.15 to 14.17 we used logistic regression to study the relationship between being rejected for military service because a recruit did not have enough teeth and age categorized into two groups, under 20 and 40 or over. Data are available for all recruits categorized into six age groups. Let's look at a logistic regression that uses all the data to predict rejection for military service based on teeth. There are six age groups: under 20, 20–25, 25–30, 30–35, 35–40, and 40 or over. We define indicator explanatory variables for the last five groups. This is similar to defining a single

indicator explanatory variable for an analysis of two groups.  **TEETH2**

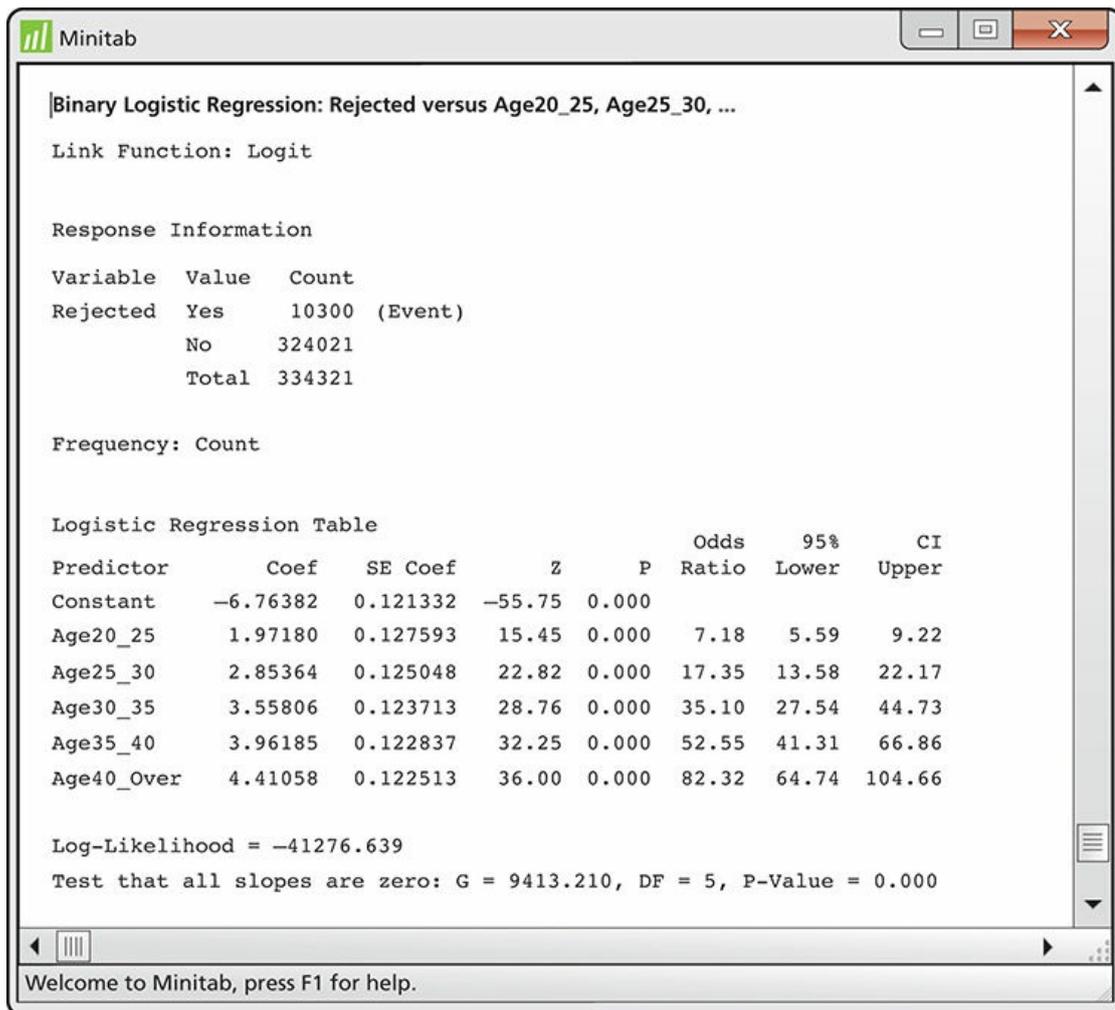
Figure 14.11 gives the Minitab output for the logistic regression to predict rejection using the five age indicator explanatory variables.

- (a) Use the output to find the fitted model.
- (b) Is there a pattern in the values of the regression slopes? If yes, describe it.

### 14.19 Inference for the multiple logistic regression model.

Refer to the previous exercise.  **TEETH2**

- (a) Describe and interpret the significance test that tests the null hypothesis that all regression coefficients are zero.
- (b) Using the information provided in the output in Figure 14.11, calculate and interpret the 95% confidence interval for each of the regression slopes.
- (c) Describe and interpret the results of the significance test for each regression slope. Be sure to give the null and alternative hypotheses, the test statistic, and the  $P$ -value with your conclusion.



**FIGURE 14.11**

Logistic regression output from Minitab for predicting recruit rejection using age in six categories, for Exercises 14.18 to 14.21.

### 14.20 Odds ratios for the multiple logistic regression model.

Refer to the two previous exercises.  **TEETH2**

- Give the odds ratio for each explanatory variable.
- Give the 95% confidence interval for each odds ratio.
- Give a brief description of the meaning of each odds ratio in this analysis.

### 14.21 Compare the multiple logistic regression analysis with the two-way table.

The data analyzed in Figure 14.11 were analyzed in Exercise 9.22 and Figure 9.7 (page 557) using a  $2 \times 6$  table of counts. Compare these two approaches to the analysis of these data. Describe some strengths and weaknesses of each approach. Which do you prefer? Give reasons for your answer.

## 14.22 Exergaming in Canada.

Exergames are active video games such as rhythmic dancing games, virtual bicycles, balance board simulators, and virtual sports simulators that require a screen and a console. A study of exergaming by students in grades 10 and 11 in Montreal, Canada, examined many factors related to participation in exergaming.<sup>4</sup> Of the 358 students who reported that they stressed about their health, 29.9% said that they were exergamers. Of the 851 students who reported that they did not stress about their health, 20.8% said that they were exergamers. Analyze these data using logistic regression and write a summary of your analytical approach, your results, and your conclusions.

## 14.23 More exergaming in Canada.

Refer to the previous exercise. Another explanatory variable reported in this study was the amount of television watched per day. Of the 54 students who reported that they watched no TV, 11.1% were exergamers; for the 776 students who watched some TV but less than two hours, 20.6% were exergamers; and for the 370 students who watched two or more hours, 31.1% were exergamers. Use logistic regression to examine the relationship between TV watching and exergaming. Write a summary of your analytical approach, your results, and your conclusions.

## 14.24 What's wrong?

For each of the following, explain what is wrong and why.

- (a) If  $b_1 = 5$  in a logistic regression analysis with one explanatory variable, we estimate that the probability of an event is multiplied by 5 when the value of the explanatory variable increases by 1 unit.
- (b) The intercept  $\beta_0$  is equal to the odds of an event when  $x = 0$ .
- (c) The odds of an event are 1 minus the probability of the event.

## 14.25 What's wrong?

For each of the following, explain what is wrong and why.

- (a) For a multiple logistic regression with 4 explanatory variables, the null hypothesis that the regression coefficients of all the explanatory variables are zero is tested with an  $F$  test.
- (b) For a logistic regression we assume that the model has a Normally distributed error term.
- (c) In logistic regression with one explanatory variable we can use a chi-square statistic to test the null hypothesis  $H_0: b_1 = 0$  versus a one-sided alternative.
- (d) In multiple logistic regression we do not need to worry about correlation among explanatory variables when interpreting model coefficient estimates.

## 14.26 Interpret the fitted model.

If we apply the exponential function to the fitted model in Example 14.6 (page 14-8), we get

$$\text{odds} = e^{-11.0391 + 3.1709x} = e^{-11.0391} \times e^{3.1709x}$$

Show that for any value of the quantitative explanatory variable  $x$ , the odds ratio for increasing  $x$  by 1,

$$\frac{\text{odds}_{x+1}}{\text{odds}_x}$$

is  $e^{3.1709} = 23.83$ . This justifies the interpretation given at the end of Example 14.6.

### 14.27 Will a movie be profitable?

In Example 14.6 (page 14-8), we developed a model to predict whether a movie is profitable based on log opening-weekend revenue. What are the predicted odds of a movie being profitable if the opening-weekend revenue is

- (a) \$25 million dollars ( $L_{\text{Opening}} = 3.219$ )?
- (b) \$45 million dollars ( $L_{\text{Opening}} = 3.807$ )?
- (c) \$65 million dollars ( $L_{\text{Opening}} = 4.174$ )?

### 14.28 Converting odds to probability.

Refer to the previous exercise. For each opening-weekend revenue, compute the estimated probability that the movie is profitable.

### 14.29 Salt intake and cardiovascular disease.

In Example 9.14 (page 549), the relative risk of developing cardiovascular disease (CVD) for people with low- and high-salt diets was estimated. Let's reanalyze these data using the methods in this chapter. Here are the data:

Developed CVD	Salt in Diet		Total
	Low	High	
Yes	88	112	200
No	1081	1134	2215
Total	1169	1246	2415

- (a) For each salt level find the probability of developing CVD.
- (b) Convert each of the probabilities that you found in part (a) to odds.
- (c) Find the log of each of the odds that you found in part (b).

### 14.30 Salt in the diet and CVD.

Refer to the previous exercise. Use  $x = 1$  for the high-salt diet and  $x = 0$  for the low-salt diet.

- (a) Find the estimates  $b_0$  and  $b_1$ .
- (b) Give the fitted logistic regression model.
- (c) What is the odds ratio for high-salt versus low-salt diet?

(d) When the probability of an event is very small, the odds ratio and relative risk are similar. Compare this odds ratio with the relative risk estimate in Example 9.14. Are they close? Explain your answer.

### 14.31 Give a 99% confidence interval for $\beta_1$ .

Refer to Example 14.9 (page 14-14). Suppose that you wanted to report a 99% confidence interval for  $\beta_1$ . Show how you would use the information provided in the outputs shown in Figure 14.7 to compute this interval.  INSECTS

### 14.32 Give a 95% confidence interval for the odds ratio.

Refer to Example 14.9 and the outputs in Figure 14.7 (page 14-14). Using the estimate  $b_1$  and its standard error, find the 95% confidence interval for the odds ratio and verify that this agrees with the interval given by the software.  INSECTS

### 14.33 $z$ and the $X^2$ statistic.

The Minitab output in Figure 14.7 (page 14-14) does not give the value of  $X^2$ . The column labeled “Z” provides similar information.  INSECTS

(a) Find the value under the heading “Z” for the predictor Lconc. Verify that Z is simply the estimated coefficient divided by its standard error. This is a  $z$  statistic that has approximately the standard Normal distribution if the null hypothesis (slope 0) is true.

(b) Show that the square of  $z$  is close to  $X^2$  (with no roundoff error, these two quantities will be equal). The two-sided  $P$ -value for  $z$  is the same as  $P$  for  $X^2$ .

(c) Draw sketches of the standard Normal distribution and the chi-square distribution with 1 degree of freedom. (*Hint*: You can use the information in Table F to sketch the chi-square distribution.) Indicate the value of the  $z$  and the  $X^2$  statistics on these sketches and use shading to illustrate the  $P$ -value.

### 14.34 Finding the best model?

In Example 14.11 (page 14-17), we looked at a multiple logistic regression for movie profitability based on three explanatory variables. Complete the analysis by looking at the 3 models that include two explanatory variables and the 3 models that include only one variable. Create a table that includes the parameter estimates and their  $P$ -values as well as the overall  $X^2$  statistic and degrees of freedom. Based on the results, which model do you feel is the best? Explain your answer.  MOVIES

### 14.35 Tipping behavior in Canada.

The Consumer Report on Eating Share Trends (CREST) contains data from all provinces of Canada detailing away-from-home food purchases by roughly 4000 households per quarter. Researchers recently restricted their attention to restaurants at which tips would normally be given.<sup>5</sup> From a total

of 73,822 observations, “high” and “low” tipping variables were created based on whether the observed tip rate was above 20% or below 10%, respectively. They then used logistic regression to identify explanatory variables associated with either “high” or “low” tips. The following table summarizes what they termed the stereotype-related variables for the low-tip analysis.

Explanatory variable	Odds ratio
Senior adult	1.099
Sunday	1.098
English as second language	1.142
French-speaking Canadian	1.163
Alcoholic drinks	0.713
Lone male	0.858

All coefficients were significant at the 0.01 level. Write a short summary explaining these results in terms of the odds of leaving a low tip.

### 14.36 What purchases will be made?

A poll of 1000 adults aged 18 or older asked about purchases they intended to make for the upcoming holiday season.<sup>6</sup> A total of 463 adults listed gift card as a planned purchase.

- What proportion of adults plan to purchase a gift card as a present?
- What are the odds that an adult will purchase a gift card as a present?
- What proportion of adults do not plan to purchase a gift card as a present?
- What are the odds that an adult will not buy a gift card as a present?
- How are your answers to parts (b) and (d) related?

### 14.37 High blood pressure and cardiovascular disease.

There is much evidence that high blood pressure is associated with increased risk of death from cardiovascular disease. A major study of this association examined 3338 men with high blood pressure and 2676 men with low blood pressure. During the period of the study, 21 men in the low-blood-pressure group and 55 in the high-blood-pressure group died from cardiovascular disease.

- Find the proportion of men who died from cardiovascular disease in the high-blood-pressure group. Then calculate the odds.
- Do the same for the low-blood-pressure group.
- Now calculate the odds ratio with the odds for the high-blood-pressure group in the numerator. Describe the result in words.

### 14.38 High blood pressure and cardiovascular disease.

Refer to the study of cardiovascular disease and blood pressure in Exercise 14.37. Computer output for a logistic regression analysis of these data gives the estimated slope  $b_1 = 0.7505$  with standard error  $SE_{b_1} = 0.2578$ .

- Give a 95% confidence interval for the slope.

- (b) Calculate the  $\chi^2$  statistic for testing the null hypothesis that the slope is zero and use Table F to find an approximate  $P$ -value.
- (c) Write a short summary of your results and conclusions.

### 14.39 High blood pressure and cardiovascular disease.

The results describing the relationship between blood pressure and cardiovascular disease are given in terms of the change in log odds in Exercise 14.38.

- (a) Transform the slope to the odds ratio and the 95% confidence interval for the slope to a 95% confidence interval for the odds ratio.
- (b) Write a conclusion using the odds to describe the results.

### 14.40 An example of Simpson's paradox.

Here is an example of Simpson's paradox, *the reversal of the direction of a comparison or an association when data from several groups are combined to form a single group*. The data concern two hospitals, A and B, and whether or not patients undergoing surgery died or survived. Here are the data for all patients:

	Hospital A	Hospital B
Died	63	16
Survived	2037	784
Total	2100	800

And here are the more detailed data where the patients are categorized as being in good condition or poor condition:

Good condition		
	Hospital A	Hospital B
Died	6	8
Survived	594	592
Total	600	600

Poor condition		
	Hospital A	Hospital B
Died	57	8
Survived	1443	192
Total	1500	200

- (a) Use a logistic regression to model the odds of death with hospital as the explanatory variable. Summarize the results of your analysis and give a 95% confidence interval for the odds ratio of Hospital A relative to Hospital B.
- (b) Rerun your analysis in part (a) using hospital and the condition of the patient as explanatory variables. Summarize the results of your analysis and give a 95% confidence interval for the odds ratio of Hospital A relative to Hospital B.
- (c) Explain Simpson's paradox in terms of your results in parts (a) and (b).

### 14.41 Reducing the number of workers.

To be competitive in global markets, many corporations are undertaking major reorganizations. Often these involve “downsizing” or a “reduction in force” (RIF), where substantial numbers of employees are terminated. Federal and various state laws require that employees be treated equally regardless of their age. In particular, employees over the age of 40 years are in a “protected” class, and many allegations of discrimination focus on comparing employees over 40 with their younger coworkers. Here are the data for a recent RIF:

Terminated	Over 40	
	No	Yes
Yes	7	41
No	504	765

- (a) Write the logistic regression model for this problem using the log odds of a RIF as the response variable and an indicator for over and under 40 years of age as the explanatory variable.
- (b) Explain the assumption concerning binomial distributions in terms of the variables in this exercise. To what extent do you think that these assumptions are reasonable?
- (c) Software gives the estimated slope  $b_1 = 1.3504$  and its standard error  $SE_{b_1} = 0.4130$ . Transform the results to the odds scale. Summarize the results and write a short conclusion.
- (d) If additional explanatory variables were available, for example, a performance evaluation, how would you use this information to study the RIF?

### 14.42 Internet use in Canada.

A recent study used data from the Canadian Internet Use Survey (CIUS) to explore the relationship between certain demographic variables and Internet use by individuals in Canada.<sup>7</sup> The response variable refers to the use of the Internet from any location within the last 12 months. Explanatory variables included age (years), income (thousands of dollars), location (1 = urban, 0 = other), sex (1 = male, 0 = female), education (1 = at least some postsecondary education, 0 = other), language (1 = English, 0 = French), and children (1 = at least one child in household, 0 = no children). The following table summarizes the results.

Explanatory variable	$b$
Age	-0.063
Income	0.013
Location	0.367
Sex	-0.222
Education	1.080
Language	0.285
Children	0.049
Intercept	2.010

All but Children were significant at the 0.05 level.

- (a) Interpret the sign of each of the coefficients (except the intercept) in terms of the probability that the individual uses the Internet.
- (b) Compute the odds ratio for each of the variables in the table.

- (c) What are the odds that a French-speaking, 23-year-old male, living alone in Montreal, and making \$50,000 a year his second year after college is using the Internet?
- (d) Convert the odds in part (c) to a probability.

### 14.43 Predicting physical activity.

Participation in physical activities typically declines between high school and young adulthood. This suggests that postsecondary institutions may be an ideal setting to address physical activity. A study looked at the association between physical activity and several behavioral and perceptual characteristics among midwestern college students.<sup>8</sup> Of 663 students who met the vigorous activity guidelines for the previous week, 169 reported eating fruit two or more times per day. Of the 471 that did not meet the vigorous activity guidelines in the previous week, 68 reported eating fruit two or more times per day. Model the log odds of vigorous activity using an indicator variable for eating fruit two or more times per day as the explanatory variable. Summarize your findings.

### 14.44 Online consumer spending.

The Consumer Behavior Report is designed to provide insight into online shopping trends.<sup>9</sup> A recent report asked the question “In the past three months, how has the current state of the economy impacted your money spending on online purchasing?” Here are the results from 3156 online consumers:

Gender	Reduced Spending	
	No	Yes
Female	586	708
Male	1074	788

- (a) What proportion of individuals reduced their spending in each gender?
- (b) What is the odds ratio for comparing female individuals to male individuals?
- (c) Write the logistic regression model for this problem using the log odds of reducing spending as the response variable and an indicator of gender as the explanatory variable.
- (d) Software gives the estimated slope  $b_1 = 0.4988$  and its standard error  $SE_{b_1} = 0.0729$ . Transform this result to the odds scale and compare it with your answer in part (b).
- (e) Construct a 95% confidence interval for the odds ratio and write a short conclusion.

### 14.45 Proximity of fast-food restaurants to schools and adolescent overweight.

A California study looked at the relationship between fast-food restaurants near schools (within a 0.5-mile radius) and overweight among middle and high school students.<sup>10</sup> Overweight was determined based on each student’s responses to the California Healthy Kids Survey (CHKS). A database of latitude-longitude coordinates for schools and restaurants was used to determine proximity. Here are the data:

Fast-food nearby	$n$	$X(\text{overweight})$
No	238,215	65,080

Use logistic regression to study the question of whether or not overweight is related to the proximity of fast-food restaurants to schools. Write a short paragraph summarizing your conclusions.

### 14.46 Overweight and fast-food restaurants, continued.

Refer to the previous exercise. In the article, the researchers state (1) “CIs were adjusted for clustering at the school level,” and (2) “All models also included controls for the following student characteristics: a female indicator, grade indicator, age indicator, race/ethnicity indicators, and physical exercise indicators. All models also included indicator variables for school location type, including large urban, midsize urban, small urban, large suburban, midsize suburban, small suburban, town, and rural.”

(a) What violation of the distribution of the response variable is Statement 1 addressing? Explain your answer.

(b) Explain why the researchers controlled for the variables described in Statement 2 when looking at the relationship between overweight and proximity.

*The following four exercises use the GPAHI data file. We examine models for relating success as measured by the GPA to several explanatory variables. In Chapter 11 we used multiple regression methods for our analysis. Here, we define an indicator variable, HIGPA, to be 1 if the GPA is 3.0 or better and 0 otherwise.*  **GPAHI**

### 14.47 Use high school grades to predict high grade point averages.

Use a logistic regression to predict HIGPA using the three high school grade summaries as explanatory variables.  **GPAHI**

(a) Summarize the results of the hypothesis test that the coefficients for all three explanatory variables are zero.

(b) Give the coefficient for high school math grades with a 95% confidence interval. Do the same for the two other predictors in this model.

(c) Summarize your conclusions based on parts (a) and (b).

### 14.48 Use SAT scores to predict high grade point averages.

Use a logistic regression to predict HIGPA using the SATM and SATCR scores as explanatory variables.

(a) Summarize the results of the hypothesis test that the coefficients for both explanatory variables are zero.

(b) Give the coefficient for the SATM score with a 95% confidence interval. Do the same for the SATCR score.

(c) Summarize your conclusions based on parts (a) and (b).

### 14.49 Use high school grades and SAT scores to predict high grade point averages.

Run a logistic regression to predict HIGPA using the three high school grade summaries and the two SAT scores as explanatory variables. We want to produce an analysis that is similar to that done for the case study in Chapter 11.  GPAHI

- (a) Test the null hypothesis that the coefficients of the three high school grade summaries are zero; that is, test  $H_0 : \beta_{HSM} = \beta_{HSS} \beta_{HSE} = 0$ .
- (b) Test the null hypothesis that the coefficients of the two SAT scores are zero; that is, test  $H_0 : \beta_{SATM} = \beta_{SATCR} = 0$ .
- (c) What do you conclude from the tests in (a) and (b)?

### 14.50 Is there an effect of gender?

In this exercise we investigate the effect of gender on the odds of getting a high GPA.  GPAHI

- (a) Use gender to predict HIGPA using a logistic regression. Summarize the results.
- (b) Perform a logistic regression using gender and the two SAT scores to predict HIGPA. Summarize the results.
- (c) Compare the results of parts (a) and (b) with respect to how gender relates to HIGPA. Summarize your conclusions.

## CHAPTER 14 Notes and Data Sources

1. Logistic regression models for the general case where there are more than two possible values for the response variable have been developed. These are considerably more complicated and are beyond the scope of our present study. For more information on logistic regression, see A. Agresti, *An Introduction to Categorical Data Analysis*, 2nd ed., Wiley, 2007; and D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd ed., Wiley, 2000.
2. This example is taken from a classic text written by a contemporary of R. A. Fisher, the person who developed many of the fundamental ideas of statistical inference that we use today. The reference is D. J. Finney, *Probit Analysis*, Cambridge University Press, 1947. Although not included in the analysis, it is important to note that the experiment included a control group that received no insecticide. No aphids died in this group. We have chosen to call the response “dead.” In Finney’s book the category is described as “apparently dead, moribund, or so badly affected as to be unable to walk more than a few steps.” This is an early example of the need to make careful judgments when defining variables to be used in a statistical analysis. An insect that is “unable to walk more than a few steps” is unlikely to eat very much of a chrysanthemum plant!
3. See [pewinternet.org/Reports/2013/in-store-mobile-commerce.aspx](http://pewinternet.org/Reports/2013/in-store-mobile-commerce.aspx).
4. Erin K. O’Loughlin et al., “Prevalence and correlates of exergaming in youth,” *Pediatrics*, 130 (2012), pp. 806–814.
5. Based on Leigh J. Maynard and Malvern Mupandawana, “Tipping behavior in Canadian restaurants,” *International Journal of Hospitality Management*, 28 (2009), pp. 597–603.
6. These results are from the Consumer Reports National Research Center, which conducted a telephone survey of a nationally representative probability sample of households with telephones. One thousand interviews were completed among adults aged at least 18 years. Interviewing took place on October 15–18, 2009.
7. Anthony A. Noce and Larry McKeown, “A new benchmark for Internet use: A logistic modeling of factors influencing Internet use in Canada, 2005,” *Government Information Quarterly*, 25 (2008), pp. 462–476.
8. Dong-Chul Seo et al., “Relations between physical activity and behavioral and perceptual correlates among midwestern college students,” *Journal of American College Health*, 56 (2007), pp. 187–197.
9. These economic trend reports can be found at [mr.pricegrabber.com](http://mr.pricegrabber.com). These results are based on the June 2009 report.
10. Brennan Davis and Christopher Carpenter, “Proximity of fast-food restaurants to schools and adolescent obesity,” *American Journal of Public Health*, 99 (2009), pp. 505–510.

# 15 Nonparametric Tests

CHAPTER



**15.1 The Wilcoxon Rank Sum Test**

**15.2 The Wilcoxon Signed Rank Test**

**15.3 The Kruskal-Wallis Test**

## Introduction

The most commonly used methods for inference about the means of quantitative response variables assume that the variables in question have Normal distributions in the population or populations from which we draw our data. In practice, of course, no distribution is exactly Normal. Fortunately, our usual methods for inference about population means (the one-sample and two-sample  $t$  procedures and analysis of variance) are quite robust. That is, the results of inference are not very sensitive to moderate lack of Normality, especially when the samples are reasonably large. Some practical guidelines for taking advantage of the *robustness* of these methods appear in Chapter 7.

### robustness

What can we do if plots suggest that the population distribution is clearly not Normal, especially when we have only a few observations? This is not a simple question. Here are the basic options:

1. If lack of Normality is due to *outliers*, it may be legitimate to remove the outliers. An outlier is an observation that may not come from the same population as the other observations. Equipment failure that produced a bad measurement, for example, entitles you to remove the outlier and analyze the remaining data. If the outlier appears to be “real data,” you can base inference on statistics that are more resistant than  $\bar{x}$  and  $s$ . Options 4 and 5 allow this.

### outlier

2. Sometimes we can **transform** our data so that their distribution is more nearly Normal. Transformations such as the logarithm that pull in the long tail of right-skewed distributions are particularly helpful. Example 7.10 (page 436) illustrates use of the logarithm.



transformations, p. 93

3. In some settings, *other standard distributions* replace the Normal distributions as models for the overall pattern in the population. We mentioned in Chapter 5 (page 315) that the Weibull distributions are common models for the lifetimes in service of equipment in statistical studies of reliability. Also, we studied the exponential distributions (page 309) and the Poisson distributions (page 339) in Chapter 5. There are inference procedures for the parameters of these distributions that replace the  $t$  procedures when we use specific non-Normal models.

## other standard distributions

4. Modern **bootstrap methods** and **permutation tests** do not require Normality or any other specific form of sampling distribution. Moreover, you can base inference on resistant statistics such as the trimmed mean. We recommend these methods unless the sample is so small that it may not represent the population well. Chapter 16 gives a full discussion.

## bootstrap methods

## permutation tests

5. Finally, there are other **nonparametric methods** that do not require any specific form for the distribution of the population. Unlike bootstrap and permutation methods, common nonparametric methods do not make use of the actual values of the observations. The *sign test* (page 438) works with *counts* of observations. This chapter presents **rank tests** based on the *rank* (place in order) of each observation in the set of all the data.

## nonparametric methods

## rank tests

This chapter concerns rank tests that are designed to replace the  $t$  tests and one-way analysis of variance when the Normality conditions for those tests are not met. Figure 15.1 presents an outline of the standard tests (based on Normal distributions) and the rank tests that compete with them.

The rank tests we will study concern the *center* of a population or populations. When a population has at least roughly a Normal distribution, we describe its center by the mean. The “Normal tests” in Figure 15.1 test hypotheses about population means. When distributions are strongly skewed, we often prefer the median to the mean as a measure of center. In simplest form, the hypotheses for rank tests just replace mean by median.

**FIGURE 15.1**

Comparison of tests based on Normal distributions with nonparametric tests for similar settings.

Setting	Normal test	Rank test
One sample	One-sample $t$ test Section 7.1	Wilcoxon signed rank test Section 15.2
Matched pairs	Apply one-sample test to differences within pairs	
Two independent samples	Two-sample $t$ test Section 7.2	Wilcoxon rank sum test Section 15.1
Several independent samples	One-way ANOVA $F$ test Chapter 12	Kruskal-Wallis test Section 15.3

We devote a section of this chapter to each of the rank procedures. Section 15.1, which discusses the most common of these tests, also contains general information about rank tests. The kind of assumptions required, the nature of the hypotheses tested, the big idea of using ranks, and the contrast between exact distributions for use with small samples and approximations for use with larger samples are common to all rank tests. Sections 15.2 and 15.3 more briefly describe other rank tests.

## 15.1 The Wilcoxon Rank Sum Test

When you complete this section, you will be able to

- Find the rank transformation for a set of data.
- Compute the Wilcoxon rank sum statistic for the comparison of two populations.
- State the null and alternative hypotheses that are used for the analysis of data using the Wilcoxon rank sum test.
- Use the two sample sizes to find the mean and the standard deviation of the sampling distribution of the Wilcoxon rank sum statistic under the null hypothesis.
- Find the  $P$ -value for the Wilcoxon rank sum significance test using the Normal approximation with the continuity correction.
- For the Wilcoxon rank sum test, use computer output to determine the results of the significance test.

Two-sample problems (see Section 7.2) are among the most common in statistics. The most useful nonparametric significance test compares two distributions. Here is an example of this setting.



two sample, p. 447

### Example

#### 15.1 Does the American League get more hits?

In 1973, the American League adopted the designated-hitter rule, which allows a substitute player to take the place of the pitcher when it is the pitcher's turn to bat. Since pitchers typically do not hit as well as other players, it was expected that the rule would produce more hits and therefore more excitement for the fans. The National League has not adopted this rule. Let's look at some data to see if we can detect a difference in hits between the American League

and the National League. Here are the number of hits for eight games played on the same spring day, four from each league.



League	Hits
American	21 18 24 20
National	19 7 11 13

The samples are too small to assess Normality adequately or to rely on the robustness of the  $t$  test. We prefer to use a test that does not require Normality.

## The rank transformation

We first rank all eight observations together. To do this, arrange them in order from smallest to largest:

7 11 13 **18** 19 **20** **21** **24**

The boldface entries in the list are the hits for the American League. The idea of rank tests is to look just at position in this ordered list. To do this, replace each observation by its order, from 1 (smallest) to 8 (largest). These numbers are the *ranks*:

Runs	7	11	13	<b>18</b>	19	<b>20</b>	<b>21</b>	<b>24</b>
Rank	1	2	3	<b>4</b>	5	<b>6</b>	<b>7</b>	<b>8</b>

It would not be surprising if we had sampled a day where more than one game had the same number of hits. We will discuss how to handle ties later in this section.

## RANKS

To rank observations, first arrange them in order from smallest to largest. The **rank** of each observation is its position in this ordered list, starting with rank 1 for the smallest observation.

Moving from the original observations to their ranks is a transformation of the data, like moving from the observations to their logarithms. The rank transformation retains only the ordering of the observations and makes no other use of their numerical values. Working with ranks allows us to dispense with specific assumptions about the shape of the distribution, such as Normality.

## USE YOUR KNOWLEDGE

### 15.1 Numbers of rooms in top spas.

A report of a readers' poll in *Condé Nast Traveler* magazine ranked 100 top resort spas.<sup>1</sup> Let Group A be the 25 top-ranked spas, and let Group B be the spas ranked 26 to 50. A simple random sample of size 5 was taken from each group, and the number of rooms in each selected spa was recorded. Here are the data:



Group A	106	145	312	60	49
Group B	190	500	1293	161	225

Rank all the observations together and make a list of the ranks for Group A and Group B.



### 15.2 The effect of Animal Kingdom on the result.

Refer to the previous exercise. Disney's Animal Kingdom in Lake Buena Vista, Florida, with 1293 rooms, was the third spa selected in Group B. Suppose, instead, a different spa, with 540 rooms, had been selected. Replace the observation 1293 in Group B by 540. Use the modified data to make a list of the ranks for Groups A and B combined. What changes?



## The Wilcoxon rank sum test

If the American League games tend to have more hits than the National League, we expect the ranks of the American League games to be higher than those for the National League games. Let's compare the *sums* of the ranks from the two treatments:

League	Sum of ranks
American	25
National	11

These sums compare the hits of the American League with those of the National League. In fact, the sum of the ranks from 1 to 8 is always equal to 36, so it is enough to report the sum for one of the two groups.

If the sum of the ranks for the American League is 25, then the ranks for the National League must be 11 because  $25 + 11 = 36$ . If there was no difference between the leagues, we would expect the sum of the ranks for each league to be 18 (half of 36). Here are the facts we need in a more general form that takes account of the fact that our two samples need not be the same size.

### THE WILCOXON RANK SUM TEST

Draw an SRS of size  $n_1$  from one population and draw an independent SRS of size  $n_2$  from a second population. There are  $N$  observations in all, where  $N = n_1 + n_2$ . Rank all  $N$  observations. The sum  $W$  of the ranks for the first sample is the **Wilcoxon rank sum statistic**. If the two populations have the same continuous distribution, then  $W$  has mean

$$\mu_W = n_1 \left( \frac{N+1}{2} \right)$$

and standard deviation

$$\sigma_W = \sqrt{n_1 n_2 (N+1) / 12}$$

The **Wilcoxon rank sum test** rejects the hypothesis that the two populations have identical distributions when the rank sum  $W$  is far from its mean.\* This test is also called the **Mann-Whitney test**.

\* This test was invented by Frank Wilcoxon (1892–1965) in 1945. Wilcoxon was a chemist who encountered statistical problems in his work at the research laboratories of American Cyanamid Company.

For the baseball question of Example 15.1, we want to test

$H_0$ : no difference in number of hits

against the one-sided alternative

$H_a$ : more hits are made in American League games than in National League games

Our test statistic is the rank sum  $W = 25$  for the American League games.

## USE YOUR KNOWLEDGE

### 15.3 Hypotheses and test statistic for top spas.

Refer to Exercise 15.1. State appropriate null and alternative hypotheses for this setting and calculate the value of  $W$ , the test statistic.



### 15.4 Effect of Animal Kingdom on the test statistic.

Refer to Exercise 15.2. Using the altered data, state appropriate null and alternative hypotheses and calculate the value of  $W$ , the test statistic.



## Example

## 15.2 Perform the significance test.

In Example 15.1,  $n_1 = 4$ ,  $n_2 = 4$ , and there are  $N = 8$  observations in all. The sum of ranks for the American League games has mean

$$\begin{aligned}\mu_W &= n_1(N+1)/2 \\ &= (4)(9)/2 = 18\end{aligned}$$

and standard deviation

$$\begin{aligned}\sigma_W &= \sqrt{n_1 n_2 (N+1)/12} \\ &= \sqrt{(4)(4)(9)/12} = 3.464\end{aligned}$$

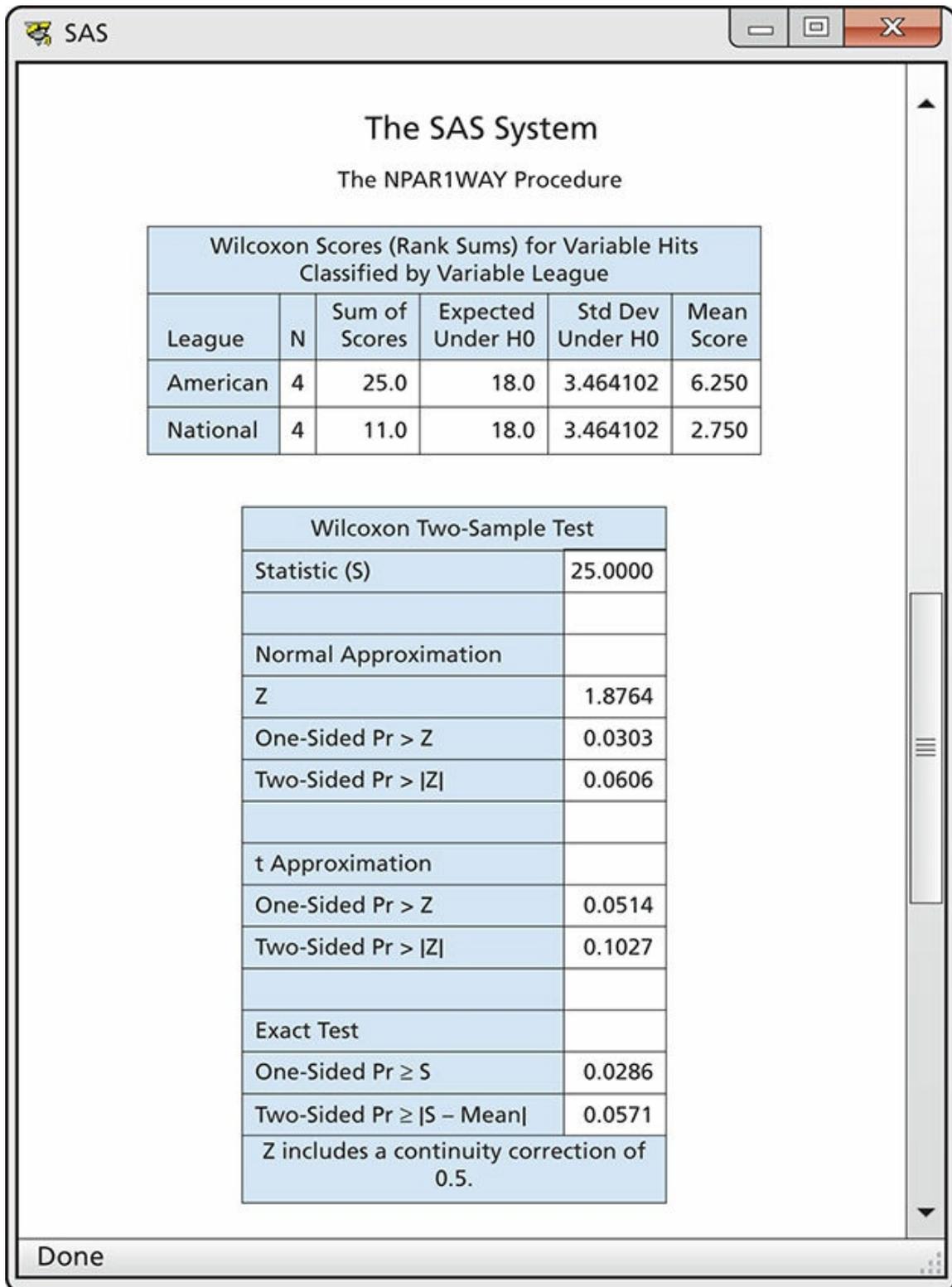
The observed sum of the ranks,  $W = 25$ , is higher than the mean, about 2 standard deviations higher  $([25 - 18]/3.464)$ . It appears that the data support our idea that American League games have more hits than National League games. The  $P$ -value for our one-sided alternative is  $P(W \geq 25)$ , the probability that  $W$  is at least as large as the value for our data when  $H_0$  is true.

To calculate the  $P$ -value  $P(W \geq 25)$ , we need to know the sampling distribution of the rank sum  $W$  when the null hypothesis is true. This distribution depends on the two sample sizes  $n_1$  and  $n_2$ . Tables are therefore a bit unwieldy, though you can find them in handbooks of statistical tables. Most statistical software will give you  $P$ -values, as well as carry out the ranking and calculate  $W$ . However, some software gives only approximate  $P$ -values. You must learn what your software offers.

## Example

### 15.3 Software output.

Figure 15.2 shows the output from software that calculates the exact sampling distribution of  $W$ . We see that the sum of the ranks (called scores in the output) for the American League is  $W = 25$ , with  $P$ -value  $P = 0.0286$  against the one-sided alternative that American League games have more hits than the National League games.



**FIGURE 15.2**

Output from SAS for the baseball hit data, for Example 15.3.

It is worth noting that the two-sample  $t$  test for the one-sided alternative gives essentially the same result as the Wilcoxon test in Example 15.3 ( $t = 2.95$ ,  $P = 0.016$ ).



two-sample  $t$  test, p. 454

## The Normal approximation

The rank sum statistic  $W$  becomes approximately Normal as the two sample sizes increase. We can then form yet another  $z$  statistic by standardizing  $W$ :

$$\begin{aligned} z &= \frac{W - \mu_W}{\sigma_W} \\ &= \frac{W - n_1(N+1)/2}{\sqrt{n_1 n_2 (N+1)/12}} \end{aligned}$$

Use standard Normal probability calculations to find  $P$ -values for this statistic. Because  $W$  takes only whole-number values, the continuity correction improves the accuracy of the approximation.



continuity correction, p. 335

### Example

#### 15.4 The continuity correction.

The standardized rank sum statistic  $W$  in our baseball example is

$$z = \frac{W - \mu_W}{\sigma_W} = \frac{25 - 183.464}{1.876} = 2.02$$

We expect  $W$  to be larger when the alternative hypothesis is true, so the approximate  $P$ -value is

$$P(Z \geq 2.02) = 0.0217$$

The continuity correction acts as if the whole number 25 occupies the entire interval from 24.5 to 25.5. We calculate the  $P$ -value  $P(W \geq 25)$  as  $P(W \geq 24.5)$  because the value 25 is included in the range whose probability we want. Here is the calculation:

$$\begin{aligned} P(W \geq 24.5) &= P\left(\frac{W - \mu_W}{\sigma_W} \geq \frac{24.5 - 183.464}{1.876}\right) \\ &= P(Z \geq 1.876) \\ &= 0.303 \end{aligned}$$

The continuity correction gives a result closer to the exact value  $P = 0.0286$  (see Figure 15.2 ).

## USE YOUR KNOWLEDGE

### 15.5 The $P$ -value for top spas.

Refer to Exercises 15.1 and 15.3 (pages 15-4 and 15-6). Find  $\mu_W$ ,  $\sigma_W$ , and the standardized rank sum statistic. Then give an approximate  $P$ -value using the Normal approximation. What do you conclude?



### 15.6 The effect of Animal Kingdom on the $P$ -value.

Refer to Exercises 15.2 and 15.4 (pages 15-4 and 15-6). Repeat the analysis in Exercise 15.5 using the altered data.



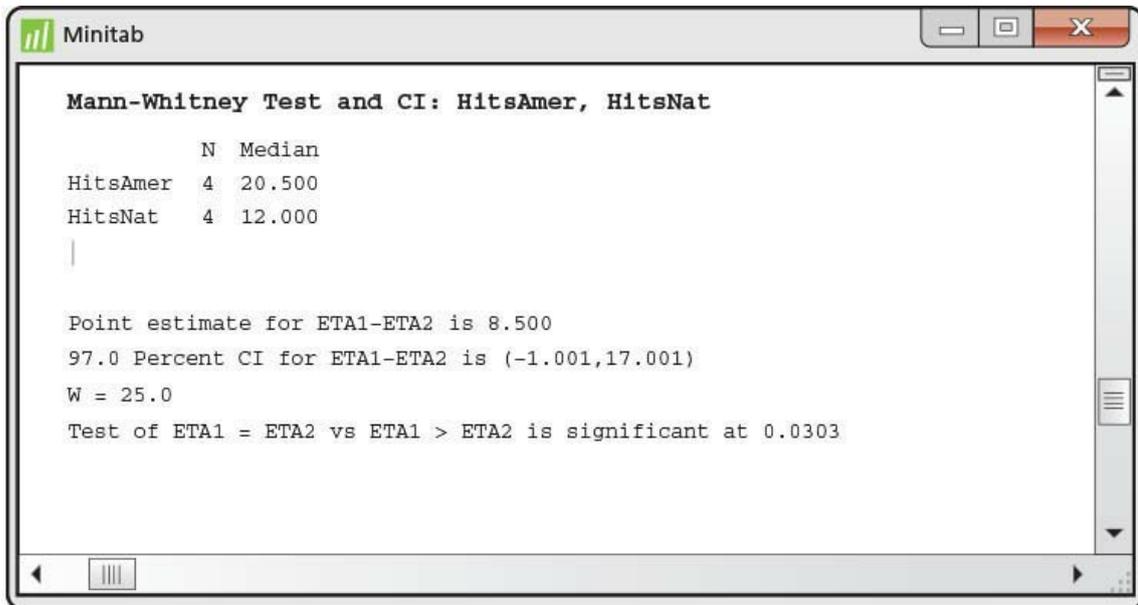
We recommend always using either the exact distribution (from software or tables) or the continuity correction for the rank sum statistic  $W$ . The exact distribution is safer for small samples. As Example 15.4 illustrates, however, the Normal approximation with the continuity correction is often adequate.

## Example

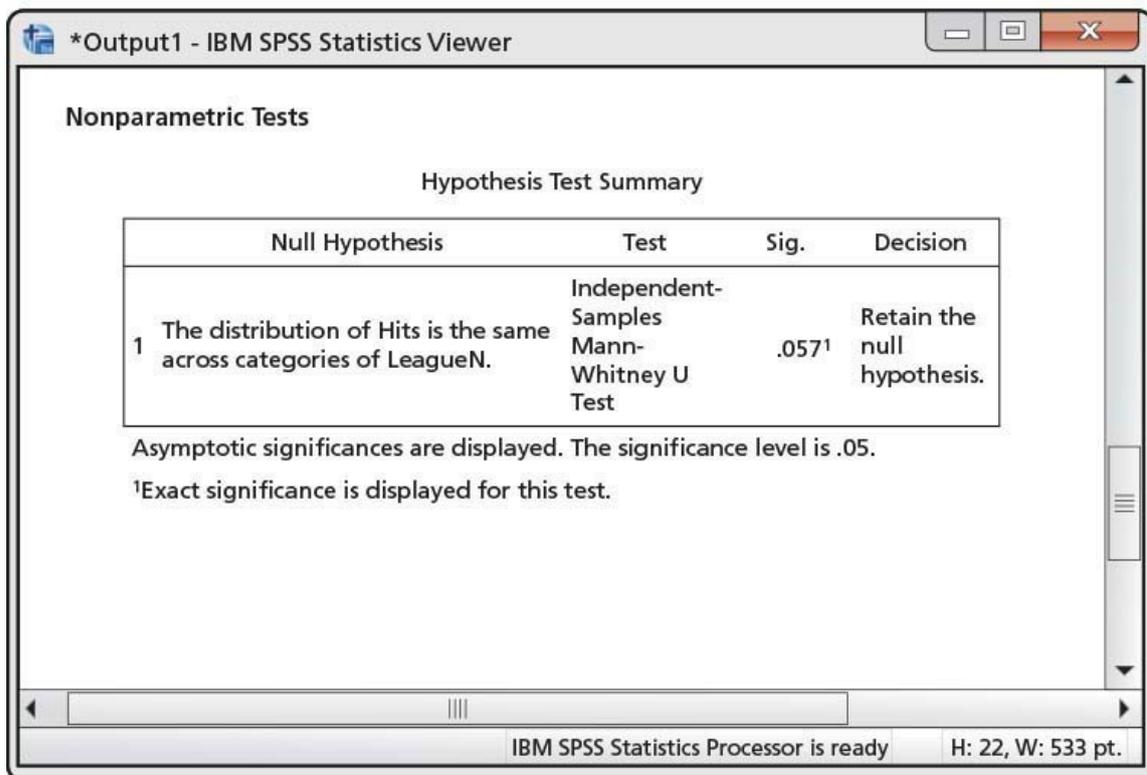
### 15.5 Software output.

Figure 15.3 shows the output for our data from two additional statistical programs. Minitab gives the Normal approximation, and it refers to the *Mann-Whitney test*. This is an alternative form of the Wilcoxon rank sum test. SPSS uses the exact calculation for the *P*-value here but tests the null hypothesis only against the two-sided alternative.

### Mann-Whitney test



(a) Minitab



(b) SPSS

### FIGURE 15.3

Output from the Minitab and SPSS statistical software for the data in Example 15.1. (a) Minitab uses the Normal approximation for the distribution of  $W$ . (b) SPSS gives the exact value for the two-sided alternative.

## What hypotheses does Wilcoxon test?

Our null hypothesis is that the distribution of hits is the same in the two leagues. Our alternative hypothesis is that there are more hits in the American League than in the National League. If we are willing to assume that hits are Normally distributed, or if we have reasonably large samples, we use the two-sample  $t$  test for means. Our hypotheses then become

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 > \mu_2$$

When the distributions may not be Normal, we might restate the hypotheses in terms of population medians rather than means:

$$H_0: \text{median}_1 = \text{median}_2$$

$$H_a: \text{median}_1 > \text{median}_2$$

*The Wilcoxon rank sum test does test hypotheses about population medians, but only if an additional assumption is met: both populations must have distributions of the same shape.* That is, the density curve for hits in the American League must look exactly like that for the National League except that it may be shifted to the left or to the right. The Minitab output in Figure 15.3(a) states the hypotheses in terms of population medians (which it calls “ETA”) and also gives a confidence interval for the difference between the two population medians.



The same-shape assumption is too strict to be reasonable in practice. Recall that our preferred version of the two-sample  $t$  test does not require that the two populations have the same standard deviation—that is, it does not make a same-shape assumption. Fortunately, the Wilcoxon test also applies in a much more general and more useful setting. It tests hypotheses that we can state in words as

$$H_0: \text{The two distributions are the same.}$$

$$H_a: \text{One distribution has values that are systematically larger.}$$

Here is a more exact statement of the *systematically larger* alternative

hypothesis. Take  $X_1$  to be hits in the American League and  $X_2$  to be hits in the National League. These hits are random variables. That is, for each game in the American League, the number of hits is a value of the variable  $X_1$ . The probability that the number of hits is more than 15 is  $P(X_1 > 15)$ . Similarly,  $P(X_2 > 15)$  is the corresponding probability for the National League. If the number of American League hits is “systematically larger” than the number of National League hits, getting more hits than 15 should be more likely in the American League. That is, we should have

systematically larger

$$P(X_1 > 15) > P(X_2 > 15)$$

The alternative hypothesis says that this inequality holds not just for 15 hits but for *any* number of hits.<sup>2</sup>

This exact statement of the hypotheses we are testing is a bit awkward. The hypotheses really are “nonparametric” because they do not involve any specific parameter such as the mean or median. If the two distributions do have the same shape, the general hypotheses reduce to comparing medians. Many texts and computer outputs state the hypotheses in terms of medians, sometimes ignoring the same-shape requirement. We recommend that you express the hypotheses in words rather than symbols. “The number of American League hits is systematically higher than the number of National League hits” is easy to understand and is a good statement of the effect that the Wilcoxon test looks for.

## Ties

The exact distribution for the Wilcoxon rank sum is obtained assuming that all observations in both samples take different values. This allows us to rank them all. In practice, however, we often find observations tied at the same value. What shall we do? The usual practice is to *assign all tied values the average of the ranks they occupy*. Here is an example with six observations:

average ranks

Observation	153	155	158	158	161	164
Rank	1	2	3.5	3.5	5	6

The tied observations occupy the third and fourth places in the ordered list, so they share rank 3.5.

The exact distribution for the Wilcoxon rank sum  $W$  changes if the data contain ties. Moreover, the standard deviation  $\sigma_W$  must be adjusted if ties are present. The Normal approximation can be used after the standard deviation is adjusted. Statistical software will detect ties, make the necessary adjustment, and switch to

the Normal approximation. In practice, software is required if you want to use rank tests when the data contain tied values.

It is sometimes useful to use rank tests on data that have very many ties because the scale of measurement has only a few values. Here is an example.

## Example

### 15.6 Exergaming in Canada.

Exergames are active video games such as rhythmic dancing games, virtual bicycles, balance board simulators, and virtual sports simulators that require a screen and a console. A study of exergaming in students from grades 10 and 11 in Montreal, Canada, examined many factors related to participation in exergaming.<sup>3</sup> In Exercise 14.23 (page 14-22) we used logistic regression to examine the relationship between exergaming and time spent viewing television. Here are the data displayed in a two-way table of counts:



Exergamer	TV time (hours per day)		
	None	Some but less than 2 hours	2 hours or more
Yes	6	160	115
No	48	616	255



## USE YOUR KNOWLEDGE

### 15.7 Analyze as a two-way table.

Analyze the exergaming data in Example 15.6 as a two-way table.



(a) Compute the percents in the three categories of TV watching for the exergamers. Do the same for those who are not exergamers. Display the percents graphically and summarize the differences in the two distributions.



chi-square test, p. 539

(b) Perform the chi-square test for the counts in the two-way table. Report the test statistic, the degrees of freedom, and the  $P$ -value. Give a brief summary of what you can conclude from this significance test.

How do we approach the analysis of these data using the Wilcoxon test? We start with the hypotheses. We have two distributions of TV viewing, one for the exergamers and one for those who are not exergamers. The null hypothesis states that these two distributions are the same. The alternative hypothesis uses the fact that the responses are ordered from no TV to 2 hours or more per day. It states that one of the exerciser groups watches more TV than the other.

$H_0$ : The amount of time spent viewing TV is the same for students who are exergamers and students who are not.

$H_a$ : One of the two groups views more TV than the other.

The alternative hypothesis is two-sided. Because the responses can take only three values, there are very many ties. All 54 students who watch no TV are tied. Similarly, all students in each of the other two columns of the table are tied. The graphical display that you prepared in Exercise 15.7 suggests that the exergamers

watch more TV than those who are not exergamers. Is this difference statistically significant?

## Example

### 15.7 Software output.

Look at Figure 15.4, which gives SAS output for the Wilcoxon test. The rank sum for the exergamers (using average ranks for ties) is  $W = 187,747.5$ . The expected rank sum under the null hypothesis is 168,740.5, so the exergamers have a higher rank sum than we would expect. The Normal approximation test statistic is  $z = 4.47$  and the two-sided  $P$ -value is reported as  $P < 0.0001$ . There is very strong evidence of a difference. Exergamers watch more TV than the students who are not exergamers.



We can use our framework of “systematically larger” (page 15-10) to summarize these data. For the exergamers, 98% watch some TV and 41% watch two or more hours per day. The corresponding percents for the students who are not exergamers are 95% and 28%.

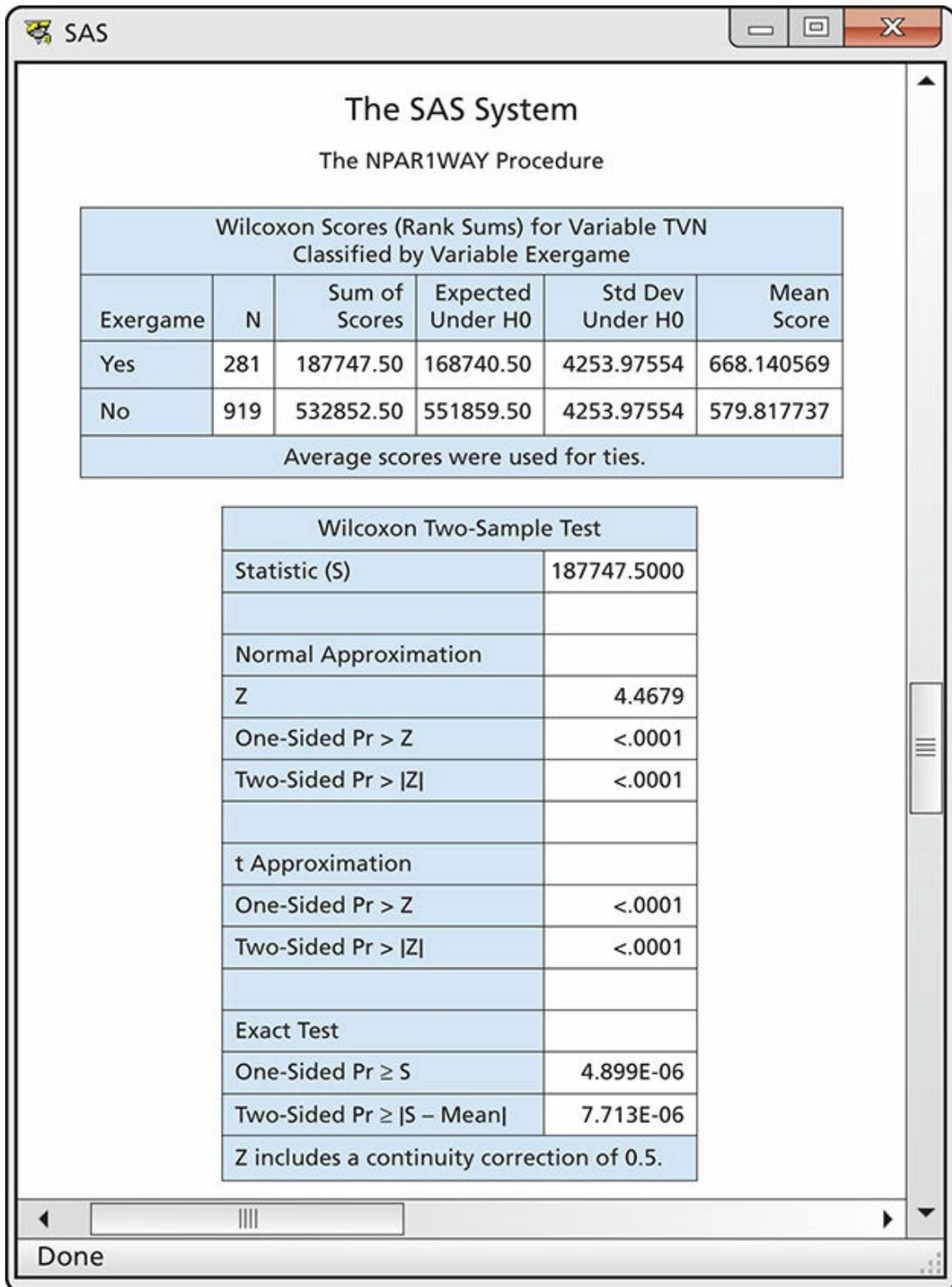
In our discussion of TV viewing and exergaming, we have expressed results in terms of the amount of TV watched. In fact, we do not have the actual hours of TV watched by each student in the study. Only data with the hours classified into three groups are available. Many government surveys summarize quantitative data categorized into ranges of values. *When summarizing the analysis of data, it is very important to explain clearly how the data are recorded.* In this setting, we have chosen to use phrases such as “watch more TV” because they express the findings based on the data available.



Note that the two-sample  $t$  test would not be appropriate in this setting. If we coded the TV-watching categories as 1, 2, and 3, the average of these coded values

would not be meaningful.

On the other hand, we frequently encounter variables measured in scales such as “strongly agree,” “agree,” “neither agree nor disagree,” “disagree,” and “strongly disagree.” In these circumstances, many would code the responses with the integers 1 to 5 and then use standard methods such as a *t* test or ANOVA. Whether to do this or not is a matter of judgment. Rank tests avoid the dilemma because they use only the order of the responses, not their actual values. *Some statisticians use t procedures when there is not a fully meaningful scale of measurement, but others avoid them.*



**FIGURE 15.4**  
Output from SAS for the exergaming data, for Example 15.7.



## Rank, $t$ , and permutation tests

The two-sample  $t$  procedures are the most common method for comparing the centers of two populations based on random samples from each. The Wilcoxon rank sum test is a competing procedure that does not start from the condition that the populations have Normal distributions. Permutation tests (Chapter 16) also avoid the need for Normality. Tests based on Normality, rank tests, and permutation tests apply in many other settings as well. How do these three approaches compare in general?

First, let's consider rank tests versus traditional tests based on Normal distributions. Both are available in almost all statistical software.

- Moving from the actual data values to their ranks allows us to find an exact sampling distribution for rank statistics such as the Wilcoxon rank sum  $W$  when the null hypothesis is true. (Most software will do this only if there are no ties and if the samples are quite small.) When our samples are small, are truly random samples from the populations, and show non-Normal distributions of the same shape, the Wilcoxon test is more reliable than the two-sample  $t$  test. In practice, the robustness of  $t$  procedures implies that we rarely encounter data that require nonparametric procedures to obtain reasonably accurate  $P$ -values. The  $t$  and  $W$  tests gave very similar results for the baseball hit data in Example 15.1, but we would not use a  $t$  procedure for the exergame data in Example 15.6.
- Normal tests compare means and are accompanied by simple confidence intervals for means or differences between means. When we use rank tests to compare medians, we can also give confidence intervals for medians. However, the usefulness of rank tests is clearest in settings when they do not simply compare medians—see the discussion “What Hypotheses Does Wilcoxon Test?” (page 15-9). Rank methods focus on significance tests, not confidence intervals.
- Inference based on ranks is largely restricted to simple settings. Normal inference extends to methods for use with complex experimental designs and multiple regression, but nonparametric tests do not. We stress Normal inference in part because it leads to more advanced statistics.

If you read Chapter 16 and use software that makes permutation tests available to you, you will also want to compare rank tests with resampling methods.

- Both rank and permutation tests are nonparametric. That is, they require no assumptions about the shape of the population distribution. A two-sample permutation test has the same null hypothesis as the Wilcoxon rank sum test: that the two population distributions are identical. Calculation of the sampling distribution under the null hypothesis is similar for both tests but is simpler for rank tests because it depends only on the sizes of the samples. As a result, software often gives exact  $P$ -values for rank tests but not for permutation tests.
- Permutation tests have the advantage of flexibility. They allow wide choice of the

statistic used to compare two samples, an advantage over both the  $t$  and Wilcoxon tests. In fact, we could apply the permutation test method to sample means (imitating  $t$ ) or to rank sums (imitating Wilcoxon), as well as to other statistics such as the trimmed mean that we used in Exercise 1.99. Permutation tests are not available in some settings, such as testing hypotheses about a single population, though bootstrap confidence intervals do allow resampling tests in these settings. Permutation tests are available for multiple regression and some other quite elaborate settings.



trimmed mean, p. 53

- An important advantage of resampling methods over both Normal and rank procedures is that we can get bootstrap confidence intervals for the parameter corresponding to whatever statistic we choose for the permutation test. If the samples are very small, however, bootstrap confidence intervals may be unreliable because the samples don't represent the population well enough to provide a good basis for bootstrapping.

In general, both Normal distribution methods and resampling methods are more useful than rank tests. *If you are familiar with resampling, we recommend rank tests only for very small samples, and even then only if your software gives exact  $P$ -values for rank tests but not for permutation tests.*



## SECTION 15.1 Summary

**Nonparametric tests** do not require any specific form for the distribution of the population from which our samples come.

**Rank tests** are nonparametric tests based on the **ranks** of observations, their positions in the list ordered from smallest (rank 1) to largest. Tied observations receive the average of their ranks.

The **Wilcoxon rank sum test** compares two distributions to assess whether one has systematically larger values than the other. The Wilcoxon test is based on the **Wilcoxon rank sum statistic  $W$** , which is the sum of the ranks of one of the samples. The Wilcoxon test can replace the **two-sample  $t$  test**.

**$P$ -values for the Wilcoxon test** are based on the sampling distribution of the rank sum statistic  $W$  when the null hypothesis (no difference in distributions) is true. You can find  $P$ -values from special tables, software, or a Normal approximation (with continuity correction).

## SECTION 15.1 Exercises

For Exercises 15.1 and 15.2, see page 15-4; for Exercises 15.3 and 15.4, see page 15-6; for Exercises 15.5 and 15.6, see page 15-8; and for Exercise 15.7, see page 15-11.

### 15.8 Time spent studying.

Students in a large first-year college class were asked how much time they spent studying on a typical weeknight. Here are the responses, in minutes, for five female students in the class:  STUDYT

120 360 115 60 170

Find the ranks for these data.

### 15.9 Find the rank sum statistic.

Refer to the previous exercise. Here are the data for six men in the class:  STUDYT

0 300 75 90 30 130

Compute the value of the Wilcoxon statistic. Take the first sample to be the women.

### 15.10 State the hypotheses.

Refer to the previous exercise. State appropriate null and alternative hypotheses for this setting.  STUDYT

### 15.11 Find the mean and standard deviation of the distribution of the statistic.

The statistic  $W$  that you calculated in Exercise 15.9 is a random variable with a sampling distribution. What are the mean and the standard deviation of this sampling distribution under the null hypothesis?  STUDYT

### 15.12 Find the $P$ -value.

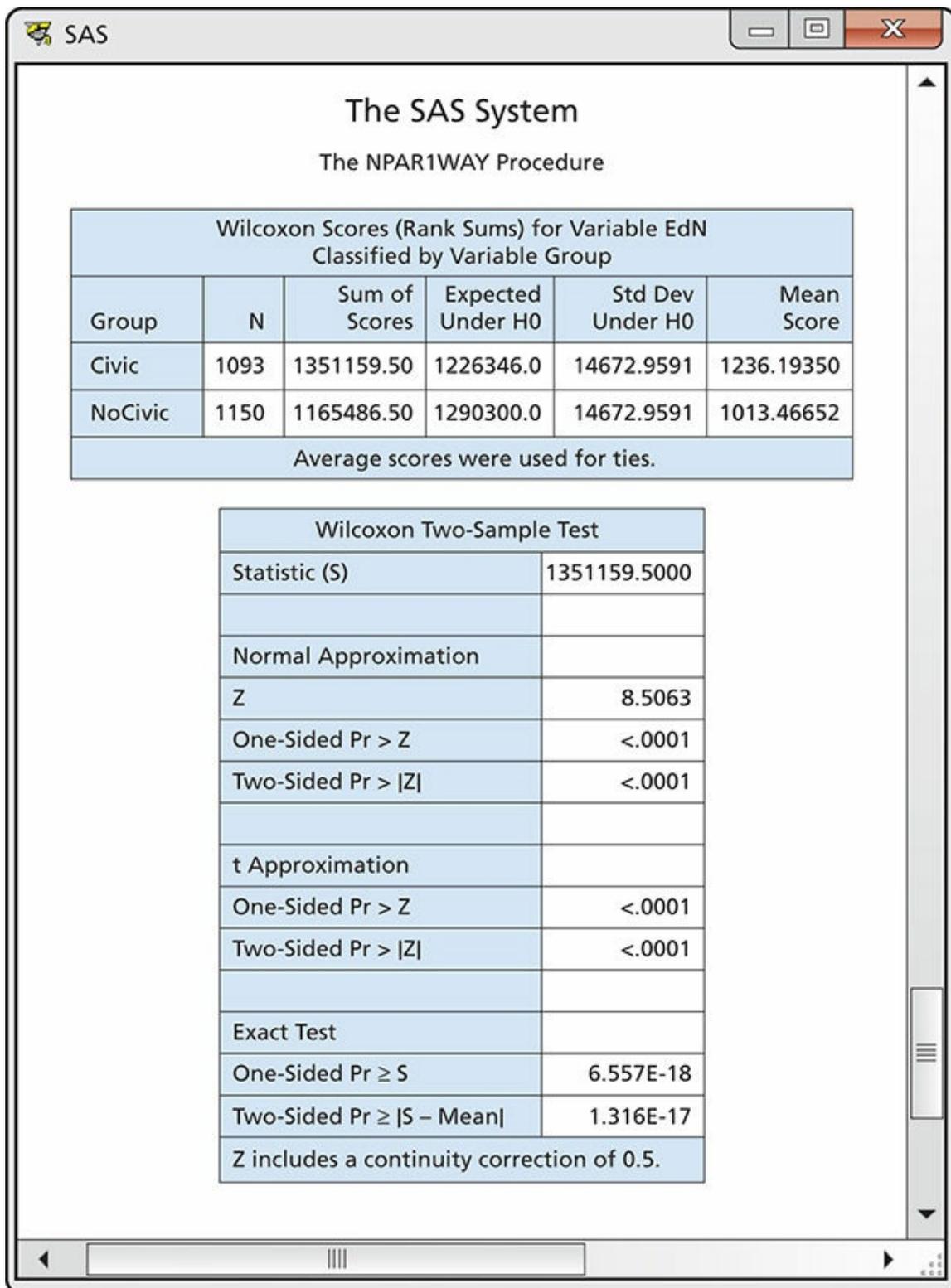
Refer to Exercises 15.8 to 15.11. Find the  $P$ -value using the Normal approximation with the continuity correction and interpret the result of the significance test.  STUDYT

### 15.13 Is civic engagement related to education?

A Pew Internet Poll of adults aged 18 and older examined factors related to civic engagement. Participants were asked whether or not they had participated in a civic group or activity in the preceding 12 months.

One analysis looked at the relationship between this variable and education. Here are the data.<sup>4</sup>  CIVIC

	Education			
Civic participation	No high school	High school	Some college	College
Civic	76	294	295	428



**FIGURE 15.5**

Output from SAS for the civic participation data, for Exercise 15.13.

Figure 15.5 gives the SAS output for analyzing these data using the Wilcoxon rank sum procedure.

(a) Describe the relevant parts of the output and write a short summary of the results.

(b) Apply the “systematically larger” framework that we used in Example 15.7 (page 15-12) to these data. Is this a useful way to describe the results of this analysis? Give reasons for your answer.

### 15.14 Do women talk more?

Conventional wisdom suggests that women are more talkative than men. One study designed to examine this stereotype collected data on the speech of 10 men and 10 women in the United States.<sup>5</sup> The variable recorded is the number of words per day. Here are the data:  **TALK10**

Men				Women			
23,871	5,180	9,951	12,460	10,592	24,608	13,739	22,376
17,155	10,344	9,811	12,387	9,351	7,694	16,812	21,066
29,920	21,791			32,291	12,320		

(a) Summarize the data for the two groups using numerical and graphical methods. Describe the two distributions.

(b) Compare the words per day spoken by the men with the words per day spoken by the women using the Wilcoxon rank sum test. Summarize your results and conclusion in a short paragraph.

### 15.15 More data for women and men talking.

The data in the previous exercise were a sample of the data collected in a larger study of 42 men and 37 women. Use the larger data set to answer the questions in the previous exercise. Discuss the advisability of using the Wilcoxon test versus the  $t$  test for this exercise and for the previous one.  **TALK**

### 15.16 Learning math through subliminal messages.

A “subliminal” message is below our threshold of awareness but may nonetheless influence us. Can subliminal messages help students learn math? A group of students who had failed the mathematics part of the City University of New York Skills Assessment Test agreed to participate in a study to find out. All received a daily subliminal message, flashed on a screen too rapidly to be consciously read. The treatment group of 10 students was exposed to “Each day I am getting better in math.” The control group of 8 students was exposed to a neutral message, “People are walking on the street.” All students participated in a summer program designed to raise their math skills, and all took the assessment test again at the end of the program. Here are data on the subjects’ scores before and after the program.<sup>6</sup>  **SUBLIM**

Treatment Group		Treatment Group	
Pretest	Posttest	Pretest	Posttest
18	24	18	29
18	25	24	29
21	33	20	24
18	29	18	26
18	33	24	38
20	36	22	27
23	34	15	22
23	36	19	31
21	34		

- (a) The study design was a randomized comparative experiment. Outline this design.
- (b) Compare the gain in scores in the two groups, using a graph and numerical descriptions. Does it appear that the treatment group's scores rose more than the scores for the control group?
- (c) Apply the Wilcoxon rank sum test to the posttest versus pretest differences. Note that there are some ties. What do you conclude?

### 15.17 Storytelling and the use of language.

A study of early childhood education asked kindergarten students to retell two fairy tales that had been read to them earlier in the week. The 10 children in the study included 5 high-progress readers and 5 low-progress readers. Each child told two stories. Story 1 had been read to them; Story 2 had been read and also illustrated with pictures. An expert listened to a recording of each child and assigned a score for certain uses of language. Here are the data:<sup>7</sup>  **STORY**

Child	Progress	Story 1 score	Story 2 score	Child	Progress	Story 1 score	Story 2 score
1	high	0.55	0.80	6	low	0.40	0.77
2	high	0.57	0.82	7	low	0.72	0.49
3	high	0.72	0.54	8	low	0.00	0.66
4	high	0.70	0.79	9	low	0.36	0.28
5	high	0.84	0.89	10	low	0.55	0.38

Is there evidence that the scores of high-progress readers are higher than those of low-progress readers when they retell a story they have heard without pictures (Story 1)?

- (a) Make Normal quantile plots for the 5 responses in each group. Are any major deviations from Normality apparent?
- (b) Carry out a two-sample  $t$  test. State hypotheses and give the two sample means, the  $t$  statistic and its  $P$ -value, and your conclusion.
- (c) Carry out the Wilcoxon rank sum test. State hypotheses and give the rank sum  $W$  for high-progress readers, its  $P$ -value, and your conclusion. Do the  $t$  and Wilcoxon tests lead you to different conclusions?

### 15.18 Repeat the analysis for Story 2.

Repeat the analysis of Exercise 15.17 for the scores when children retell a story they have heard and seen illustrated with pictures (Story 2).  **STORY**

### 15.19 Do the calculations by hand.

Use the data in Exercise 15.17 for children telling Story 2 to carry out by hand the steps in the Wilcoxon rank sum test.  **STORY**

- (a) Arrange the 10 observations in order and assign ranks. There are no ties.
- (b) Find the rank sum  $W$  for the 5 high-progress readers. What are the mean and standard deviation of  $W$  under the null hypothesis that low-progress and high-progress readers do not differ?

(c) Standardize  $W$  to obtain a  $z$  statistic. Do a Normal probability calculation with the continuity correction to obtain a one-sided  $P$ -value.

(d) The data for Story 1 contain tied observations. What ranks would you assign to the 10 scores for Story 1?

## 15.2 The Wilcoxon Signed Rank Test

When you complete this section, you will be able to

- For a set of paired sample data, take the differences between the pairs, take the absolute values of the differences, put the absolute values of the differences in order, from smallest to largest with an indication of which absolute differences were from positive differences.
- Compute the Wilcoxon signed rank statistic  $W^+$  from an ordered list of differences with an indication of which absolute differences were from positive differences.
- State the null and alternative hypotheses that are used for the analysis of data using the Wilcoxon signed rank test.
- Using the sample size (that is, the number of pairs), find the mean and the standard deviation of the sampling distribution of the Wilcoxon signed rank statistic under the null hypothesis.
- Find the  $P$ -value for the Wilcoxon signed rank significance test using the Normal approximation with the continuity correction.
- For the Wilcoxon signed rank test, use computer output to determine the results of the significance test.
- Test a hypothesis about the median of a distribution using the Wilcoxon signed rank test.

We use the one-sample  $t$  procedures for inference about the mean of one population or for inference about the mean difference in a matched pairs setting. The matched pairs setting is more important because good studies are generally comparative. We will now meet a rank test for this setting.

### Example

#### 15.8 Storytelling and reading.

A study of early childhood education asked kindergarten students to retell two fairy tales that had been read to them earlier in the week. The first (Story 1)

had been read to them, and the second (Story 2) had been read but also illustrated with pictures. An expert listened to recordings of the children retelling each story and assigned a score for certain uses of language. Here are the data for five “low-progress” readers in a pilot study:<sup>8</sup> Higher scores are better.



Child	1	2	3	4	5
Story 2	0.77	0.49	0.66	0.28	0.38
Story 1	0.40	0.72	0.00	0.36	0.55
Difference	0.37	-0.23	0.66	-0.08	-0.17

We wonder if illustrations improve how the children retell a story. We would like to test the hypotheses

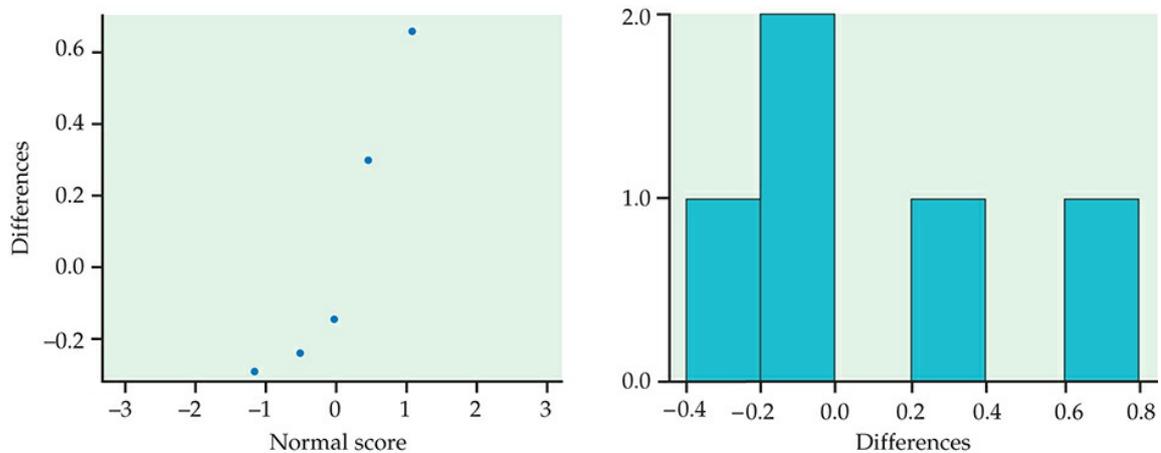


$H_0$ : Scores have the same distribution for both stories.

$H_a$ : Scores are systematically higher for Story 2.

Because this is a matched pairs design, we base our inference on the differences. The matched pairs  $t$  test gives  $t = 0.635$  with one-sided  $P$ -value  $P$

$= 0.280$ . Displays of the data (Figure 15.6) suggest some lack of Normality. We would therefore like to use a rank test.



**FIGURE 15.6**

Normal quantile plot and histogram for the five differences in story scores, for Example 15.8.

Positive differences in Example 15.8 indicate that the child performed better telling Story 2. If scores are generally higher with illustrations, the positive differences should be farther from zero in the positive direction than the negative differences are in the negative direction. We therefore compare the *absolute values* of the differences, that is, their magnitudes without a sign. Here they are, with boldface indicating the positive values:

absolute value

**0.37** 0.23 **0.66** 0.08 0.17

Arrange these in increasing order and assign ranks, keeping track of which values were originally positive. Tied values receive the average of their ranks. If there are cases with zero differences, discard them before ranking.

Absolute value	0.08	0.17	0.23	<b>0.37</b>	<b>0.66</b>
Rank	1	2	3	<b>4</b>	<b>5</b>

The test statistic is the sum of the ranks of the positive differences. (We could equally well use the sum of the ranks of the negative differences.) This is the *Wilcoxon signed rank statistic*. Its value here is  $W^+ = 9$ .

## THE WILCOXON SIGNED RANK TEST FOR MATCHED PAIRS

Draw an SRS of size  $n$  from a population for a matched pairs study and take the differences in responses within pairs. Rank the absolute values of these

differences. The sum  $W^+$  of the ranks for the positive differences is the **Wilcoxon signed rank statistic**. If the distribution of the responses is not affected by the different treatments within pairs, then  $W^+$  has mean

$$\mu_{W^+} = n(n+1)/4$$

and standard deviation

$$\sigma_{W^+} = \sqrt{n(n+1)(2n+1)/24}$$

The **Wilcoxon signed rank test** rejects the hypothesis that there are no systematic differences within pairs when the rank sum  $W^+$  is far from its mean.

## USE YOUR KNOWLEDGE

### 15.20 Service and food provided by top 25 spas.



SPAS3

The readers' poll in *Condé Nast Traveler* magazine that ranked 100 top resort spas and that was described in Exercise 15.1 also reported scores on service and on food. Here are the scores for a random sample of 7 spas that ranked in the top 25:

Spa	1	2	3	4	5	6	7
Service	89.6	89.8	87.3	94.2	95.8	87.9	91.0
Food	83.1	88.1	85.8	92.9	95.7	80.7	83.6

Is service more important than food for a top ranking? Formulate this question in terms of null and alternative hypotheses. Then compute the differences and find the value of the Wilcoxon signed rank statistic,  $W^+$ .

### 15.21 Scores for the next 25 spas.



SPAS4

Refer to the previous exercise. Here are the scores for a random sample of 7 spas that ranked between 26 and 50:

Spa	1	2	3	4	5	6	7
Service	90.6	87.2	95.0	88.4	91.5	88.2	91.2
Food	86.6	74.4	89.1	81.0	85.7	83.2	93.1

Answer the questions from the previous exercise for this setting.

## Example

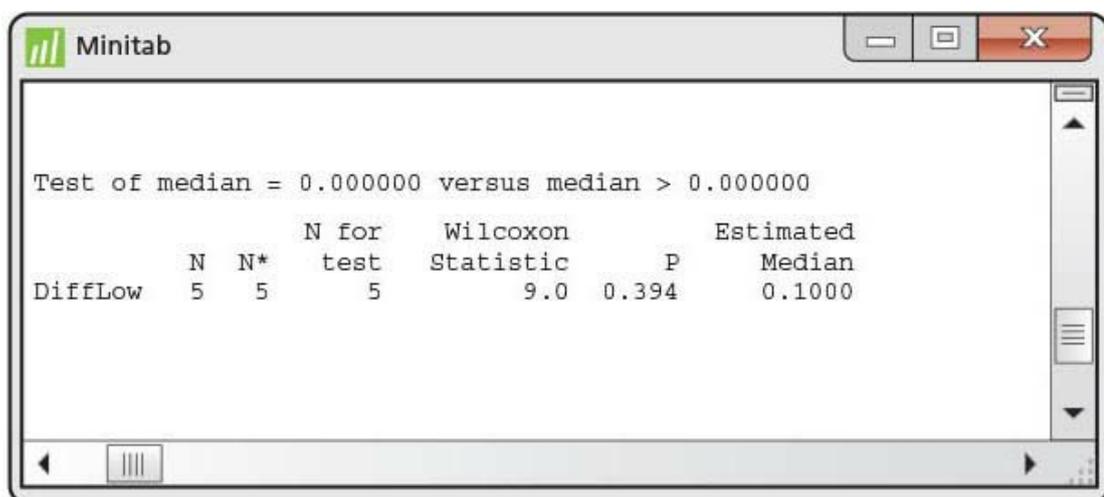
### 15.9 Software output.



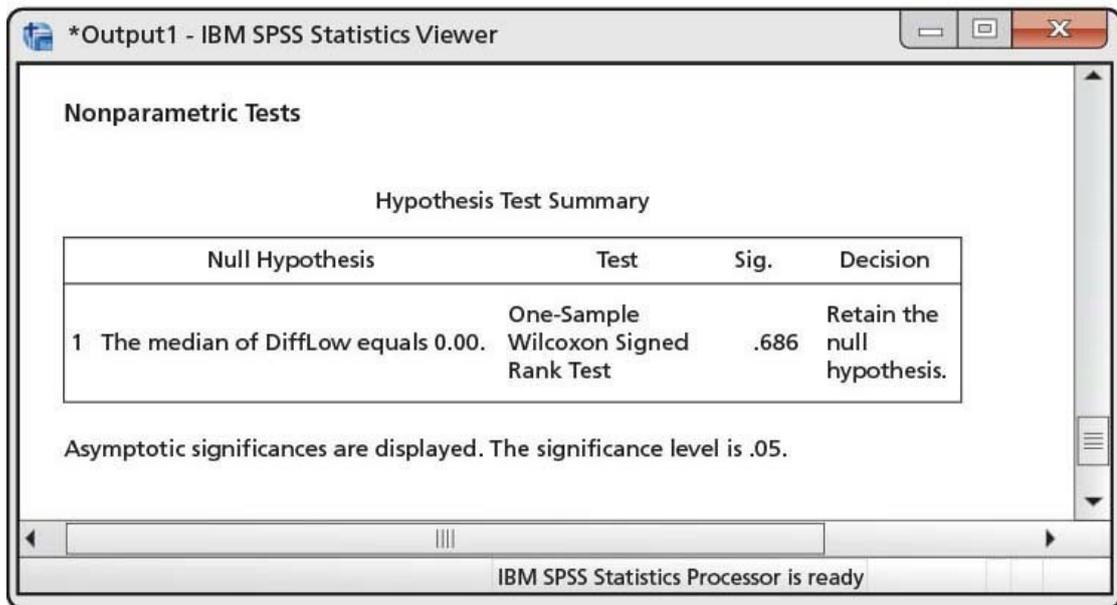
In the storytelling study of Example 15.8,  $n = 5$ . If the null hypothesis (no systematic effect of illustrations) is true, the mean of the signed rank statistic is

$$\mu_{W^+} = n(n+1)/4 = (5)(6)/4 = 7.5$$

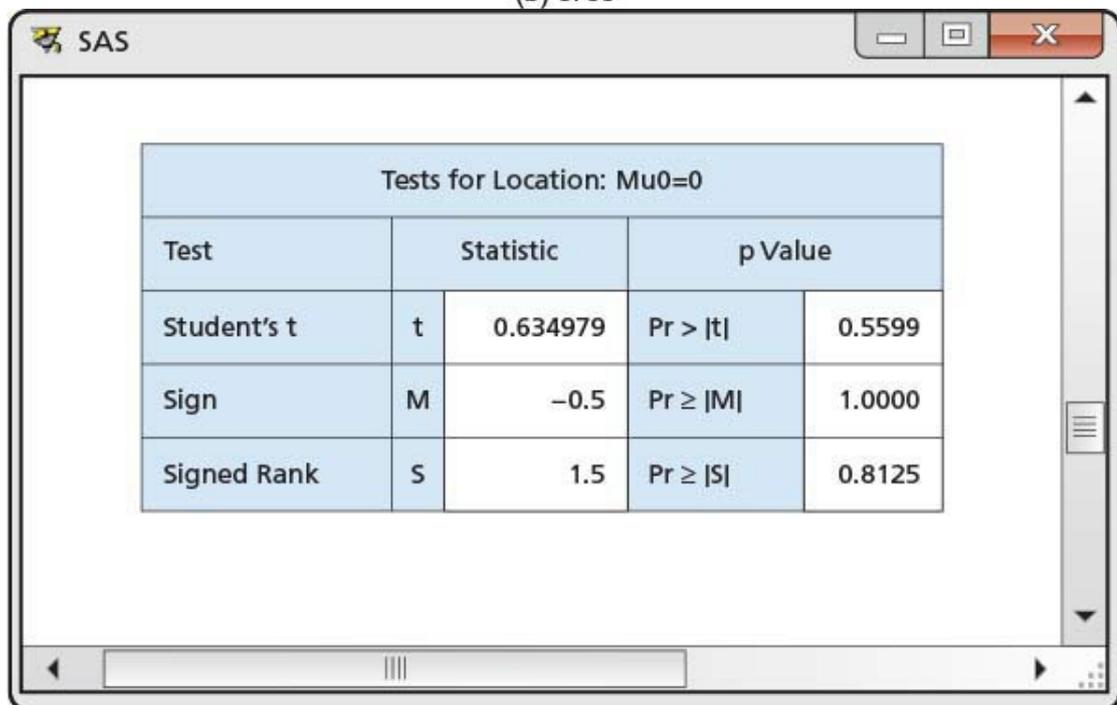
Our observed value  $W^+ = 9$  is only slightly larger than this mean. The one-sided  $P$ -value is  $P(W^+ \geq 9)$ .



(a) Minitab



(b) SPSS



(c) SAS

**FIGURE 15.7**

Output from (a) Minitab, (b) SPSS, and (c) SAS for the storytelling data, for Example 15.9.

Most statistical software uses the differences between the two variables, *with the signs*, as input. Alternatively, the differences can sometimes be calculated within the software. Figure 15.7 displays the output from three statistical programs. Each does things a little differently. The Minitab output in Figure 15.7(a) gives  $P = 0.394$  for the one-sided Wilcoxon signed rank test with  $n = 5$  observations and  $W^+ = 9$ . In Figure 15.7(b), the SPSS output gives  $P = 0.686$  for testing the two-sided alternative. The results from SAS in Figure

15.7(c) are part of the usual output for the analysis of a single variable. The two-sided alternative is used. The test statistic for the signed rank test is given as  $S = 1.5$ . This quantity is  $W^+$  minus its expected value  $\mu_{W^+} = 7.5$ ,  $S = W^+ - \mu_{W^+}$ . The  $P$ -value is given as  $P = 0.8125$ .

Results reported in the three outputs lead us to the same qualitative conclusion: the data do not provide evidence to support the idea that the Story 2 scores are higher than (or not equal to) the Story 1 scores. Different methods and approximations are used to compute the  $P$ -values. With larger sample sizes, we would not expect so much variation in the  $P$ -values. Note that the  $t$  test results reported in SAS also give the same conclusion,  $P = 0.5599$ .

When the sampling distribution of a test statistic is symmetric, we can use output that gives a  $P$ -value for a two-sided alternative to compute a  $P$ -value for a one-sided alternative. Check that the effect is in the direction specified by the one-sided alternative and then divide the  $P$ -value by 2.

## The Normal approximation

The distribution of the signed rank statistic when the null hypothesis (no difference) is true becomes approximately Normal as the sample size becomes large. We can then use Normal probability calculations (with the continuity correction) to obtain approximate  $P$ -values for  $W^+$ . Let's see how this works in the storytelling example, even though  $n = 5$  is certainly not a large sample.

### Example

#### 15.10 The Normal approximation.

For  $n = 5$  observations, we saw in Example 15.9 that  $\mu_{W^+} = 7.5$ . The standard deviation of  $W^+$  under the null hypothesis is

$$\begin{aligned}\sigma_{W^+} &= \sqrt{n(n+1)(2n+1)/24} \\ &= \sqrt{(5)(6)(11)/24} \\ &= 3.75 \approx 3.708\end{aligned}$$

The continuity correction calculates the  $P$ -value  $P(W^+ \geq 9)$  as  $P(W^+ \geq 8.5)$ , treating the value  $W^+ = 9$  as occupying the interval from 8.5 to 9.5. We find

the Normal approximation for the  $P$ -value by standardizing and using the standard Normal table:

$$\begin{aligned}P(W^+ \geq 8.5) &= P(W^+ - 7.53.708 \geq 8.5 - 7.53.708) \\ &= P(Z \geq 0.27) \\ &= 0.394\end{aligned}$$

Despite the small sample size, the Normal approximation gives a result quite close to the exact value  $P = 0.4062$ . Figure 15.7(b) shows that the approximation is much less accurate without the continuity correction. *This output reminds us not to trust software unless we know exactly what it does.*



## USE YOUR KNOWLEDGE

### 15.22 Significance test for top-ranked spas.

Refer to Exercise 15.20 (page 15-20). Find  $\mu_{W^+}$ ,  $\sigma_{W^+}$  and the Normal approximation for the  $P$ -value for the Wilcoxon signed rank test.



SPAS3

### 15.23 Significance test for lower-ranked spas.

Refer to Exercise 15.21 (page 15-20). Find  $\mu_{W^+}$ ,  $\sigma_{W^+}$ , and the Normal approximation for the  $P$ -value for the Wilcoxon signed rank test.



SPAS4

**Ties**

Ties among the absolute differences are handled by assigning average ranks. A tie *within* a pair creates a difference of zero. Because these are neither positive nor negative, the usual procedure simply drops such pairs from the sample. *This amounts to dropping observations that favor the null hypothesis (no difference). If there are many ties, the test may be biased in favor of the alternative hypothesis.* As in the case of the Wilcoxon rank sum, ties complicate finding a  $P$ -value. Most software no longer provides an exact distribution for the signed rank statistic  $W^+$ , and the standard deviation  $\sigma_{W^+}$  must be adjusted for the ties before we can use the Normal approximation. Software will do this. Here is an example.



## Example

### 15.11 Golf scores of a women's golf team.

Here are the golf scores of 12 members of a college women's golf team in two rounds of tournament play. (A golf score is the number of strokes required to complete the course, so that low scores are better.)



Player	1	2	3	4	5	6	7	8	9	10	11	12
Round 2	94	85	89	89	81	76	107	89	87	91	88	80
Round 1	89	90	87	95	86	81	102	105	83	88	91	79
Difference	5	-5	2	-6	-5	-5	5	-16	4	3	-3	1

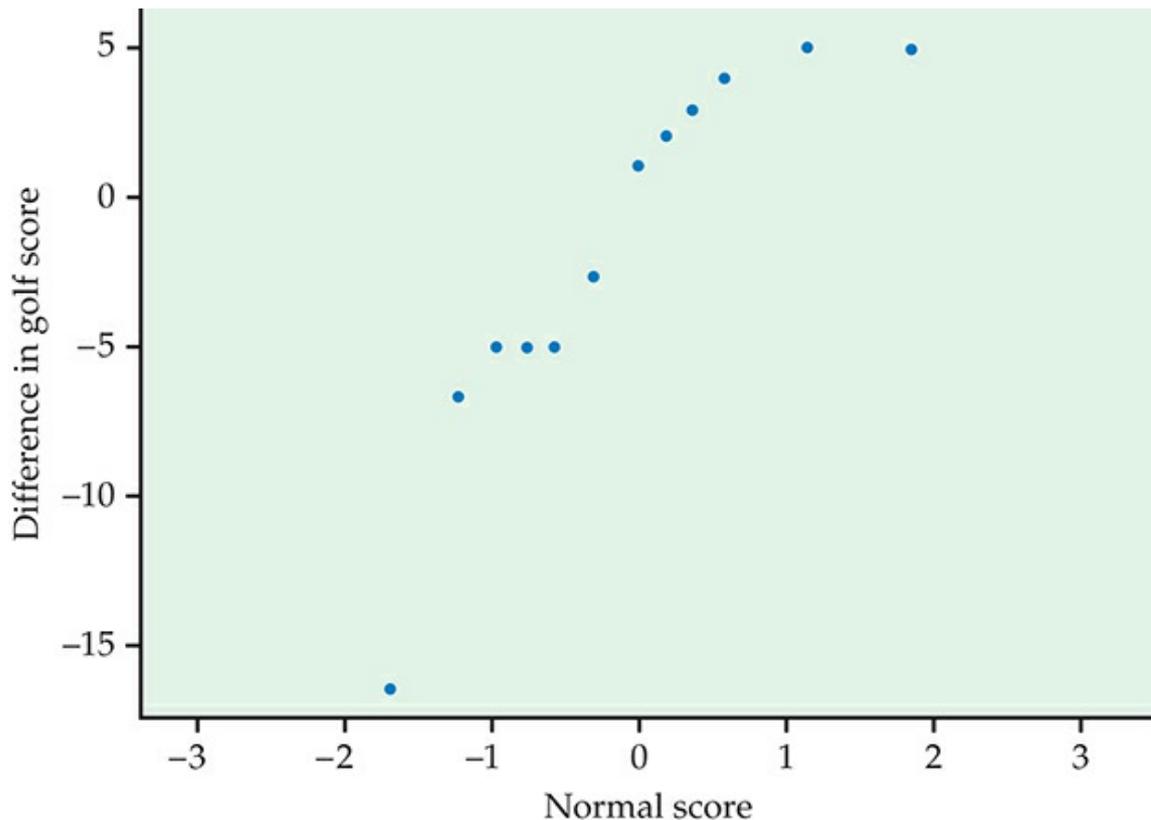
Negative differences indicate better (lower) scores on the second round. We

see that 6 of the 12 golfers improved their scores. We would like to test the hypotheses that in a large population of collegiate women golfers

$H_0$ : Scores have the same distribution in Rounds 1 and 2.

$H_a$ : Scores are systematically lower or higher in Round 2.

A Normal quantile plot of the differences (Figure 15.8) shows some irregularity and a low outlier. We will use the Wilcoxon signed rank test.



**FIGURE 15.8**

Normal quantile plot of the difference in scores for two rounds of a golf tournament, for Example 15.11.

The absolute values of the differences, with boldface indicating those that are negative, are

5 5 **2** **6** **5** **5** **5** **16** 4 3 **3** 1

Arrange these in increasing order and assign ranks, keeping track of which values were originally negative. Tied values receive the average of their ranks.

Absolute value	1	2	<b>3</b>	<b>3</b>	4	<b>5</b>	<b>5</b>	<b>5</b>	<b>5</b>	<b>5</b>	<b>6</b>	<b>16</b>
Rank	1	2	<b>3.5</b>	<b>3.5</b>	5	<b>8</b>	<b>8</b>	<b>8</b>	<b>8</b>	<b>8</b>	<b>11</b>	<b>12</b>

The Wilcoxon signed rank statistic is the sum of the ranks of the negative differences. (We could equally well use the sum of the ranks of the positive

differences.) Its value is  $W^+ = 50.5$ .

## Example

### 15.12 Software output.

Here are the two-sided  $P$ -values for the Wilcoxon signed rank test for the golf score data from three statistical programs:

Program	$P$ -value
Minitab	$P = 0.388$
SAS	$P = 0.388$
SPSS	$P = 0.363$

All lead to the same practical conclusion: these data give no evidence for a systematic change in scores between rounds. However, the  $P$ -value reported by SPSS differs a bit from the other two. The reason for the variation is that the programs use slightly different versions of the approximate calculations needed when ties are present. The exact result depends on which version the software programmer chose to use.

For the golf data, the matched pairs  $t$  test gives  $t = 0.9314$  with  $P = 0.3716$ . Once again,  $t$  and  $W^+$  lead to the same conclusion.

## Testing a hypothesis about the median of a distribution

Let's take another look at how the Wilcoxon signed rank test works. We have data for a pair of variables measured on the same individuals. The analysis starts with the differences between the two variables. These differences are what we input to statistical software.

At this stage we can think of our data as consisting of a single variable. The Wilcoxon signed rank test tests the null hypothesis that the population median of the differences is zero. The alternative is that the median is not zero.

Think about starting the analysis at the stage where we have a single variable and we are interested in testing a hypothesis about the median. The null hypothesis does not necessarily need to be zero. If it is some other value, we simply subtract that value from each observation before we start the analysis. Exercise 15.35 (page 15-27) leads you through the steps needed for this analysis.

## SECTION 15.2 Summary

The **Wilcoxon signed rank test** applies to matched pairs studies. It tests the null hypothesis that there is no systematic difference within pairs against alternatives that assert a systematic difference (either one-sided or two-sided).

The test is based on the **Wilcoxon signed rank statistic**  $W^+$ , which is the sum of the ranks of the positive (or negative) differences when we rank the absolute values of the differences. The **matched pairs  $t$  test** and the **sign test** are alternative tests in this setting.

**$P$ -values for the signed rank test** are based on the sampling distribution of  $W^+$  when the null hypothesis is true. You can find  $P$ -values from special tables, software, or a Normal approximation (with continuity correction).

## SECTION 15.2 Exercises

For Exercises 15.20 and 15.21, see page 15-20; and for Exercises 15.22 and 15.23, see page 15-23.

### 15.24 Fuel efficiency.

Computers in some vehicles calculate various quantities related to performance. One of these is the fuel efficiency, or gas mileage, usually expressed as miles per gallon (mpg). For one vehicle equipped in this way, the mpg were recorded each time the gas tank was filled, and the computer was then reset. In addition to the computer calculating mpg, the driver also recorded the mpg by dividing the miles driven by the number of gallons at fill-up.<sup>9</sup>

The driver wants to determine if these calculations are different.  **MPG8**

Fill-up	1	2	3	4	5	6	7	8
Computer	41.5	50.7	36.6	37.3	34.2	45.0	48.0	43.2
Driver	36.5	44.2	37.2	35.6	30.5	40.5	40.0	41.0

- For each of the eight fill-ups find the difference between the computer mpg and the driver mpg.
- Find the absolute values of the differences you found in part (a).
- Order the absolute values of the differences that you found in part (b) from smallest to largest, and underline those absolute differences that came from positive differences in part (a).

### 15.25 Find the Wilcoxon signed rank statistic.

Using the work that you performed in the previous exercise, find the value of the Wilcoxon signed rank statistic  $W^+$

### 15.26 State the hypotheses.

Refer to Exercise 15.24. State the null hypothesis and the alternative hypothesis for this setting.

### 15.27 Find the mean and the standard deviation.

Refer to Exercise 15.24. Use the sample size to find the mean and the standard deviation of the sampling distribution of the Wilcoxon signed rank statistic  $W^+$  under the null hypothesis.

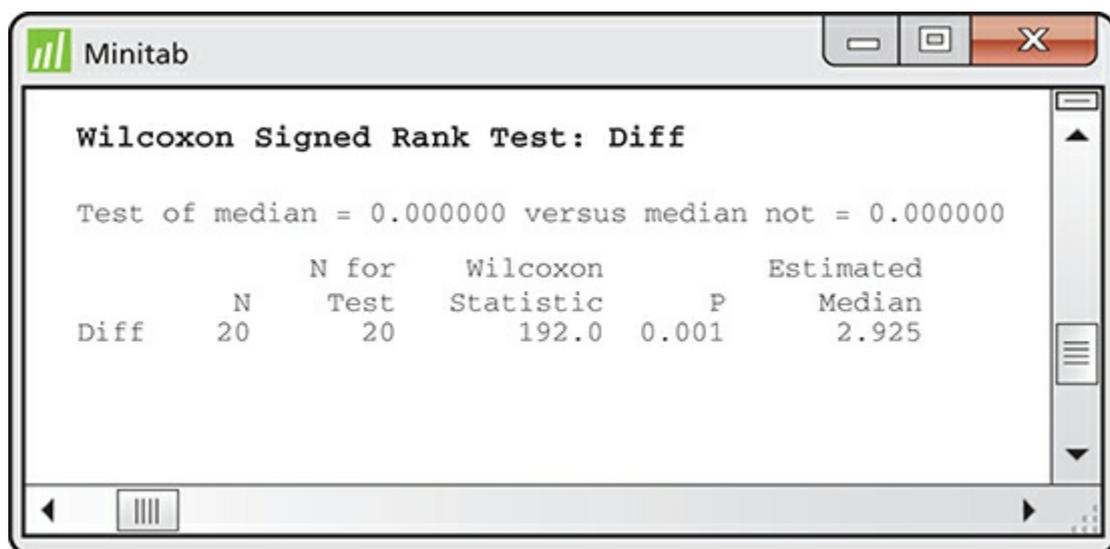
### 15.28 Find the $P$ -value.

Refer to Exercises 15.24 to 15.27. Find the  $P$ -value for the Wilcoxon signed rank statistic using the Normal approximation with the continuity correction.

### 15.29 Read the output.

The data in Exercise 15.24 are a subset of a larger set of data. Figure 15.9 gives Minitab output for the analysis of this larger set of data.  **MPGCOMP**

- (a) How many pairs of observations are in the larger data set?
- (b) What is the value of the Wilcoxon signed rank statistic  $W^+$ ?
- (c) Report the  $P$ -value for the significance test and give a brief statement of your conclusion.



**FIGURE 15.9**

Minitab output for the fuel efficiency data, for Exercise 15.29.

- (d) The output reports an estimated median. Explain how this statistic is calculated from the data.

### 15.30 Number of friends on Facebook.

Facebook recently examined all active Facebook users (more than 10% of the global population) and determined that the average user has 190 friends. This distribution takes only integer values, so it is certainly not Normal. It is also highly skewed to the right, with a median of 100 friends.<sup>10</sup> Consider the following SRS of  $n = 30$  Facebook users from your large university.  **FACEFR**

594	60	417	120	132	176	516	319	734	8
31	325	52	63	537	27	368	11	12	190
85	165	288	65	57	81	257	24	297	148

(a) Use the Wilcoxon signed rank procedure to test the null hypothesis that the median number of Facebook friends for Facebook users at your university is 190. Describe the steps in the procedure and summarize the results.

(b) Exercise 7.26 (page 442) asked you to analyze these data using the  $t$  procedure. Perform this analysis and compare the results with those that you found in part (a).

### 15.31 The full moon and behavior.

Can the full moon influence behavior? A study observed 15 nursing-home patients with dementia. The number of incidents of aggressive behavior was recorded each day for 12 weeks. Call a day a “moon day” if it is the day of a full moon or the day before or after a full moon. Here are the average numbers of aggressive incidents for moon days and other days for each subject:<sup>11</sup>  MOON

Patient	Moon days	Other days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26
6	3.67	0.11
7	4.67	0.30
8	2.67	0.40
9	6.00	1.59
10	4.33	0.60
11	3.33	0.65
12	0.67	0.69
13	1.33	1.26
14	0.33	0.23
15	2.00	0.38

The matched pairs  $t$  test (Example 7.7, page 429) gives  $P < 0.000015$ , and a permutation test (Example 16.14, page 16-50) gives  $P = 0.0001$ . Does the Wilcoxon signed rank test, based on ranks rather than means, agree that there is strong evidence that there are more aggressive incidents on moon days?

### 15.32 Comparison of two energy drinks.

Consider the following study to compare two popular energy drinks. For each subject, a coin was flipped to determine which drink to rate first. Each drink was rated on a 0 to 100 scale, with 100 being the highest rating.  ENERDR6

Drink	Subject					
	1	2	3	4	5	6
A	43	83	66	87	78	67
B	45	78	64	79	71	62

(a) Inspect the data. Is there a tendency for these subjects to prefer one of the two energy drinks?

(b) Use the matched pairs  $t$  test of Chapter 7 (page 429) to compare the two drinks.

- (c) Use the Wilcoxon signed rank test to compare the two drinks.
- (d) Write a summary of your results and explain why the two tests give different conclusions.

### 15.33 Comparison of two energy drinks with an additional subject.

Refer to the previous exercise. Let's suppose that there is an additional subject who expresses a strong preference for energy drink "A." Here is the new data set:  ENERDR7

Drink	Subject						
	1	2	3	4	5	6	7
A	43	83	66	87	78	67	90
B	45	78	64	79	71	62	60

Answer the questions given in the previous exercise. Write a summary comparing this exercise with the previous one. Include a discussion of what you have learned regarding the choice of the  $t$  test versus the Wilcoxon signed rank test for different sets of data.

### 15.34 A summer language institute for teachers.

A matched pairs study of the effect of a summer language institute on the ability of teachers to comprehend spoken French had these improvements in scores between the pretest and the posttest for 20 teachers:  SUMLANG

2 0 6 6 3 3 2 3 -6 6  
6 6 3 0 1 1 0 2 3 3

(Exercise 7.45, page 446, applies the  $t$  test to these data; Exercise 16.59, page 16-49, applies a permutation test based on the means.) Show the assignment of ranks and the calculation of the signed rank statistic  $W^+$  for these data. Remember that zeros are dropped from the data before ranking, so that  $n$  is the number of nonzero differences within pairs.

### 15.35 Radon detectors.

How accurate are radon detectors of a type sold to homeowners? To answer this question, university researchers placed 12 detectors in a chamber that exposed them to 105 picocuries per liter (pCi/l) of radon.<sup>12</sup> The detector readings are as follows:  RADON

91.9 97.8 111.4 122.3 105.4 95.0  
103.8 99.6 96.6 119.3 104.8 101.7

We wonder if the median reading differs significantly from the true value 105.

- (a) Graph the data, and comment on skewness and outliers. A rank test is appropriate.
- (b) We would like to test hypotheses about the median reading from home radon detectors:

$$H_0: \text{median} = 105$$

$$H_a: \text{median} \neq 105$$

To do this, apply the Wilcoxon signed rank statistic to the differences between the observations and 105. (This is the one-sample version of the test.) What do you conclude?

### 15.36 Vitamin C in wheat-soy blend.

The U.S. Agency for International Development provides large quantities of wheat-soy blend (WSB) for development programs and emergency relief in countries throughout the world. One study collected data on the vitamin C content of 5 bags of WSB at the factory and five months later in Haiti.<sup>13</sup> Here are the data:  WSBVITC

Sample	1	2	3	4	5
Before	73	79	86	88	78
After	20	27	29	36	17

We want to know if vitamin C has been lost during transportation and storage. Describe what the data show about this question. Then use a rank test to see whether there has been a significant loss.

## 15.3 The Kruskal-Wallis Test\*

When you complete this section, you will be able to

- Describe the setting where the Kruskal-Wallis test can be used.
- Specify the null and alternative hypotheses for the Kruskal-Wallis test.
- For the Kruskal-Wallis test, use computer output to determine the results of the significance test.

---

\* Because this test is an alternative to the one-way analysis of variance  $F$  test, you should first read Chapter 12.

We have now considered alternatives to the matched pairs and two-sample  $t$  tests for comparing the magnitude of responses to two treatments. To compare more than two treatments, we use one-way analysis of variance (ANOVA) if the distributions of the responses to each treatment are at least roughly Normal and have similar spreads. What can we do when these distribution requirements are violated?

### Example

#### 15.13 Weeds and corn yield.

Lamb's-quarter is a common weed that interferes with the growth of corn. A researcher planted corn at the same rate in 16 small plots of ground and then randomly assigned the plots to four groups. He weeded the plots by hand to allow a fixed number of lamb's-quarter plants to grow in each meter of corn row. These numbers were 0, 1, 3, and 9 in the four groups of plots. No other weeds were allowed to grow, and all plots received identical treatment except for the weeds. Here are the yields of corn (bushels per acre) in each of the plots:<sup>14</sup>



## WEEDS

Weeds per meter	Corn yield						
0	166.7	1	166.2	3	158.6	9	162.8
0	172.2	1	157.3	3	176.4	9	142.4
0	165.0	1	166.7	3	153.1	9	162.7
0	176.9	1	161.1	3	156.0	9	162.4

The summary statistics are

Weeds	<i>n</i>	Mean	Std. dev.
0	4	170.200	5.422
1	4	162.825	4.469
3	4	161.025	10.493
9	4	157.575	10.118

The sample standard deviations do not satisfy our rule of thumb that for safe use of ANOVA the largest should not exceed twice the smallest. A careful look at the data suggests that there may be some outliers in the 3 and 9 weeds per meter groups. These are the correct yields for their plots, so we have no justification for removing them. Let's use a rank test that is not sensitive to outliers.

## Hypotheses and assumptions

The ANOVA  $F$  test concerns the means of the several populations represented by our samples. For Example 15.13, the ANOVA hypotheses are

$$H_0: \mu_0 = \mu_1 = \mu_3 = \mu_9$$

$$H_a: \text{not all four means are equal}$$

Here,  $\mu_0$  is the mean yield in the population of all corn planted under the conditions of the experiment with no weeds present. The data should consist of four independent random samples from the four populations, all Normally distributed with the same standard deviation.

The **Kruskal-Wallis test** is a rank test that can replace the ANOVA  $F$  test. The assumption about data production (independent random samples from each population) remains important, but we can relax the Normality assumption. We assume only that the response has a continuous distribution in each population. The hypotheses tested in our example are

$$H_0: \text{Yields have the same distribution in all groups.}$$

$$H_a: \text{Yields are systematically higher in some groups than in others.}$$

If all the population distributions have the same shape (Normal or not), these hypotheses take a simpler form. The null hypothesis is that all four populations have the same *median* yield. The alternative hypothesis is that not all four median yields are equal.

## The Kruskal-Wallis test

Recall the analysis of variance idea: we write the total observed variation in the responses as the sum of two parts, one measuring variation among the groups (sum of squares for groups, SSG) and one measuring variation among individual observations within the same group (sum of squares for error, SSE). The ANOVA  $F$  test rejects the null hypothesis that the mean responses are equal in all groups if SSG is large relative to SSE.

The idea of the Kruskal-Wallis rank test is to rank all the responses from all groups together and then apply one-way ANOVA to the ranks rather than to the original observations. If there are  $N$  observations in all, the ranks are always the whole numbers from 1 to  $N$ . The total sum of squares for the ranks is therefore a fixed number no matter what the data are. So we do not need to look at both SSG and SSE. Although it isn't obvious without some unpleasant algebra, the Kruskal-Wallis test statistic is essentially just SSG for the ranks. We give the formula, but you should rely on software to do the arithmetic. When SSG is large, that is evidence that the groups differ.

### THE KRUSKAL-WALLIS TEST

Draw independent SRSs of sizes  $n_1, n_2, \dots, n_I$  from  $I$  populations. There are  $N$  observations in all. Rank all  $N$  observations and let  $R_i$  be the sum of the ranks for the  $i$ th sample. The **Kruskal-Wallis statistic** is

$$H = \frac{12}{N(N+1)} \sum R_i^2 n_i - 3(N+1)$$

When the sample sizes  $n_i$  are large and all  $I$  populations have the same continuous distribution,  $H$  has approximately the chi-square distribution with  $I - 1$  degrees of freedom.

The **Kruskal-Wallis test** rejects the null hypothesis that all populations have the same distribution when  $H$  is large.

We now see that, like the Wilcoxon rank sum statistic, the Kruskal-Wallis statistic is based on the sums of the ranks for the groups we are comparing. The more different these sums are, the stronger is the evidence that responses are systematically larger in some groups than in others.

The exact distribution of the Kruskal-Wallis statistic  $H$  under the null

hypothesis depends on all the sample sizes  $n_1$  to  $n_J$ , so tables are awkward. The calculation of the exact distribution is so time-consuming for all but the smallest problems that even most statistical software uses the chi-square approximation to obtain  $P$ -values. As usual, there is no usable exact distribution when there are ties among the responses. We again assign average ranks to tied observations.

## Example

### 15.14 Perform the significance test.



In Example 15.13, there are  $I = 4$  populations and  $N = 16$  observations. The sample sizes are equal,  $n_i = 4$ . The 16 observations arranged in increasing order, with their ranks, are

Yield	142.4	153.1	156.0	157.3	158.6	161.1	162.4	162.7
Rank	1	2	3	4	5	6	7	8
Yield	162.8	165.0	166.2	166.7	166.7	172.2	176.4	176.9
Rank	9	10	11	12.5	12.5	14	15	16

There is one pair of tied observations. The ranks for each of the four treatments are

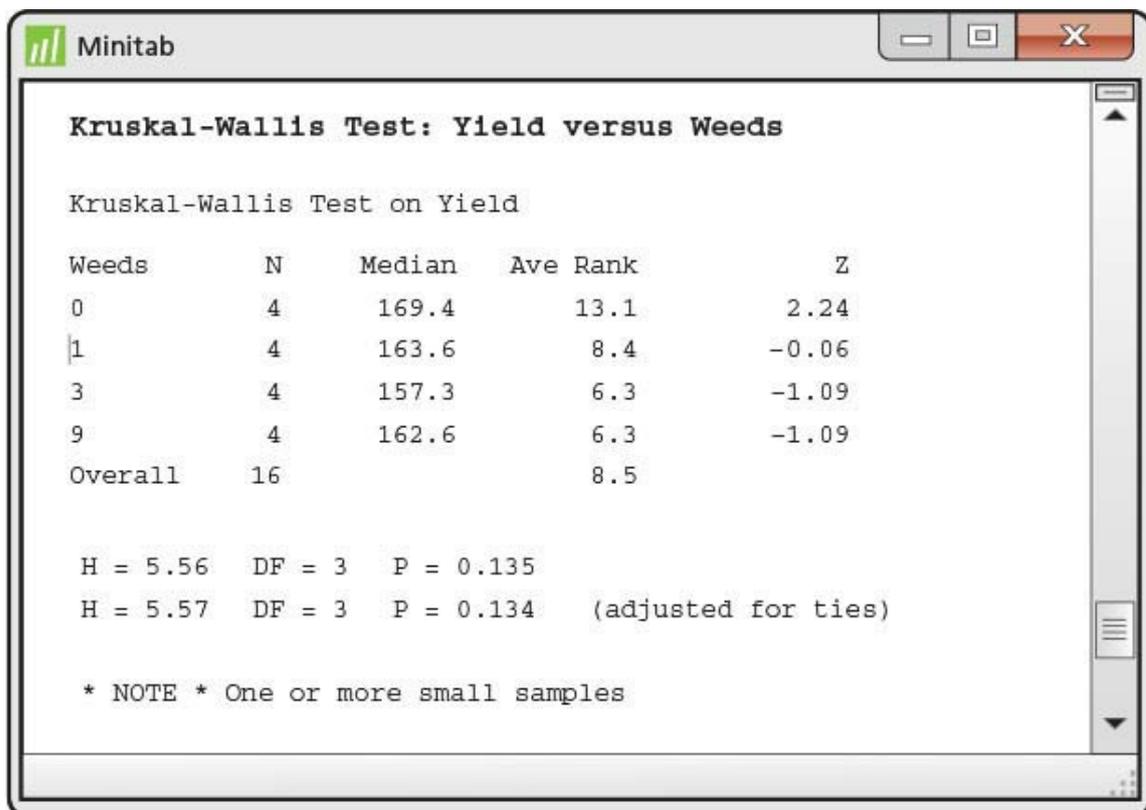
Weeds	Ranks				Rank sums
0	10	12.5	14	16	52.5
1	4	6	11	12.5	33.5
3	2	3	5	15	25.0
9	1	7	8	9	25.0

The Kruskal-Wallis statistic is therefore

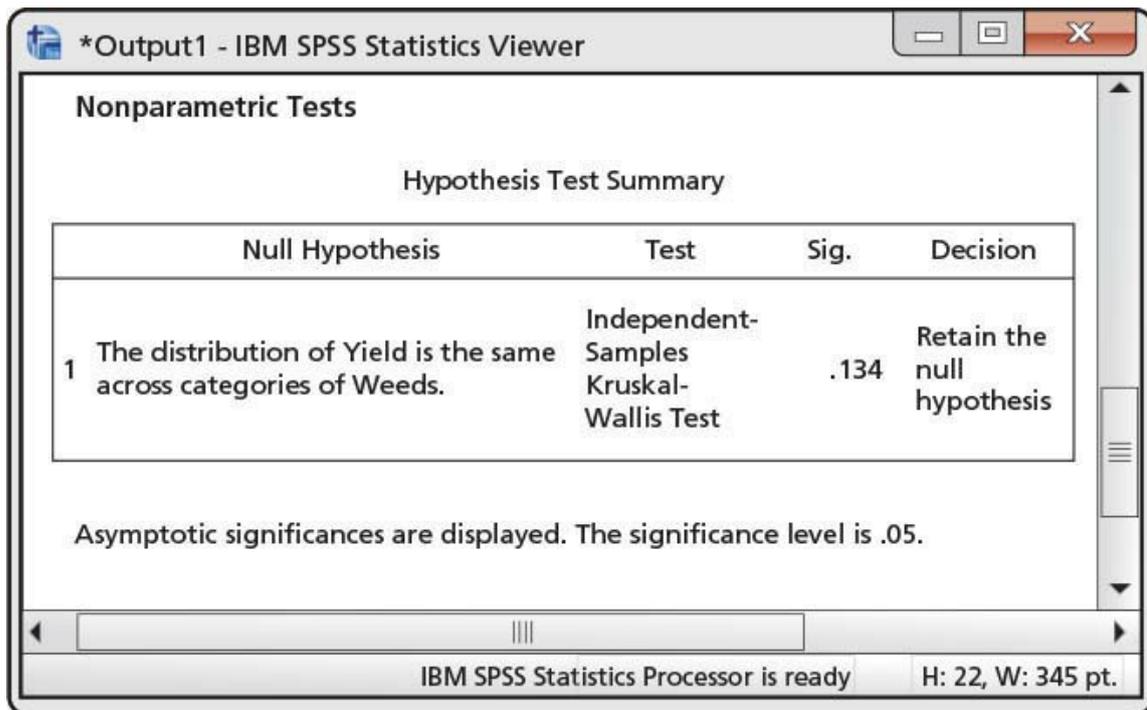
$$\begin{aligned}
 H &= 12N(N+1) \sum R_i^2 n_i - 3(N+1) \\
 &= 12(16)(17)(52.524 + 33.524 + 25.24 + 25.24) - (3)(17) \\
 &= 12272(1282.125) - 51 \\
 &= 5.56
 \end{aligned}$$

Referring to the table of chi-square critical points (Table F) with  $df = 3$ , we find that the  $P$ -value lies in the interval  $0.10 < P < 0.15$ . This small experiment suggests that more weeds decrease yield but does not provide convincing evidence that weeds have an effect.

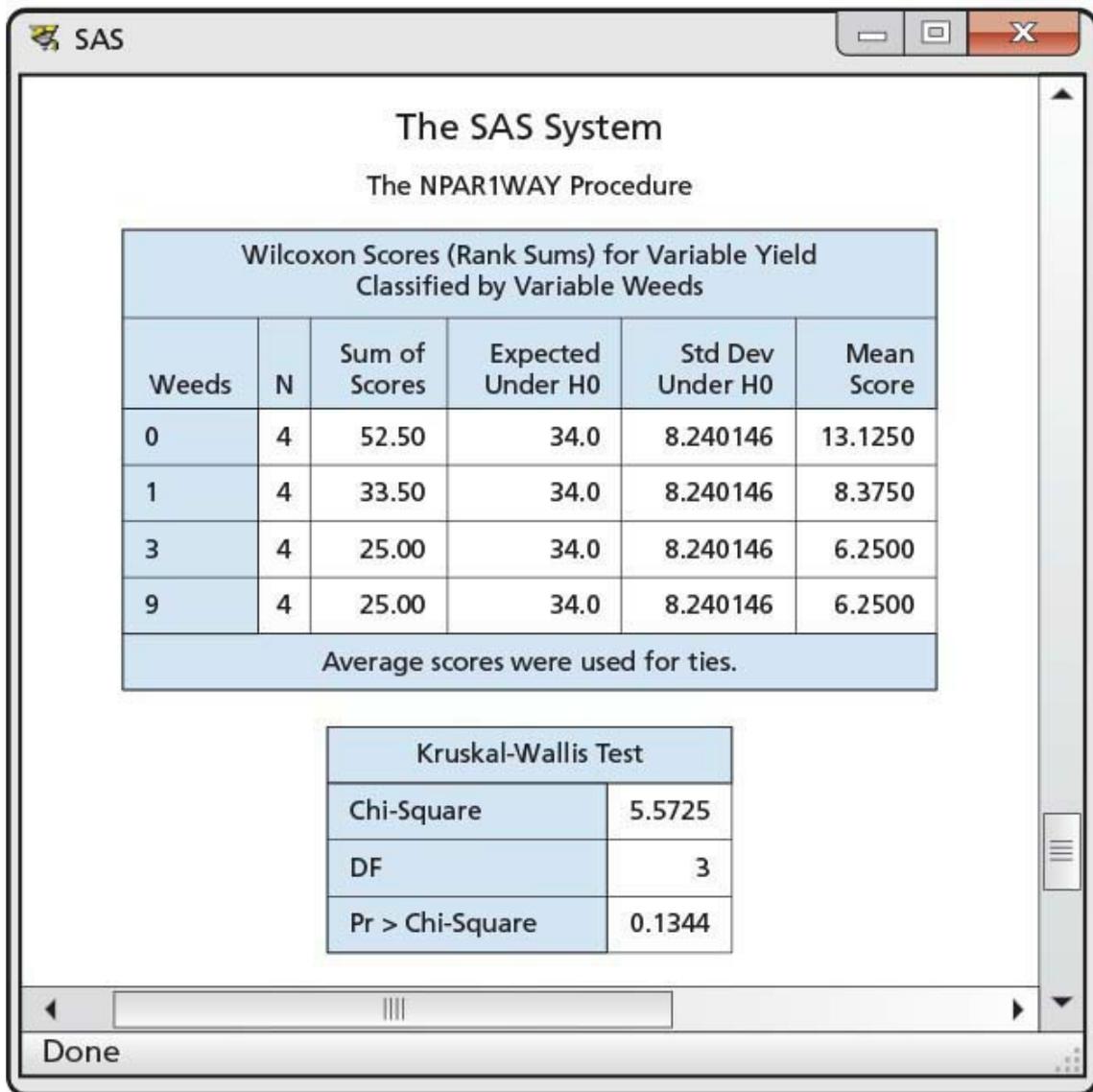
Figure 15.10 displays the output from Minitab, SPSS, and SAS for the analysis of the data in Example 15.14. Minitab gives the  $H$  statistic adjusted for ties as  $H = 5.57$  with 3 degrees of freedom and  $P = 0.134$ . SPSS reports the same  $P$ -value. SAS reports a chi-square statistic with 3 degrees of freedom and  $P = 0.1344$ . All agree that there is not sufficient evidence in the data to reject the null hypothesis that the number of weeds per meter has no effect on the yield.



(a) Minitab



(b) SPSS



(c) SAS

**FIGURE 15.10**

Output from (a) Minitab, (b) SPSS, and (c) SAS for the Kruskal-Wallis test applied to the weed data, for Example 15.14.

### SECTION 15.3 Summary

The **Kruskal-Wallis test** compares several populations on the basis of independent random samples from each population. This is the **one-way analysis of variance** setting.

The null hypothesis for the Kruskal-Wallis test is that the distribution of the response variable is the same in all the populations. The alternative hypothesis is that responses are systematically larger in some populations than in others.

The **Kruskal-Wallis statistic  $H$**  can be viewed in two ways. It is essentially the result of applying one-way ANOVA to the ranks of the observations. It is also a comparison of the sums of the ranks for the several samples.

When the sample sizes are not too small and the null hypothesis is true,  $H$  for comparing  $I$  populations has approximately the chi-square distribution with  $I - 1$  degrees of freedom. We use this approximate distribution to obtain  $P$ -values.

## SECTION 15.3 Exercises

### 15.37 Number of Facebook friends.

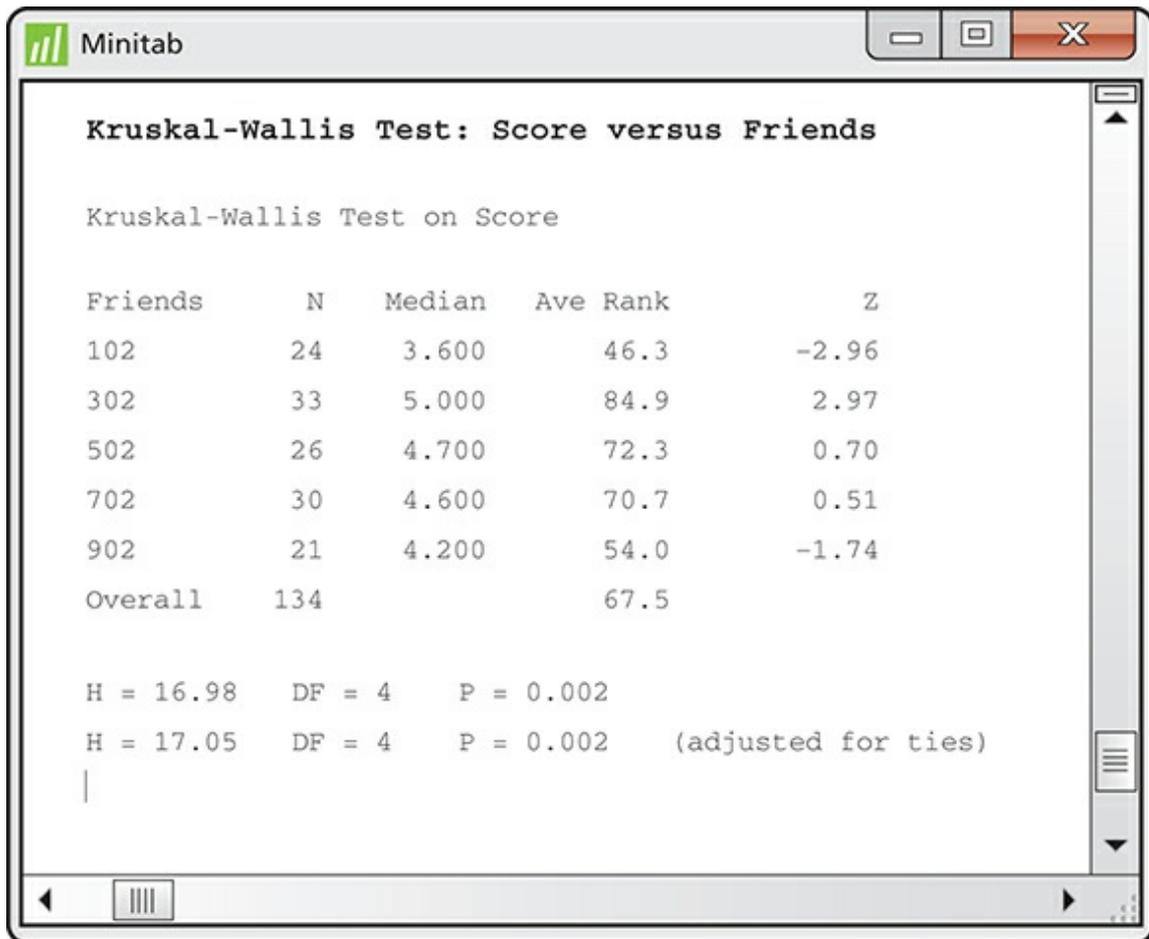
An experiment was run to examine the relationship between the number of Facebook friends and the user's perceived social attractiveness.<sup>15</sup> A total of 134 undergraduate participants were randomly assigned to observe one of five Facebook profiles. Everything about the profile was the same except the number of friends, which appeared on the profile as 102, 302, 502, 702, or 902. After viewing the profile, each participant was asked to fill out a questionnaire on the physical and social attractiveness of the profile user. Each attractiveness score is an average of several seven-point questionnaire items, ranging from 1 (strongly disagree) to 7 (strongly agree). In Example 12.3 (page 648), we analyzed these data using a one-way ANOVA. Explain the setting for this problem. Include the number of groups to be compared, assumptions about independence, and the distribution of the distributions.  FRIENDS

### 15.38 What are the hypotheses?

Refer to the previous exercise. What are the null hypothesis and the alternative hypothesis? Explain why a nonparametric procedure is appropriate in this setting.

### 15.39 Read the output.

Figure 15.11 gives the Minitab output for the analysis of the data described in Exercise 15.37. Describe the results given in the output and write a short summary of your conclusions from the analysis.



**FIGURE 15.11**

Output from Minitab for the Kruskal-Wallis test applied to the Facebook data, for Exercise 15.39.

### 15.40 Do we experience emotions differently?

In Exercise 12.37 (page 684) you analyzed data related to the way people from different cultures experience emotions. The study subjects were 410 college students from five different cultures. They were asked to record, on a 1 (never) to 7 (always) scale, how much of the time they typically felt eight specific emotions. These were averaged to produce the global emotion score for each participant. Analyze the data using the Kruskal-Wallis test and write a summary of your analysis and conclusions. Be sure to include your assumptions, hypotheses, and the results of the significance test.  **EMOTION**

### 15.41 Do isoflavones increase bone mineral density?

In Exercise 12.45 (page 686) you investigated the effects of isoflavones from kudzu on bone mineral density (BMD). The experiment randomized rats to three diets: control, low isoflavones, and high isoflavones. Here are the data:  **KUDZU**

Treatment	BMD (g/cm <sup>2</sup> )									
Control	0.228	0.207	0.234	0.220	0.217	0.228	0.209	0.221		
	0.204	0.220	0.203	0.219	0.218	0.245	0.210			
Low dose	0.211	0.220	0.211	0.233	0.219	0.233	0.226	0.228		

	0.216	0.225	0.200	0.208	0.198	0.208	0.203
High dose	0.250	0.237	0.217	0.206	0.247	0.228	0.245
	0.267	0.261	0.221	0.219	0.232	0.209	0.255

- (a) Use the Kruskal-Wallis test to compare the three diets.
- (b) How do these results compare with what you find using the ANOVA  $F$  statistic?

### 15.42 Vitamins in bread.

Does bread lose its vitamins when stored? Here are data on the vitamin C content (milligrams per 100 grams of flour) in bread baked from the same recipe and stored for 1, 3, 5, or 7 days.<sup>16</sup> The 10 observations are from 10 different loaves of bread.  **BREAD**

Condition	Vitamin C (mg/100 g)	
Immediately after baking	47.62	49.79
One day after baking	40.45	43.46
Three days after baking	21.25	22.34
Five days after baking	13.18	11.65
Seven days after baking	8.51	8.13

The loss of vitamin C over time is clear, but with only 2 loaves of bread for each storage time we wonder if the differences among the groups are significant.

- (a) Use the Kruskal-Wallis test to assess significance and then write a brief summary of what the data show.
- (b) Because there are only 2 observations per group, we suspect that the common chi-square approximation to the distribution of the Kruskal-Wallis statistic may not be accurate. The exact  $P$ -value (from SAS software) is  $P = 0.0011$ . Compare this with your  $P$ -value from part (a). Is the difference large enough to affect your conclusion?

### 15.43 Jumping and strong bones.

In Exercise 12.47 (page 687) you studied the effects of jumping on the bones of rats. Ten rats were assigned to each of three treatments: a 60-centimeter “high jump,” a 30-centimeter “low jump,” and a control group with no jumping.<sup>17</sup> Here are the bone densities (in milligrams per cubic centimeter) after eight weeks of 10 jumps per day:  **JUMP**

Group	Bone density (mg/cm <sup>3</sup> )				
Control	611	621	614	593	593
	653	600	554	603	569
Low jump	635	605	638	594	599
	632	631	588	607	596
High jump	650	622	626	626	631
	622	643	674	643	650

- (a) The study was a randomized comparative experiment. Outline the design of this experiment.

(b) Make side-by-side stemplots for the three groups, with the stems lined up for easy comparison. The distributions are a bit irregular but not strongly non-Normal. We would usually use analysis of variance to assess the significance of the difference in group means.

(c) Do the Kruskal-Wallis test. Explain the distinction between the hypotheses tested by Kruskal-Wallis and ANOVA.

(d) Write a brief statement of your findings. Include a numerical comparison of the groups as well as your test result.

### 15.44 Do poets die young?

In Exercise 12.46 (page 686) you analyzed the age at death for female writers. They were classified as novelists, poets, and nonfiction writers. The data are given in Table 12.1 (page 686).  **POETS**

(a) Use the Kruskal-Wallis test to compare the three groups of female writers.

(b) Compare these results with what you find using the ANOVA  $F$  statistic.

## CHAPTER 15 Exercises

### 15.45 Plants and hummingbirds.

Different varieties of the tropical flower *Heliconia* are fertilized by different species of hummingbirds. Over time, the lengths of the flowers and the forms of the hummingbirds' beaks have evolved to match each other. Here are data on the lengths in millimeters of three varieties of these flowers on the island of Dominica:<sup>18</sup>  **HBIRDS**

<i>H. bihai</i>					
47.12	46.75	46.81	47.12	46.67	47.43
46.44	46.64	48.07	48.34	48.15	50.26
50.12	46.34	46.94	48.36		
<i>H. caribaea red</i>					
41.90	42.01	41.93	43.09	41.47	41.69
39.78	40.57	39.63	42.18	40.66	37.87
39.16	37.40	38.20	38.07	38.10	37.97
38.79	38.23	38.87	37.78	38.01	
<i>H. caribaea yellow</i>					
36.78	37.02	36.52	36.11	36.03	35.45
38.13	37.10	35.17	36.82	36.66	35.68
36.03	34.57	34.63			

Do a complete analysis that includes description of the data and a rank test for the significance of the differences in lengths among the three species.

### 15.46 Time spent studying.

In Exercise 1.173 (page 50) you compared the time spent studying by men and women. The students in a large first-year college class were asked how many minutes they studied on a typical weeknight.

Here are the responses of random samples of 30 women and 30 men from the class:  **STIME**

Women	Men
170 120 180 360 240	80 120 30 90 200
120 180 120 240 170	90 45 30 120 75
150 120 180 180 150	150 120 60 240 300
200 150 180 150 180	240 60 120 60 30
120 60 120 180 180	30 230 120 95 150
90 240 180 115 120	0 200 120 120 180

(a) Summarize the data numerically and graphically.

(b) Use the Wilcoxon rank sum test to compare the men and women. Write a short summary of your results.

- (c) Use a two-sample  $t$  test to compare the men and women. Write a short summary of your results.
- (d) Which procedure is more appropriate for these data? Give reasons for your answer.

### 15.47 Response times for telephone repair calls.

A study examined the time required for the telephone company Verizon to respond to repair calls from its own customers and from customers of a CLEC, another phone company that pays Verizon to use its local lines. Here are the data, which are rounded to the nearest hour:  **TREPAIR**

Verizon										
1	1	1	1	2	2	1	1	1	2	2
1	1	1	1	2	2	1	1	1	2	3
1	1	1	1	2	3	1	1	1	2	3
1	1	1	1	2	3	1	1	1	2	3
1	1	1	1	2	3	1	1	1	2	4
1	1	1	1	2	5	1	1	1	2	5
1	1	1	1	2	6	1	1	1	2	8
1	1	1	1	2	15	1	1	1	2	2
CLEC										
1	1	5	5	5	1	5	5	5	5	

- (a) Does Verizon appear to give CLEC customers the same level of service as its own customers? Compare the data using graphs and descriptive measures and express your opinion.
- (b) We would like to see if times are significantly longer for CLEC customers than for Verizon customers. Why would you hesitate to use a  $t$  test for this purpose? Carry out a rank test. What can you conclude?
- (c) Explain why a nonparametric procedure is appropriate in this setting.

*Iron-deficiency anemia is the most common form of malnutrition in developing countries. Does the type of cooking pot affect the iron content of food? We have data from a study in Ethiopia that measured the iron content (milligrams per 100 grams of food) for three types of food cooked in each of three types of pots.*<sup>19</sup>  **COOK**

Type of Pot	Iron Content			
	Meat			
Aluminum	1.77	2.36	1.96	2.14
Clay	2.27	1.28	2.48	2.68
Iron	5.27	5.17	4.06	4.22
	Legumes			
Aluminum	2.40	2.17	2.41	2.34
Clay	2.41	2.43	2.57	2.48
Iron	3.69	3.43	3.84	3.72
	Vegetables			
Aluminum	1.03	1.53	1.07	1.30
Clay	1.55	0.79	1.68	1.82
Iron	2.45	2.99	2.80	2.92

Exercises 15.48 to 15.50 use these data.

### 15.48 Cooking vegetables in different pots.

Does the vegetable dish vary in iron content when cooked in aluminum, clay, and iron pots?   
**COOK**

- (a) What do the data appear to show? Check the conditions for one-way ANOVA. Which requirements are a bit dubious in this setting?
- (b) Instead of ANOVA, do a rank test. Summarize your conclusions about the effect of pot material on the iron content of the vegetable dish.

### 15.49 Cooking meat and legumes in aluminum and clay pots.

There appears to be little difference between the iron content of food cooked in aluminum pots and food cooked in clay pots. Is there a significant difference between the iron content of meat cooked in aluminum and clay? Is the difference between aluminum and clay significant for legumes? Use rank tests.  **COOK**

### 15.50 Iron in food cooked in iron pots.

The data show that food cooked in iron pots has the highest iron content. They also suggest that the three types of food differ in iron content. Is there significant evidence that the three types of food differ in iron content when all are cooked in iron pots?  **COOK**

### 15.51 Multiple comparisons for plants and hummingbirds.

As in ANOVA, we often want to carry out a **multiple-comparisons** procedure following a Kruskal-Wallis test to tell us *which* groups differ significantly.<sup>20</sup> The Bonferroni method (page 670) is a simple method: If we carry out  $k$  tests at fixed significance level  $0.05/k$ , the probability of *any* false rejection among the  $k$  tests is always no greater than 0.05. That is, to get overall significance level 0.05 for all of  $k$  comparisons, do each individual comparison at the  $0.05/k$  level. In Exercise 15.45 you found a significant difference among the lengths of three varieties of the flower *Heliconia*. Now we will explore multiple comparisons.  **HBIRDS**

- (a) Write down all the pairwise comparisons we can make, for example, *bihai* versus *caribaea* red. There are three possible pairwise comparisons.
- (b) Carry out three Wilcoxon rank sum tests, one for each of the three pairs of flower varieties. What are the three two-sided  $P$ -values?
- (c) For purposes of multiple comparisons, any of these three tests is significant if its  $P$ -value is no greater than  $0.05/3 = 0.0167$ . Which pairs differ significantly at the overall 0.05 level?

### 15.52 Multiple comparisons for cooking pots.

The previous exercise outlines how to use the Wilcoxon rank sum test several times for multiple comparisons with overall significance level 0.05 for all comparisons together. Apply this procedure

to the data used in each of Exercises 15.48 to 15.50.  **COOK**

## CHAPTER 15 Notes and Data Sources

1. *Condé Nast Traveler* readers poll data for 2013, from [cntraveler.com/spas/2013/03/best-spas-united-states-caribbean-mexico-cruise-ships](http://cntraveler.com/spas/2013/03/best-spas-united-states-caribbean-mexico-cruise-ships).

2. For purists, here is the precise definition:  $X_1$  is *stochastically larger* than  $X_2$  if

$$P(X_1 > a) \geq P(X_2 > a)$$

for all  $a$ , with strict inequality for at least one  $a$ . The Wilcoxon rank sum test is effective against this alternative in the sense that the power of the test approaches 1 (that is, the test becomes more certain to reject the null hypothesis) as the number of observations increases.

3. Erin K. O’Loughlin et al., “Prevalence and correlates of exergaming in youth,” *Pediatrics*, 130 (2012), pp. 806–814.

4. From the PEW Internet and American Life website, [pewinternet.org/Reports/2013/Civic-Engagement.aspx](http://pewinternet.org/Reports/2013/Civic-Engagement.aspx).

5. From Matthias R. Mehl et al., “Are women really more talkative than men?,” *Science*, 317, No. 5834 (2007), p. 82. The raw data were provided by Matthias Mehl.

6. Data provided by Warren Page, New York City Technical College, from a study done by John Hudesman.

7. Data provided by Susan Stadler, Purdue University.

8. *Ibid.*

9. The vehicle is a 2002 Toyota Prius owned by the third author.

10. Statistics regarding Facebook usage can be found at [facebook.com/notes/facebook-data-team/anatomy-of-facebook/10150388519243859](https://www.facebook.com/notes/facebook-data-team/anatomy-of-facebook/10150388519243859).

11. These data were collected as part of a larger study of dementia patients conducted by Nancy Edwards, School of Nursing, and Alan Beck, School of Veterinary Medicine, Purdue University.

12. Data provided by Diana Schellenberg, Purdue University School of Health Sciences.

13. These data are from “Results report on the vitamin C pilot program,” prepared by SUSTAIN (Sharing United States Technology to Aid in the Improvement of Nutrition) for the U.S. Agency for International Development. The report was used by the Committee on International Nutrition of the National Academy of Sciences/Institute of Medicine to make recommendations on whether or not the vitamin C content of food commodities used in U.S. food aid programs should be increased. The program was directed by Peter Ranum and Françoise Chomé. The second author was a member of the committee.

14. Data provided by Sam Phillips, Purdue University.

15. See Note 10.

16. Data provided by Helen Park. See H. Park et al., “Fortifying bread with each of three antioxidants,” *Cereal Chemistry*, 74 (1997), pp. 202–206.

17. Data provided by Jo Welch, Purdue University Department of Foods and Nutrition.

18. We thank Ethan J. Temeles of Amherst College for providing the data. His work is described in Ethan J. Temeles and W. John Kress, “Adaptation in a plant-hummingbird association,” *Science*, 300 (2003), pp. 630–633.

19. Based on A. A. Adish et al., “Effect of consumption of food cooked in iron pots on iron status and growth of young children: A randomised trial,” *The Lancet*, 353 (1999), pp. 712–716.

**20.** For more details on multiple comparisons, see M. Hollander and D. A. Wolfe, *Nonparametric Statistical Methods*, 2nd ed., Wiley, 1999. This book is a useful reference on applied aspects of nonparametric inference in general.

# 16 Bootstrap Methods and Permutation Tests\*

CHAPTER



- 16.1 The Bootstrap Idea
- 16.2 First Steps in Using the Bootstrap
- 16.3 How Accurate Is a Bootstrap Distribution?
- 16.4 Bootstrap Confidence Intervals
- 16.5 Significance Testing Using Permutation Tests

---

\* The original version of this chapter was written by Tim Hesterberg, David S. Moore, Shaun Monaghan, Ashley Clipson, and Rachel Epstein, with support from the National Science Foundation under grant DMI-0078706. Revisions have been made by Bruce A. Craig and George P. McCabe. Special thanks to Bob Thurman, Richard Heiberger, Laura Chihara, Tom Moore, and Gudmund Iversen for helpful

comments on an earlier version.

## Introduction

The continuing revolution in computing is having a dramatic influence on statistics. The exploratory analysis of data is becoming easier as more graphs and calculations are automated. The statistical study of very large and very complex data sets is now feasible. Another impact of this fast and inexpensive computing is less obvious: new methods apply previously unthinkable amounts of computation to produce confidence intervals and tests of significance in settings that don't meet the conditions for safe application of the usual methods of inference.

Consider the commonly used  $t$  procedures for inference about means (Chapter 7) and for relationships between quantitative variables (Chapter 10). All these methods rest on the use of Normal distributions for data. While no data are exactly Normal, the  $t$  procedures are useful in practice because they are *robust*. Nonetheless, we cannot use  $t$  confidence intervals and tests if the data are strongly skewed, unless our samples are quite large.



**robust, p. 432**

Other procedures cannot be used on non-Normal data even when the samples are large. Inference about spread based on Normal distributions is *not robust* and therefore of little use in practice.



**$F$  test for equality of spread, p. 474**

Finally, what should we do if we are interested in, say, a *ratio* of means, such as the ratio of average men's salary to average women's salary? There is no simple traditional inference method for this setting.

The methods of this chapter—bootstrap confidence intervals and permutation tests—apply the power of the computer to relax some of the conditions needed for traditional inference and to do inference in new settings. The big ideas of statistical inference remain the same. The fundamental reasoning is still based on asking, "What would happen if we applied this method many times?" Answers to this question are still given by confidence levels and  $P$ -values based on the sampling distributions of statistics.

The most important requirement for trustworthy conclusions about a population is still that our data can be regarded as random samples from the population—not even the computer can rescue voluntary response samples or confounded experiments. But the new methods set us free from the need for Normal data or large samples. They work the same way for many different statistics in many

different settings. They can, with sufficient computing power, give results that are more accurate than those from traditional methods.

Bootstrap intervals and permutation tests are conceptually simple because they appeal directly to the basis of all inference: the sampling distribution that shows what would happen if we took very many samples under the same conditions. The new methods do have limitations, some of which we will illustrate. But their effectiveness and range of use are so great that they are now widely used in a variety of settings.

## Software

Bootstrapping and permutation tests are feasible in practice only with software that automates the heavy computation that these methods require. If you are sufficiently expert, you can program at least the basic methods yourself. It is easier to use software that offers bootstrap intervals and permutation tests preprogrammed, just as most software offers the various  $t$  intervals and tests. You can expect the new methods to become more common in standard statistical software.

This chapter primarily uses R, the software choice of many statisticians doing research on resampling methods.<sup>1</sup> There are several packages of functions for resampling in R. We will focus on the `boot` package, which offers the most capabilities. Unlike software such as Minitab and SPSS, R is not menu driven and requires command line requests to load data and access various functions. All commands used in this chapter are available on the text website.

SPSS and SAS also offer preprogrammed bootstrap and permutation methods. SPSS has an auxiliary bootstrap module that contains most of the methods described in this chapter. In SAS, the SURVEYSELECT procedure can be used to do the necessary resampling. The bootstrap macro contains most of the confidence interval methods offered by R. You can find links for downloading these modules or macros on the text website.

## 16.1 The Bootstrap Idea

When you complete this section, you will be able to

- Randomly select bootstrap resamples from a small sample using software and a table of random numbers.
- Find the bootstrap standard error from a collection of resamples.
- Use computer output to describe the results of a bootstrap analysis of the mean.

Here is the example we will use to introduce these methods.

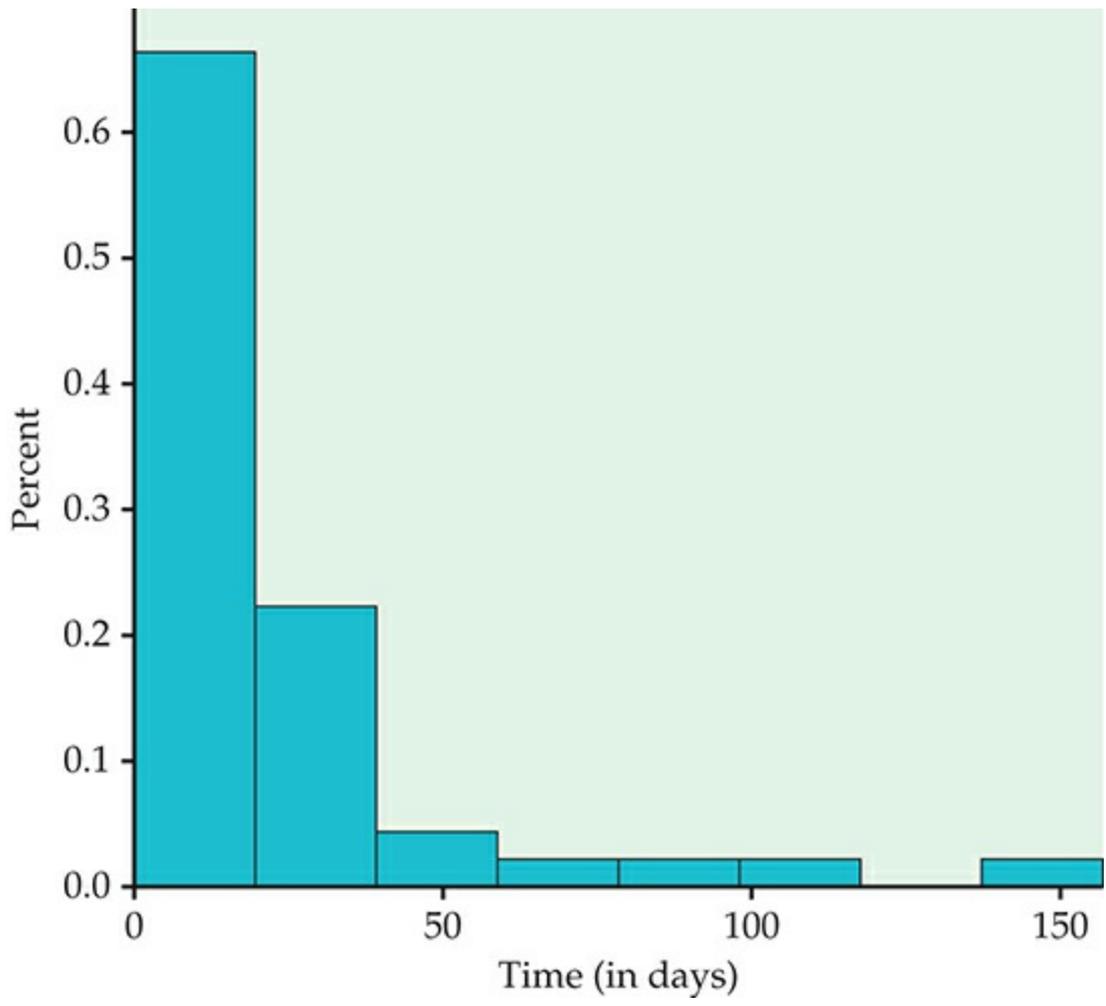
### EXAMPLE

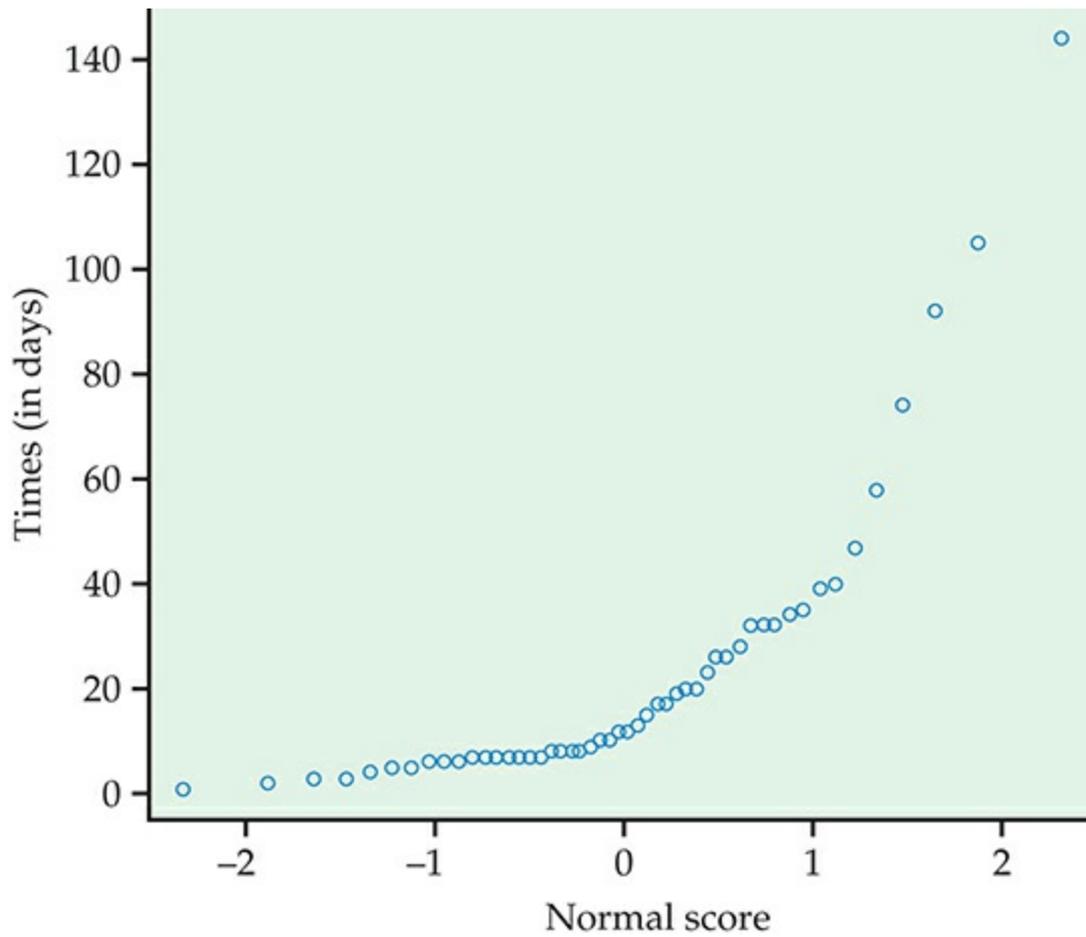
#### 16.1 Time to start a business.



The World Bank collects information about starting businesses throughout the world. They have determined the time, in days, to complete all the procedures required to start a business. For this example, we use the times to start a business for a random sample of 50 countries included in the World Bank survey.

Figure 16.1(a) gives a histogram and Figure 16.1(b) gives the Normal quantile plot. The data are strongly skewed to the right. The median is 12 days and the mean is almost twice as large, 23.26 days. We have some concerns about using the  $t$  procedures for these data.





**FIGURE 16.1**

(a) The distribution of 50 times to start a business. (b) Normal quantile plot of the times to start a business, for Example 16.1. The distribution is strongly right-skewed.

## The big idea: resampling and the bootstrap distribution

Statistical inference is based on the sampling distributions of sample statistics. A sampling distribution is based on many random samples from the population. The bootstrap is a way of finding the sampling distribution, at least approximately, from just one sample. Here is the procedure:



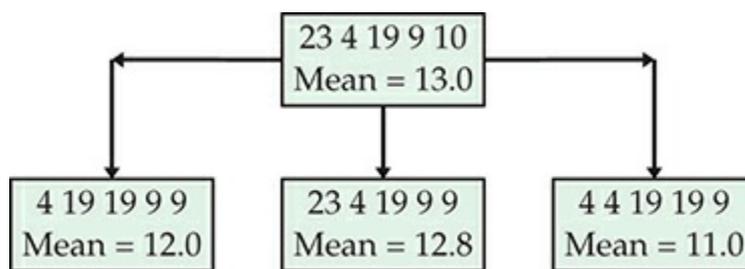
**sampling distribution, p. 302**

**Step 1: Resampling.** In Example 16.1, we have just one random sample. In place of many samples from the population, create many **resamples** by repeatedly sampling *with replacement* from this one random sample. Each resample is the same size as the original random sample.

resamples

**Sampling with replacement** means that after we randomly draw an observation from the original sample, we put it back before drawing the next observation. Think of drawing a number from a hat and then putting it back before drawing again. As a result, any number can be drawn more than once. If we sampled *without* replacement, we'd get the same set of numbers we started with, though in a different order. Figure 16.2 illustrates three resamples from a sample of five observations. In practice, we draw hundreds or thousands of resamples, not just three.

sampling with replacement



**FIGURE 16.2**

The resampling idea. The top box is a sample of size  $n = 5$  from the time to start a business data. The three lower boxes are three resamples from this original sample. Some values from the original sample are repeated in the resamples because each resample is formed by sampling with replacement. We calculate the statistic of interest, the sample mean in this example, for the original sample and each resample.

**Step 2: Bootstrap distribution.** The sampling distribution of a statistic collects the values of the statistic from the many samples of the population. The **bootstrap distribution** of a statistic collects its values from the many resamples. The bootstrap distribution gives information about the sampling distribution.

bootstrap distribution

## THE BOOTSTRAP IDEA

The original sample is representative of the population from which it was drawn. Thus, resamples from this original sample represent what we would get if we took many samples from the population. The **bootstrap distribution** of a statistic, based on the resamples, represents the sampling distribution of the statistic.

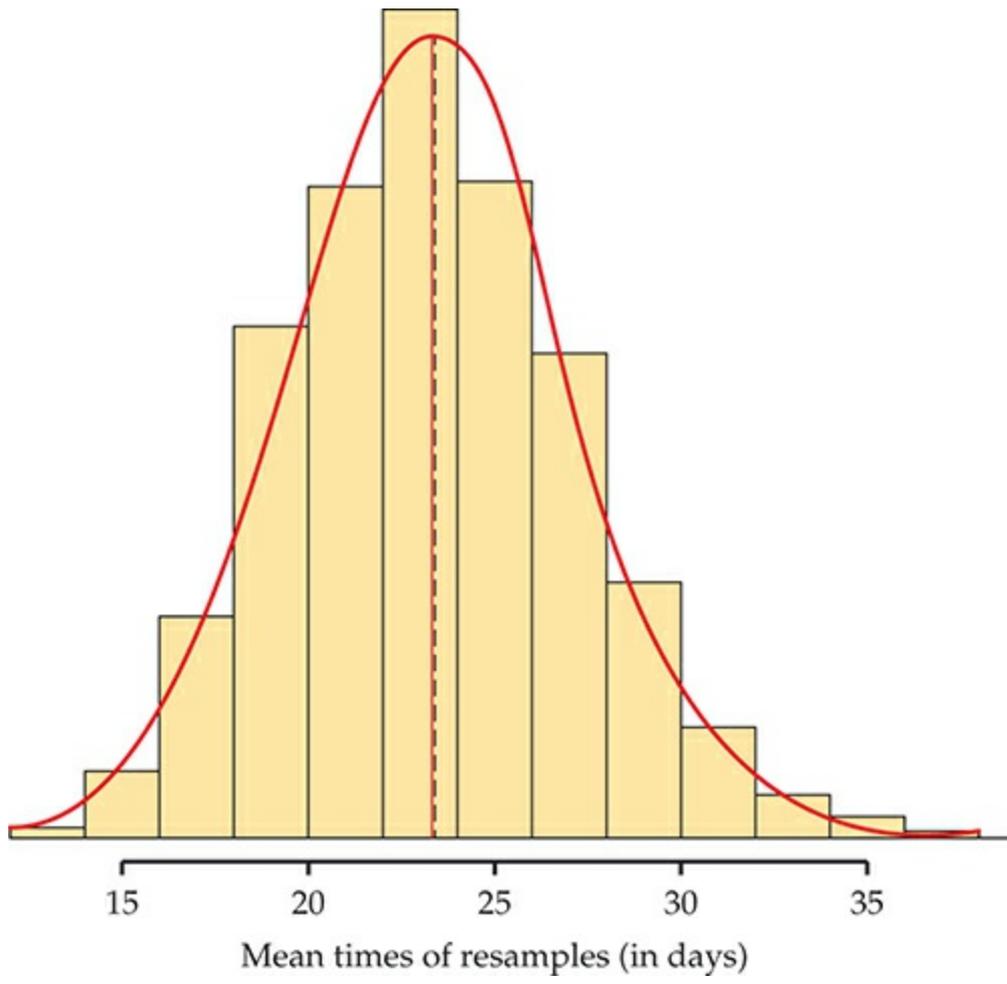
## EXAMPLE

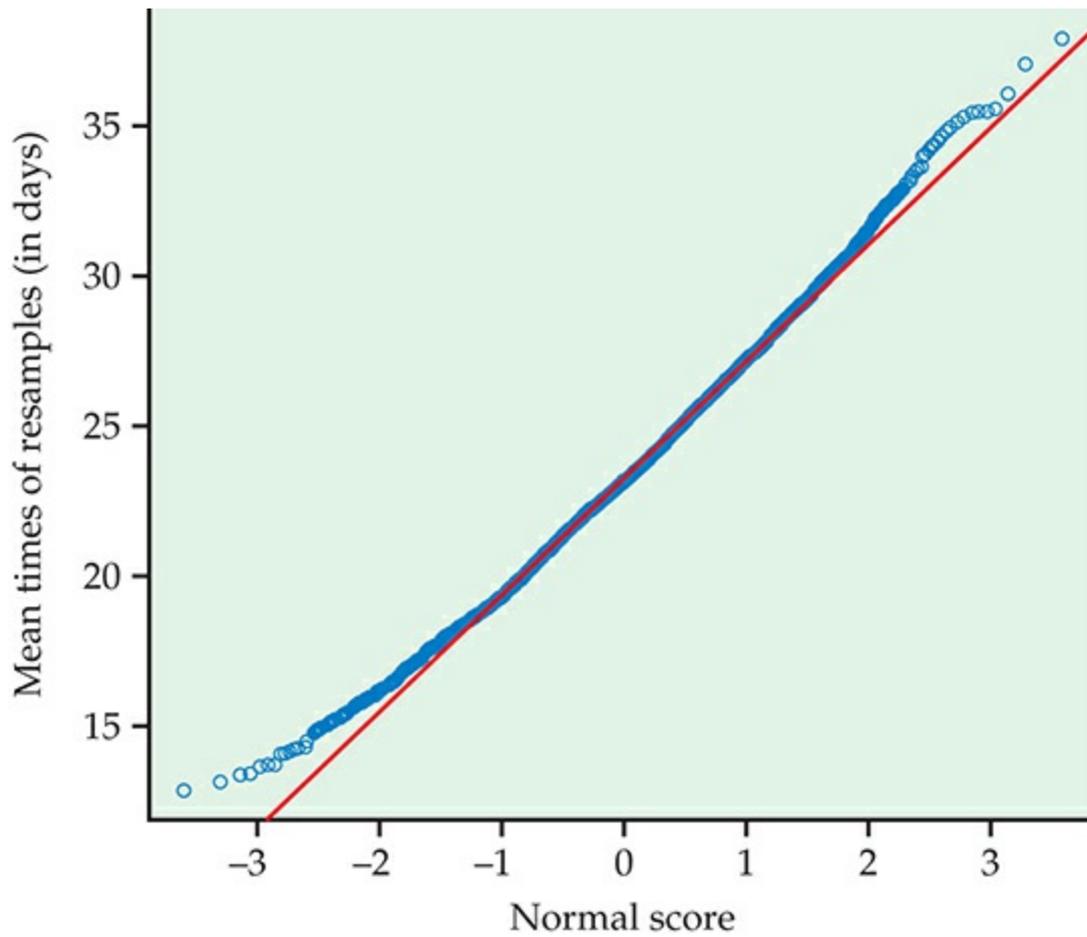
### 16.2 Bootstrap distribution of mean time to start a business.



In Example 16.1, we want to estimate the population mean time to start a business,  $\mu$ , so the statistic is the sample mean  $\bar{x}$ . For our one random sample of 50 times,  $\bar{x} = 23.26$  days. When we resample, we get different values of  $\bar{x}$ , just as we would if we took new samples from the population of all times to start a business.

We randomly generated 3000 resamples for these data. The mean for the resamples is 23.30 days and the standard deviation is 3.85. Figure 16.3(a) gives a histogram of the bootstrap distribution of the means of 3000 resamples from the time to start a business data. The Normal density curve with the mean 23.30 and standard deviation 3.85 is superimposed on the histogram. A Normal quantile plot is given in Figure 16.3(b). The distribution of the resample means is approximately Normal, although a small amount of skewness is still evident.





**FIGURE 16.3**

(a) The bootstrap distribution of 3000 resample means from the sample of times to start a business. The smooth curve is the Normal density function for the distribution that matches the mean and standard deviation of the distribution of the resample means. (b) The Normal quantile plot confirms that the bootstrap distribution is somewhat skewed to the right but fits the Normal distribution quite well.

According to the bootstrap idea, the bootstrap distribution represents the sampling distribution. Let's compare the bootstrap distribution with what we know about the sampling distribution.

← **LOOK BACK**

**central limit theorem, p. 307**

**Shape:** We see that the bootstrap distribution is nearly Normal. The central limit theorem says that the sampling distribution of the sample mean  $\bar{x}$  is approximately Normal if  $n$  is large. So the bootstrap distribution shape is close to the shape we expect the sampling distribution to have.

← **LOOK BACK**

**mean and standard deviation of  $\bar{x}$ , p. 306**

**Center:** The bootstrap distribution is centered close to the mean of the original sample, 23.30 days versus 23.26 days for the original sample. Therefore, the mean of the bootstrap distribution has little bias as an estimator of the mean of the original sample. We know that the sampling distribution of  $\bar{x}$  is centered at the population mean  $\mu$ , that is, that  $\bar{x}$  is an unbiased estimate of  $\mu$ . So the resampling distribution behaves (starting from the original sample) as we expect the sampling distribution to behave (starting from the population).

**Spread:** The histogram and density curve in Figure 16.3(a) picture the variation among the resample means. We can get a numerical measure by calculating their standard deviation. Because this is the standard deviation of the 3000 values of  $\bar{x}$  that make up the bootstrap distribution, we call it the **bootstrap standard error** of  $\bar{x}$ . The numerical value is 3.85. In fact, we know that the standard deviation of  $\bar{x}$  is  $\sigma/n$ , where  $\sigma$  is the standard deviation of individual observations in the population. Our usual estimate of this quantity is the standard error of  $\bar{x}$ ,  $s/n$ , where  $s$  is the standard deviation of our one random sample. For these data,  $s = 28.20$  and

bootstrap standard error

$$s/n = 28.2050/7.07 = 3.99$$

The bootstrap standard error 3.85 is relatively close to the theory-based estimate 3.99.

In discussing Example 16.2, we took advantage of the fact that statistical theory tells us a great deal about the sampling distribution of the sample mean  $\bar{x}$ . We found that the bootstrap distribution created by resampling matches the properties of this sampling distribution. The heavy computation needed to produce the bootstrap distribution replaces the heavy theory (central limit theorem, mean, and standard deviation of  $\bar{x}$ ) that tells us about the sampling distribution.

*The great advantage of the resampling idea is that it often works even when theory fails.* Of course, theory also has its advantages: we know exactly when it works. We don't know exactly when resampling works, so that "When can I safely bootstrap?" is a somewhat subtle issue.

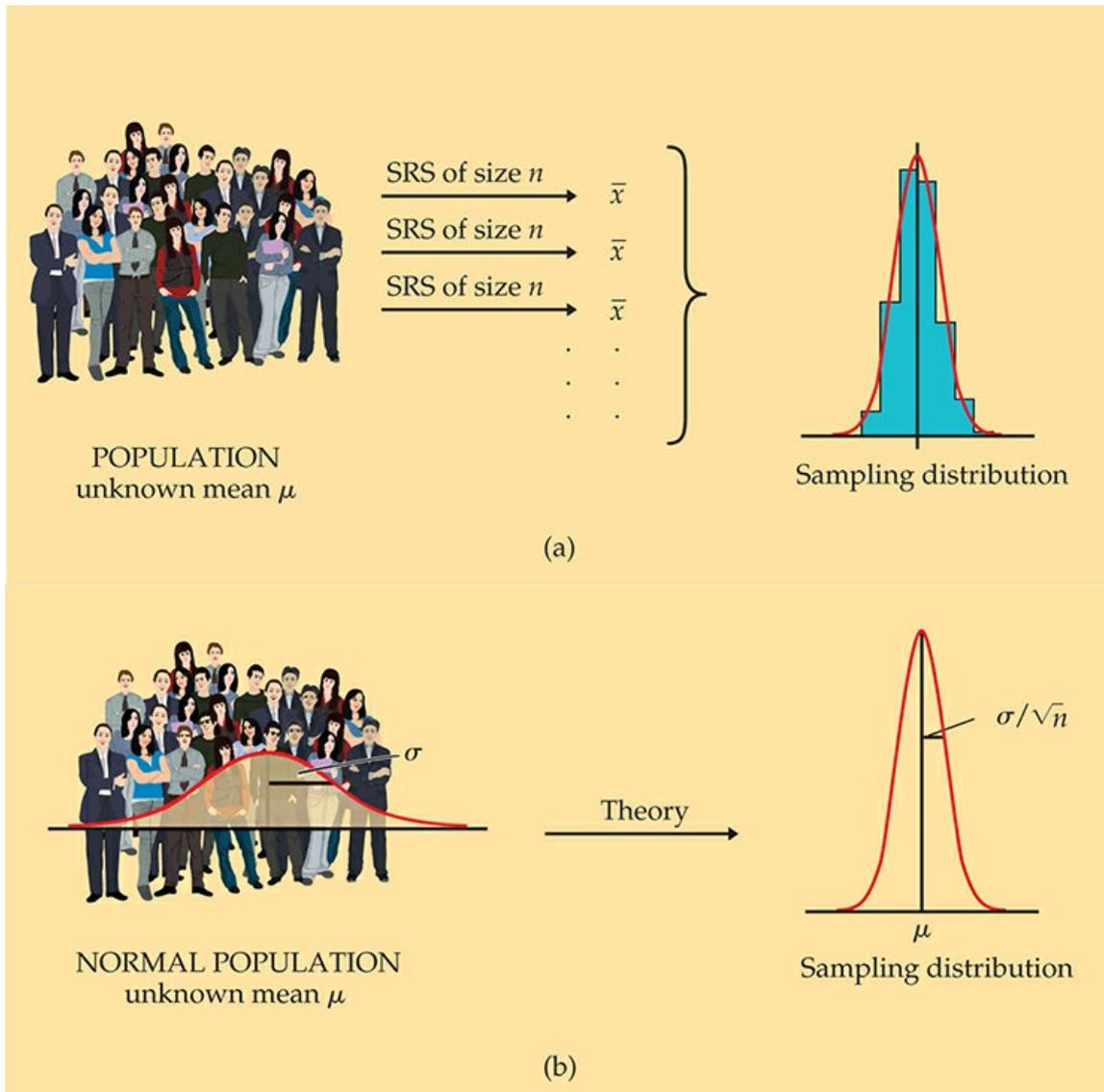
Figure 16.4 illustrates the bootstrap idea by comparing three distributions. Figure 16.4(a) shows the idea of the sampling distribution of the sample mean  $\bar{x}$ : take many random samples from the population, calculate the mean  $\bar{x}$  for each sample, and collect these  $\bar{x}$ -values into a distribution.

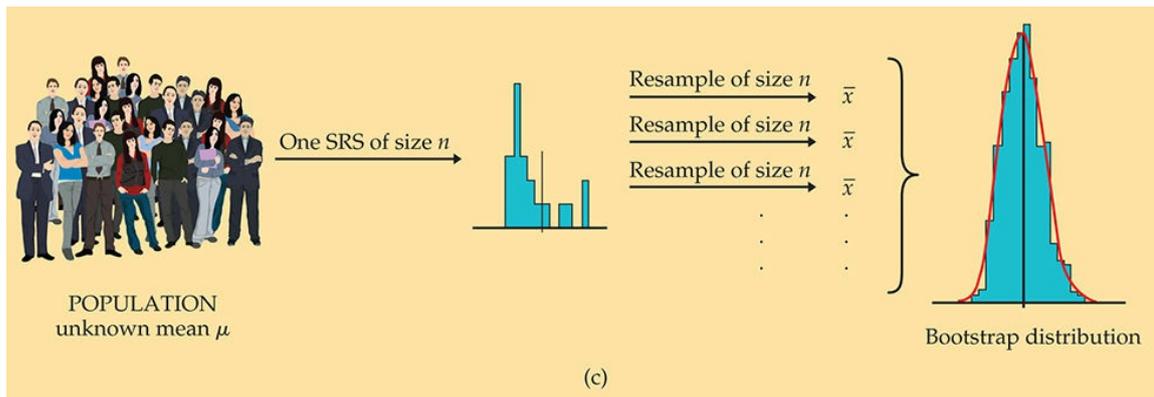
Figure 16.4(b) shows how traditional inference works: statistical theory tells us that if the population has a Normal distribution, then the sampling distribution of  $\bar{x}$  is also Normal. If the population is not Normal but our sample is large, we can use the central limit theorem. If  $\mu$  and  $\sigma$  are the mean and standard deviation of the population, the sampling distribution of  $\bar{x}$  has mean  $\mu$  and standard deviation  $\sigma/n$ . When it is available, theory is wonderful: we know the sampling distribution without the impractical task of actually taking many samples from the population.

← LOOK BACK

central limit theorem, p. 307

Figure 16.4(c) shows the bootstrap idea: we avoid the task of taking many samples from the population by instead taking many resamples from a single sample. The values of  $\bar{x}$  from these resamples form the bootstrap distribution. We use the bootstrap distribution rather than theory to learn about the sampling distribution.





**FIGURE 16.4**

(a) The idea of the sampling distribution of the sample mean  $\bar{x}$ : take very many samples, collect the  $\bar{x}$ -values from each, and look at the distribution of these values. (b) The theory shortcut: if we know that the population values follow a Normal distribution, theory tells us that the sampling distribution of  $\bar{x}$  is also Normal. (c) The bootstrap idea: when theory fails and we can afford only one sample, that sample stands in for the population, and the distribution of  $\bar{x}$  in many resamples stands in for the sampling distribution.

## USE YOUR KNOWLEDGE

### 16.1 A small bootstrap example.

To illustrate the bootstrap procedure, let's bootstrap a small random subset of the time to start a business data:

8 3 10 47 7 32

- Sample *with replacement* from this initial SRS by rolling a die. Rolling a 1 means select the first member of the SRS, a 2 means select the second member, and so on. (You can also use Table B of random digits, responding only to digits 1 to 6.) Create 20 resamples of size  $n = 6$ .
- Calculate the sample mean for each of the resamples.
- Make a stemplot of the means of the 20 resamples. This is the bootstrap distribution.
- Calculate the bootstrap standard error.

### 16.2 Standard deviation versus standard error.

Explain the difference between the standard deviation of a sample and the standard error of a statistic such as the sample mean.

## Thinking about the bootstrap idea

It might appear that resampling creates new data out of nothing. This seems suspicious. Even the name “bootstrap” comes from the impossible image of “pulling yourself up by your own bootstraps.”<sup>2</sup> But the resampled observations are not used as if they were new data. The bootstrap distribution of the re-sample means is used only to estimate how the sample mean of one actual sample of size 50 would vary because of random sampling.

Using the same data for two purposes—to estimate a parameter and also to estimate the variability of the estimate—is perfectly legitimate. We do exactly this when we calculate  $\bar{x}$  to estimate  $\mu$  and then calculate  $s/\sqrt{n}$  from the same data to estimate the variability of  $\bar{x}$ .

What is new? First of all, we don’t rely on the formula  $s/\sqrt{n}$  to estimate the standard deviation of  $\bar{x}$ . Instead, we use the ordinary standard deviation of the many  $\bar{x}$ -values from our many resamples.<sup>3</sup> Suppose that we take  $B$  resamples and call the means of these resamples  $\bar{x}^*$  to distinguish them from the mean  $\bar{x}$  of the original sample. We would then find the mean and standard deviation of the  $\bar{x}^*$ ’s in the usual way.

To make clear that these are the mean and standard deviation of the means of the  $B$  resamples rather than the mean  $\bar{x}$  and standard deviation  $s$  of the original sample, we use a distinct notation:



**describing distributions with numbers, p. 30**

$$\text{meanboot} = \frac{1}{B} \sum \bar{x}^*$$

$$\text{SEboot} = \sqrt{\frac{1}{B-1} \sum (\bar{x}^* - \text{meanboot})^2}$$

These formulas go all the way back to Chapter 1. Once we have the values  $\bar{x}^*$ , we can just ask our software for their mean and standard deviation.

Because we will often apply the bootstrap to statistics other than the sample mean, here is the general definition for the bootstrap standard error.

**BOOTSTRAP STANDARD ERROR**

The **bootstrap standard error**  $\text{SE}_{\text{boot}}$  of a statistic is the standard deviation of the bootstrap distribution of that statistic.

Another thing that is new is that we don’t appeal to the central limit theorem or other theory to tell us that a sampling distribution is roughly Normal. We look at the bootstrap distribution to see if it is roughly Normal (or not). In most cases, the bootstrap distribution has approximately the same shape and spread as the sampling distribution, but it is centered at the original sample statistic value rather

than the parameter value.

In summary, the bootstrap allows us to calculate standard errors for statistics for which we don't have formulas and to check Normality for statistics that theory doesn't easily handle. To apply the bootstrap idea, we must start with a statistic that estimates the parameter we are interested in. We come up with a suitable statistic by appealing to another principle that we have often applied without thinking about it.

## THE PLUG-IN PRINCIPLE

To estimate a parameter, a quantity that describes the population, use the statistic that is the corresponding quantity for the sample.

The plug-in principle tells us to estimate a population mean  $\mu$  by the sample mean  $\bar{x}$  and a population standard deviation  $\sigma$  by the sample standard deviation  $s$ . Estimate a population median by the sample median and a population regression line by the least-squares line calculated from a sample. The bootstrap idea itself is a form of the plug-in principle: substitute the data for the population and then draw samples (resamples) to mimic the process of building a sampling distribution.

## Using software

Software is essential for bootstrapping in practice. Here is an outline of the program you would write if your software can choose random samples from a set of data but does not have bootstrap functions:

```
Repeat B times {  
  Draw a resample with replacement from the data.  
  Calculate the resample statistic.  
  Save the resample statistic into a variable.  
}  
Make a histogram and Normal quantile plot of the B  
  resample statistics.  
Calculate the standard deviation of the B statistics.
```

## EXAMPLE

## 16.3 Using software.

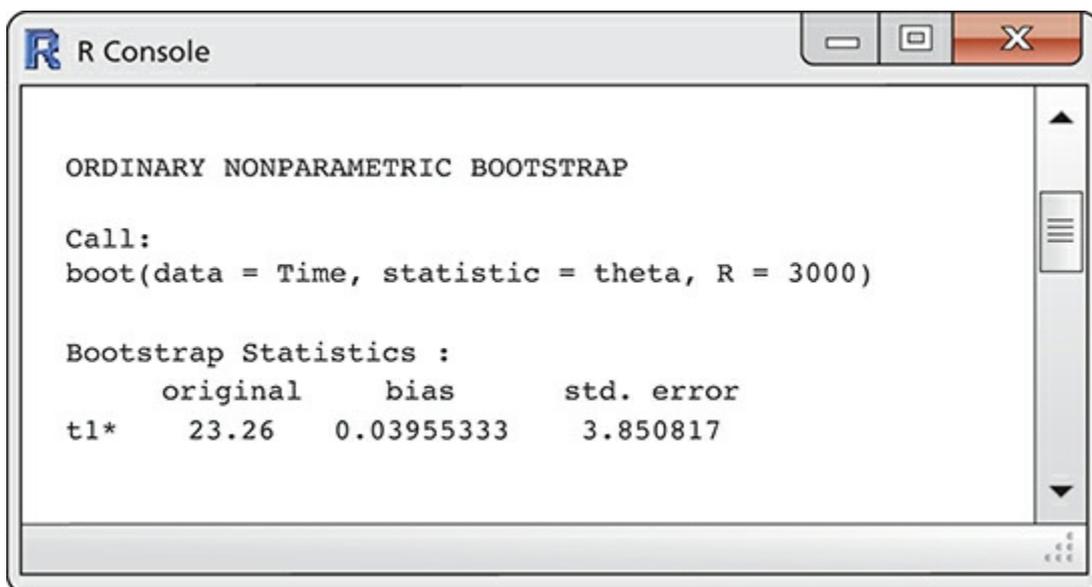


R has packages that contain various bootstrap functions so we do not have to write them ourselves. If the 50 times to start a business times are saved as a variable, we can use functions to resample from the data, calculate the means of the resamples, and request both graphs and printed output. We can also ask that the bootstrap results be saved for later access.

The function `plot.boot` will generate graphs similar to those in Figure 16.3 so you can assess Normality. Figure 16.5 contains the default output from a call of the function `boot`. The variable `Time` contains the 50 starting times, the function `theta` is specified to be the mean, and we request 3000 resamples. The original entry gives the mean  $\bar{x}=23.26$  of the original sample. Bias is the difference between the mean of the resample means and the original mean. If we add the entries for bias and original we get the mean of the resample means, `meanboot`:

$$23.26 + 0.04 = 23.30$$

The bootstrap standard error is displayed under `std.error`. All these values except original will differ a bit if you take another 3000 resamples, because resamples are drawn at random.



```
ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = Time, statistic = theta, R = 3000)

Bootstrap Statistics :
      original      bias      std. error
t1*    23.26    0.03955333    3.850817
```

**FIGURE 16.5**

R output for the time to start a business bootstrap, for Example 16.3.

## SECTION 16.1 Summary

To bootstrap a statistic such as the sample mean, draw hundreds of **resamples** with replacement from a single original sample, calculate the statistic for each resample, and inspect the **bootstrap distribution** of the resample statistics.

A bootstrap distribution approximates the sampling distribution of the statistic. This is an example of the **plug-in principle**: use a quantity based on the sample to approximate a similar quantity from the population.

A bootstrap distribution usually has approximately the same shape and spread as the sampling distribution. It is centered at the statistic (from the original sample) when the sampling distribution is centered at the parameter (of the population).

Use graphs and numerical summaries to determine whether the bootstrap distribution is approximately Normal and centered at the original statistic, and to get an idea of its spread. The **bootstrap standard error** is the standard deviation of the bootstrap distribution.

The bootstrap does not replace or add to the original data. We use the bootstrap distribution as a way to estimate the variation in a statistic based on the original data.

## SECTION 16.1 Exercises

For Exercises 16.1 and 16.2, see page 16-8.

### 16.3 Gosset's data on double stout sales.

William Sealy Gosset worked at the Guinness Brewery in Dublin and made substantial contributions to the practice of statistics. In Exercise 1.61 (page 48), we examined Gosset's data on the change in the double stout market before and after World War I (1914–1918). For various regions in England and Scotland, he calculated the ratio of sales in 1925, after the war, as a percent of sales in 1913, before the war. Here are the data for a sample of six of the regions in the original data:  **STOUT6**

Bristol	94	Glasgow	66
English P	46	Liverpool	140
English Agents	78	Scottish	24

- Do you think that these data appear to be from a Normal distribution? Give reasons for your answer.
- Select five resamples from this set of data.
- Compute the mean for each resample.

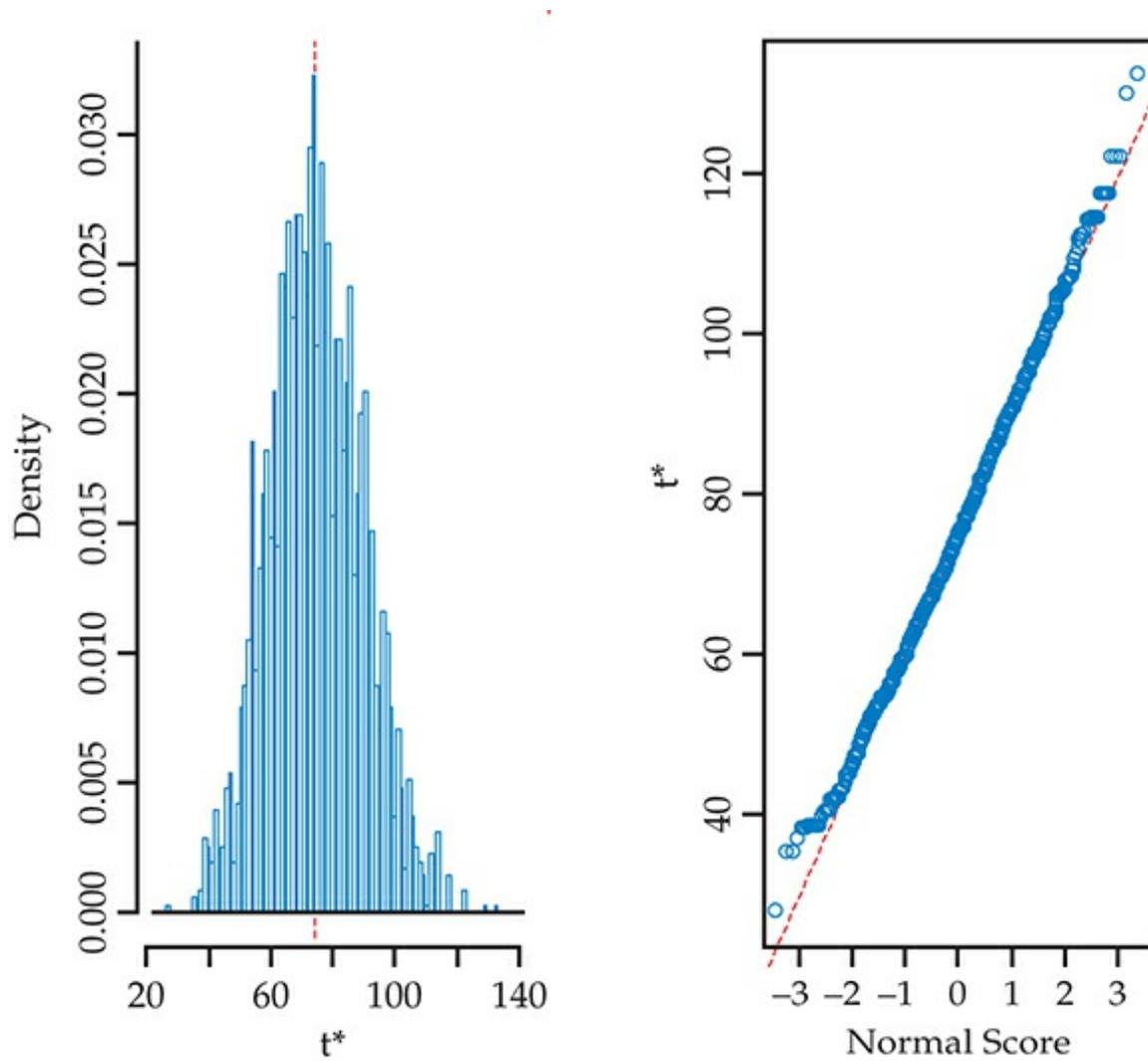
### 16.4 Find the bootstrap standard error.

Refer to your work in the previous exercise.  **STOUT6**

- Would you expect the bootstrap standard error to be larger, smaller, or approximately equal to the standard deviation of the original sample of six regions? Explain your answer.
- Find the bootstrap standard error.

## 16.5 Read the output.

Figure 16.6 gives a histogram and a Normal quantile plot for 3000 resample means from R. Interpret these plots.



**FIGURE 16.6**

R output for the change in double stout sales bootstrap, for Exercise 16.5.

```

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = stout, statistic = theta, R = 3000)

Bootstrap Statistics :
      original      bias    std.error
t1*      74.66667  -0.2038889  14.90047

```

**FIGURE 16.7**

R output for the change in double stout sales bootstrap, for Exercise 16.6.

## 16.6 Read the output.

Figure 16.7 gives output from R for the sample of regions in Exercise 16.3. Summarize the results of the analysis using this output.

## 16.7 What's wrong?

Explain what is wrong with each of the following statements.

- (a) The standard deviation of the bootstrap distribution will be approximately the same as the standard deviation of the original sample.
- (b) The bootstrap distribution is created by resampling without replacement from the original sample.
- (c) When generating the resamples, it is best to use a sample size smaller than the size of the original sample.
- (d) The bootstrap distribution is created by resampling with replacement from the population.

*Inspecting the bootstrap distribution of a statistic helps us judge whether the sampling distribution of the statistic is close to Normal. Bootstrap the sample mean  $\bar{x}$  for each of the data sets in Exercises 16.8 to 16.12 using 2000 resamples. Construct a histogram and a Normal quantile plot to assess Normality of the bootstrap distribution. On the basis of your work, do you expect the sampling distribution of  $\bar{x}$  to be close to Normal? Save your bootstrap results for later analysis.*

## 16.8 Bootstrap distribution of average IQ score.

The distribution of the 60 IQ test scores in Table 1.1 (page 16) is roughly Normal (see Figure 1.9) and the sample size is large enough that we expect a Normal sampling distribution. 

## 16.9 Bootstrap distribution of StubHub! prices.

We examined the distribution of the 186 tickets for the National Collegiate Athletic Association (NCAA) Women's Final Four Basketball Championship in New Orleans posted for sale on StubHub! on January 2, 2013, in Example 1.48 (page 71). The distribution is clearly not Normal; it has three peaks possibly corresponding to three types of seats. We view these data as coming from a process that gives seat prices for an event such as this.  STUBHUB

## 16.10 Bootstrap distribution of time spent watching videos on a cell phone.

The hours per month spent watching videos on cell phones in a random sample of eight cell phone subscribers (Example 7.1, page 421) are

11.9 2.8 3.0 6.2 4.7 9.8 11.1 7.8

The distribution has no outliers, but we cannot assess Normality from such a small sample.  VIDEO

## 16.11 Bootstrap distribution of *Titanic* passenger ages.

In Example 1.36 (page 54) we examined the distribution of the ages of the passengers on the *Titanic*. There is a single mode around 25, a short left tail, and a long right tail. We view these data as coming from a process that would generate similar data.  TITANIC

## 16.12 Bootstrap distribution of average audio file length.

The lengths (in seconds) of audio files found on an iPod (Table 7.3, page 437) are skewed. We previously transformed the data prior to using  $t$  procedures.  SONGS

## 16.13 Standard error versus the bootstrap standard error.

We have two ways to estimate the standard deviation of a sample mean  $\bar{x}$ : use the formula  $s/n$  for the standard error, or use the bootstrap standard error.

- Find the sample standard deviation  $s$  for the 60 IQ test scores in Exercise 16.8 and use it to find the standard error  $s/n$  of the sample mean. How closely does your result agree with the bootstrap standard error from your resampling in Exercise 16.8?
- Find the sample standard deviation  $s$  for the StubHub! ticket price data in Exercise 16.9 and use it to find the standard error  $s/n$  of the sample mean. How closely does your result agree with the bootstrap standard error from your resampling in Exercise 16.9?
- Find the sample standard deviation  $s$  for the eight video-watching times in Exercise 16.10 and use it to find the standard error  $s/n$  of the sample mean. How closely does your result agree with the bootstrap standard error from your resampling in Exercise 16.10?

## 16.14 Service center call lengths.

Table 1.2 (page 19) gives the service center call lengths for a sample of 80 calls. See Example 1.15 (page

18) for more details about these data.  CALLS80

(a) Make a histogram of the call lengths. The distribution is strongly skewed.

(b) The central limit theorem says that the sampling distribution of the sample mean  $\bar{x}$  becomes Normal as the sample size increases. Is the sampling distribution roughly Normal for  $n = 80$ ? To find out, bootstrap these data using 1000 resamples and inspect the bootstrap distribution of the mean. The central part of the distribution is close to Normal. In what way do the tails depart from Normality?

### 16.15 More on service center call lengths.

Here is an SRS of 10 of the service center call lengths from Exercise 16.14:  CALLS10

104 102 35 211 56 325 67 9 179 59

We expect the sampling distribution of  $\bar{x}$  to be less close to Normal for samples of size 10 than for samples of size 80 from a skewed distribution.

(a) Create and inspect the bootstrap distribution of the sample mean for these data using 1000 resamples. Compared with your distribution from the previous exercise, is this distribution closer to or farther away from Normal?

(b) Compare the bootstrap standard errors for your two sets of resamples. Why is the standard error larger for the smaller SRS?

## 16.2 First Steps in Using the Bootstrap

When you complete this section, you will be able to

- Determine when it is appropriate to use the bootstrap standard error and the  $t$  distribution to find a confidence interval.
- Use the bootstrap standard error and the  $t$  distribution to find a confidence interval.

To introduce the key ideas of resampling and bootstrap distributions, we studied an example in which we knew quite a bit about the actual sampling distribution. We saw that the bootstrap distribution agrees with the sampling distribution in *shape* and *spread*.

The *center* of the bootstrap distribution is not the same as the center of the sampling distribution. The sampling distribution of a statistic used to estimate a parameter is centered at the actual value of the parameter in the population, plus any bias. The bootstrap distribution is centered at the value of the statistic for the original sample, plus any bias. The key fact is that the two biases are similar even though the two centers may not be.



**bias, p. 179**

The bootstrap method is most useful in settings where we don't know the sampling distribution of the statistic. The principles are

- **Shape:** Because the shape of the bootstrap distribution approximates the shape of the sampling distribution, we can use the bootstrap distribution to check Normality of the sampling distribution.
- **Center:** A statistic is biased as an estimate of the parameter if its sampling distribution is not centered at the true value of the parameter. We can check bias by seeing whether the bootstrap distribution of the statistic is centered at the value of the statistic for the original sample.

More precisely, the bias of a statistic is the difference between the mean of its sampling distribution and the true value of the parameter. The **bootstrap estimate of bias** is the difference between the mean of the bootstrap distribution and the value of the statistic in the original sample.

bootstrap estimate of bias

- **Spread:** The bootstrap standard error of a statistic is the standard deviation of its bootstrap distribution. The bootstrap standard error estimates the standard deviation of the sampling distribution of the statistic.

## Bootstrap $t$ confidence intervals

If the bootstrap distribution of a statistic shows a Normal shape and small bias, we can get a confidence interval for the parameter by using the bootstrap standard error and the familiar  $t$  distribution. An example will show how this works.

### EXAMPLE

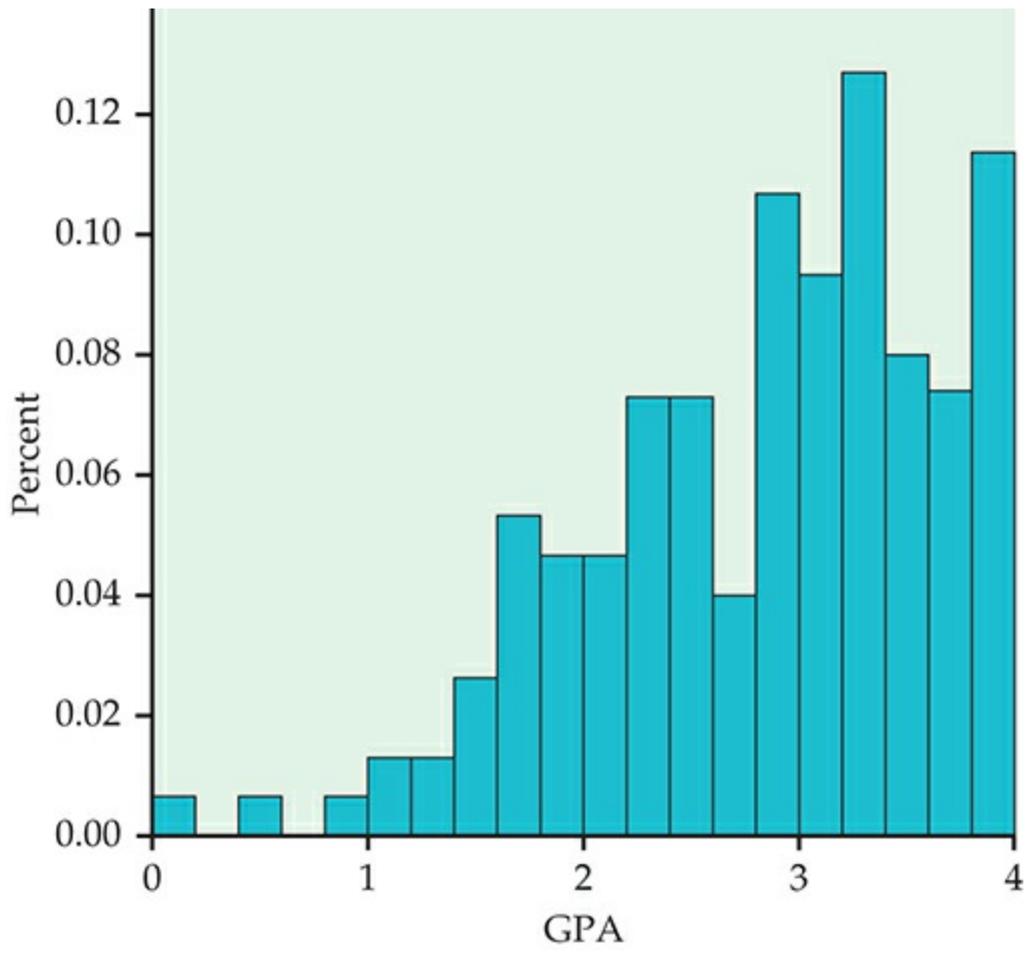
#### 16.4 Grade point averages.

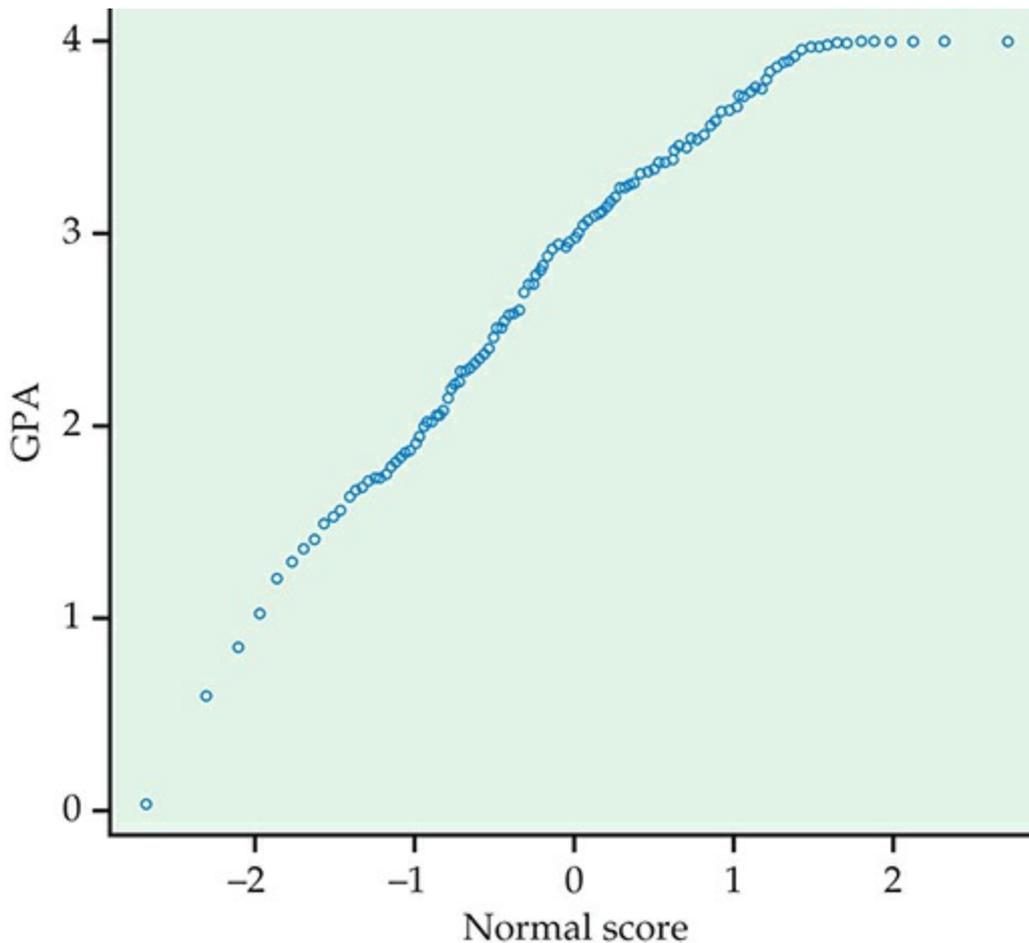


GPA

A study of college students at a large university looked at grade point average (GPA) after three semesters of college as a measure of success. In Example 11.1 (page 612) we examined predictors of GPA. Let's take a look at the distribution of the GPA for the 150 students in this study.

A histogram is given in Figure 16.8(a). The Normal quantile plot is given in Figure 16.8(b). The distribution is strongly skewed to the left. The Normal quantile plot suggests that there are several students with perfect (4.0) GPAs and one at the lower end of the distribution (0.0). These data are not Normally distributed.





**FIGURE 16.8**

Histogram and Normal quantile plot for 150 grade point averages, for Example 16.4. The distribution is strongly skewed.



### trimmed mean, p. 53

The first step is to abandon the mean as a measure of center in favor of a statistic that focuses on the central part of the distribution. We might choose the median, but in this case we will use the 25% trimmed mean, the mean of the middle 50% of the observations. The median is the middle observation or the mean of the two middle observations. The trimmed mean often does a better job of representing the average of typical observations than does the median.

Our *parameter* is the 25% trimmed mean of the population of college student GPAs after three semesters at this large university. By the plug-in principle, the *statistic* that estimates this parameter is the 25% trimmed mean of the sample of 150 students. Because 25% of 150 is 37.5, we drop the 37 lowest and 37 highest GPAs and find the mean of the remaining 76 GPAs. The statistic is

$$\bar{x}^{-25\%}=2.950$$

Given the relatively large sample size from this strongly skewed distribution,

we can use the central limit theorem to argue that the sampling distribution would be approximately Normal with mean near 2.950. Estimating its standard deviation, however, is a more difficult task. We can't simply use the standard error of the sample mean based on the remaining 76 observations, as that will underestimate the true variability.

Fortunately, we don't need any distribution facts to use the bootstrap. We bootstrap the 25% trimmed mean just as we bootstrapped the sample mean: draw 3000 resamples of size 150 from the 150 GPAs, calculate the 25% trimmed mean for each resample, and form the bootstrap distribution from these 3000 values.

Figure 16.9 shows the bootstrap distribution of the 25% trimmed mean. Here is the summary output from R:

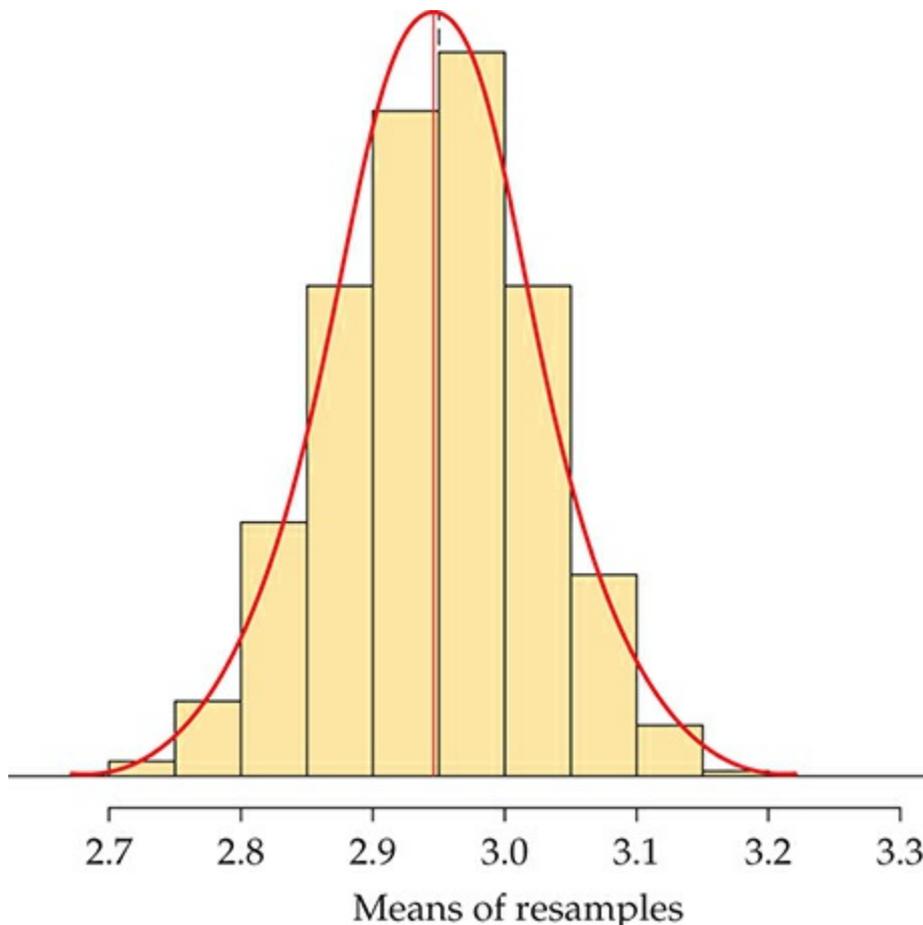
#### ORDINARY NONPARAMETRIC BOOTSTRAP

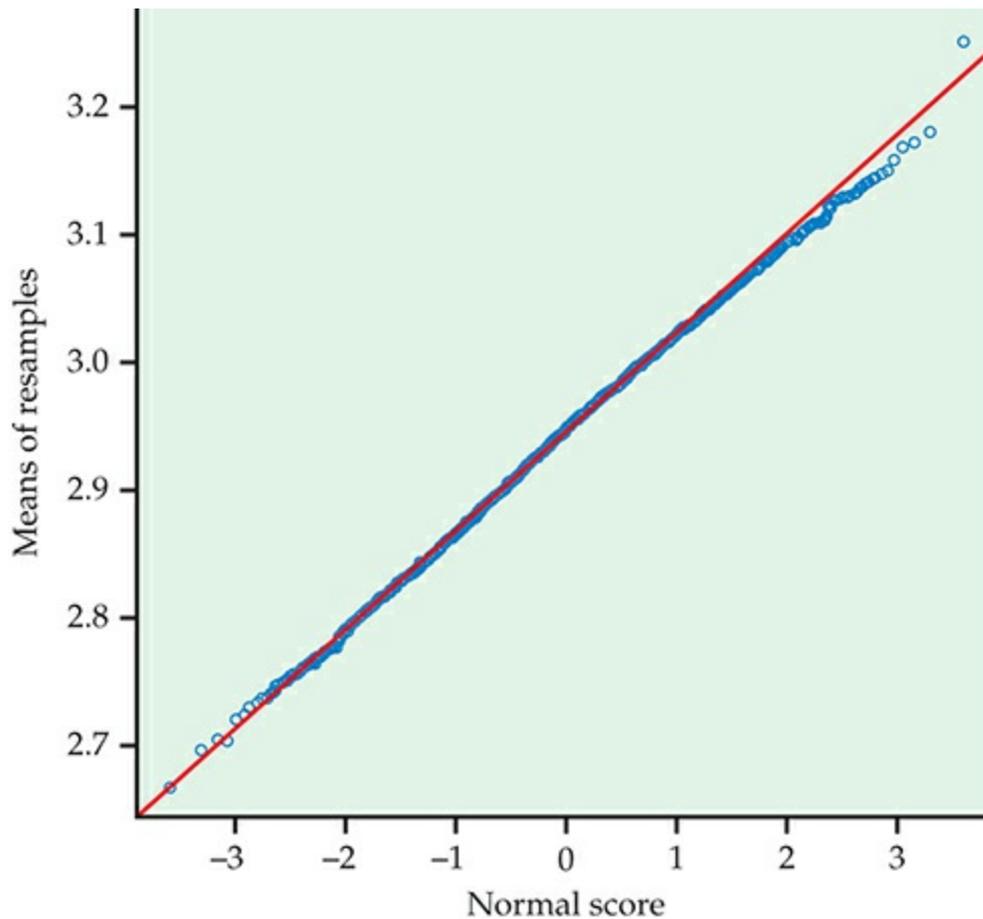
Call:

```
boot(data = GPA, statistic = theta, R = 3000)
```

Bootstrap Statistics :

	original	bias	std. error
t1*	2.949605	-0.002912	0.0778597





**FIGURE 16.9**

The bootstrap distribution of the 25% trimmed means for 3000 resamples from the GPA data in Example 16.4. The bootstrap distribution is approximately Normal.

What do we see?

**Shape:** The bootstrap distribution is close to Normal. This suggests that the sampling distribution of the trimmed mean is also close to Normal.

**Center:** The bootstrap estimate of bias is  $-0.003$ , which is small relative to the value  $2.950$  of the statistic. So the statistic (the trimmed mean of the sample) has small bias as an estimate of the parameter (the trimmed mean of the population).

**Spread:** The bootstrap standard error of the statistic is

$$SE_{\text{boot}} = 0.078$$

This is an estimate of the standard deviation of the sampling distribution of the trimmed mean.

Recall the familiar one-sample  $t$  confidence interval (page 421) for the mean of a Normal population:

$$\bar{x} \pm t^* SE = \bar{x} \pm t^* s/n$$

This interval is based on the Normal sampling distribution of the sample mean  $\bar{x}$  and the formula  $SE = s/n$  for the standard error of  $\bar{x}$ . When a bootstrap distribution is approximately Normal and has small bias, we can essentially use the same idea

with the bootstrap standard error to get a confidence interval for any parameter.

## BOOTSTRAP $t$ CONFIDENCE INTERVAL

Suppose that the bootstrap distribution of a statistic from an SRS of size  $n$  is approximately Normal and that the bootstrap estimate of bias is small. An approximate **level  $C$  confidence interval** for the parameter that corresponds to this statistic by the plug-in principle is

$$\text{statistic} \pm t^* SE_{\text{boot}}$$

where  $SE_{\text{boot}}$  is the bootstrap standard error for this statistic and  $t^*$  is the critical value of the  $t(n - 1)$  distribution with area  $C$  between  $-t^*$  and  $t^*$ .

## EXAMPLE

### 16.5 Bootstrap distribution of the trimmed mean.



GPA

We want to estimate the 25% trimmed mean of the population of all college student GPAs after three semesters at this large university. We have an SRS of size  $n = 150$ . The software output above shows that the trimmed mean of this sample is  $\bar{x}_{25\%} = 2.950$  and that the bootstrap standard error of this statistic is  $SE_{\text{boot}} = 0.078$ . A 95% confidence interval for the population trimmed mean is therefore

$$\begin{aligned}\bar{x}_{25\%} \pm t^* SE_{\text{boot}} &= 2.950 \pm (2.000)(0.078) \\ &= 2.950 \pm 0.156 \\ &= (2.794, 3.106)\end{aligned}$$

Because Table D does not have entries for  $[n - 2(37)] - 1 = 75$  degrees of freedom, we used  $t^* = 2.000$ , the entry for 60 degrees of freedom.

We are 95% confident that the 25% trimmed mean (the mean of the middle

50%) for the population of college student GPAs after three semesters at this large university is between 2.794 and 3.106.

## USE YOUR KNOWLEDGE

### 16.16 Bootstrap $t$ confidence interval.

Recall Example 16.2 (page 16-4). Suppose that a bootstrap distribution was created using 3000 resamples and that the mean and standard deviation of the resample means were 23.29 and 3.90, respectively.

- What is the bootstrap estimate of the bias?
- What is the bootstrap standard error of  $\bar{x}$ ?
- Assume that the bootstrap distribution is reasonably Normal. Since the bias is small relative to the observed  $\bar{x}$ , the bootstrap  $t$  confidence interval for the population mean  $\mu$  is justified. Give the 95% bootstrap  $t$  confidence interval for  $\mu$ .

### 16.17 Bootstrap $t$ confidence interval for average audio file length.



Return to or create the bootstrap distribution resamples on the sample mean for audio file length in Exercise 16.12 (page 16-12). In Example 7.10 (page 437), the  $t$  confidence interval was applied to the logarithm of the time measurements.

- Inspect the bootstrap distribution. Is a bootstrap  $t$  confidence interval appropriate? Explain why or why not.
- Construct the 95% bootstrap  $t$  confidence interval.
- Compare the bootstrap results with the  $t$  confidence interval reported in Example 7.11 (page 438).

## Bootstrapping to compare two groups



two-sample  $t$  significance test, p. 454

Two-sample problems are among the most common statistical settings. In a two-sample problem, we wish to compare two populations, such as male and female college students, based on separate samples from each population. When both populations are roughly Normal, the two-sample  $t$  procedures compare the two population means. The bootstrap can also compare two populations, without the Normality condition and without the restriction to comparison of means. The most important new idea is that bootstrap resampling must mimic the “separate samples” design that produced the original data.

## BOOTSTRAP FOR COMPARING TWO POPULATIONS

Given independent SRSs of sizes  $n$  and  $m$  from two populations:

1. Draw a resample of size  $n$  with replacement from the first sample and a separate resample of size  $m$  from the second sample. Compute a statistic that compares the two groups, such as the difference between the two sample means.
2. Repeat this resampling process thousands of times.
3. Construct the bootstrap distribution of the statistic. Inspect its shape, bias, and bootstrap standard error in the usual way.

## EXAMPLE

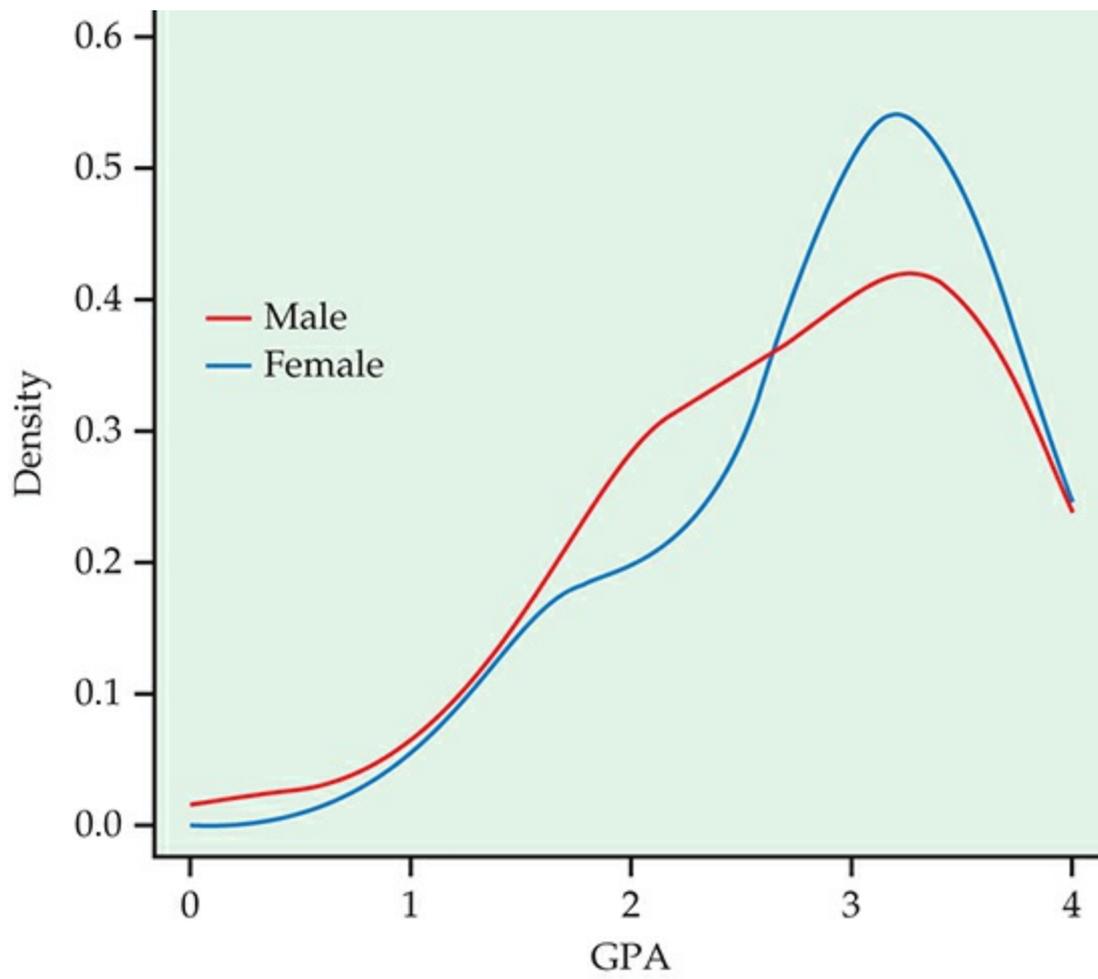
### 16.6 Bootstrap comparison of GPAs.

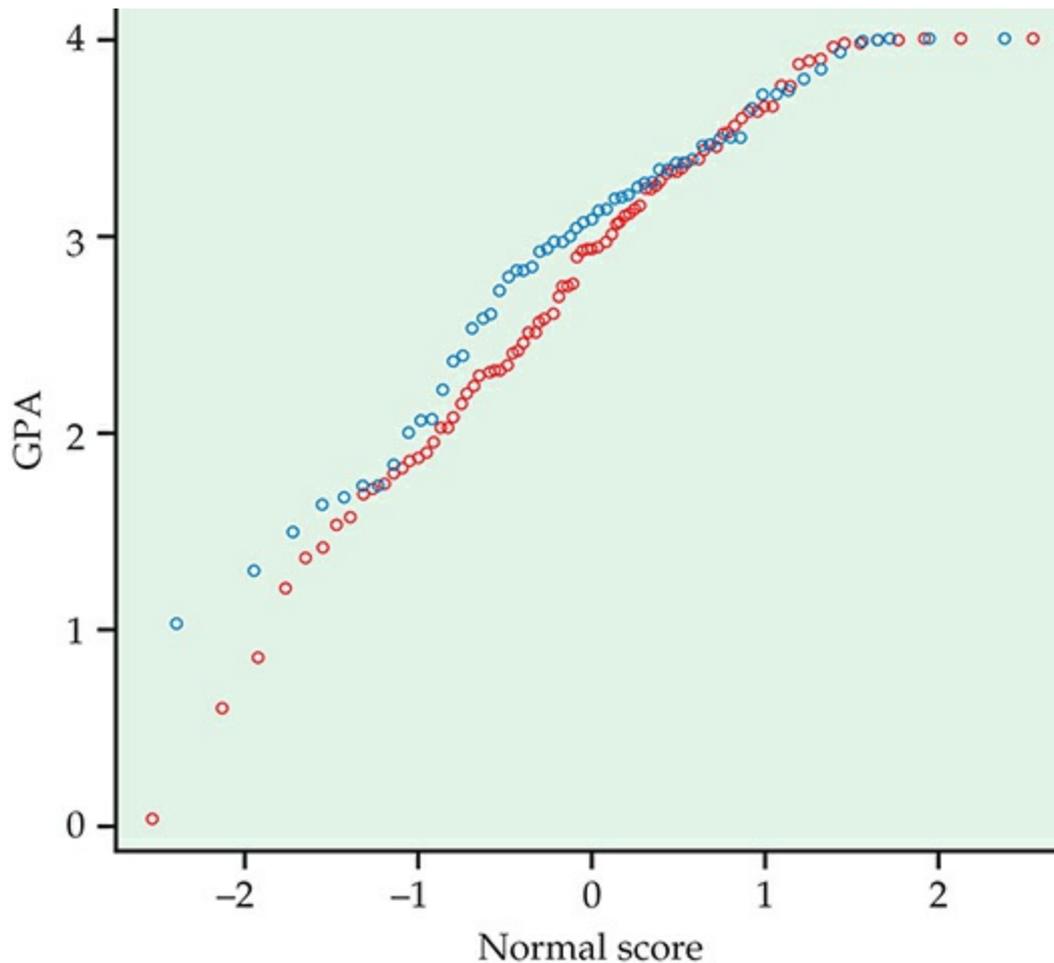
In Example 16.4 we looked at grade point average (GPA) after three semesters of college as a measure of success. How do GPAs compare between men and women? Figure 16.10 shows density curves and Normal quantile plots for the GPAs of 91 males and 59 females. The distributions are both far from Normal. Here are some summary statistics:

Gender	$n$	$\bar{x}$	$s$
Male	91	2.784	0.859
Female	59	2.933	0.748
Difference		-0.149	

The data suggest that GPAs tend to be slightly higher for females. The mean

GPA for females is roughly 0.15 higher than the mean for males.





**FIGURE 16.10**

Density curves and Normal quantile plots of the distributions of GPA for males and females, for Example 16.6.

In the setting of Example 16.6 we want to estimate the difference between population means,  $\mu_1 - \mu_2$ . We might be somewhat reluctant to use the two-sample  $t$  confidence interval because both samples are very skewed. To compute the bootstrap standard error for the difference in sample means  $\bar{x}_1 - \bar{x}_2$ , resample separately from the two samples. Each of our 3000 resamples consists of two group resamples, one of size 91 drawn with replacement from the male data and one of size 59 drawn with replacement from the female data.

For each combined resample, compute the statistic  $\bar{x}_1 - \bar{x}_2$ . The 3000 differences form the bootstrap distribution. The bootstrap standard error is the standard deviation of the bootstrap distribution.

The boot function in R automates this bootstrap procedure. Here is the R output:

## STRATIFIED BOOTSTRAP

Call:

```
boot(data = gpa, statistic = meanDiff, R = 3000,  
      strata = sex)
```

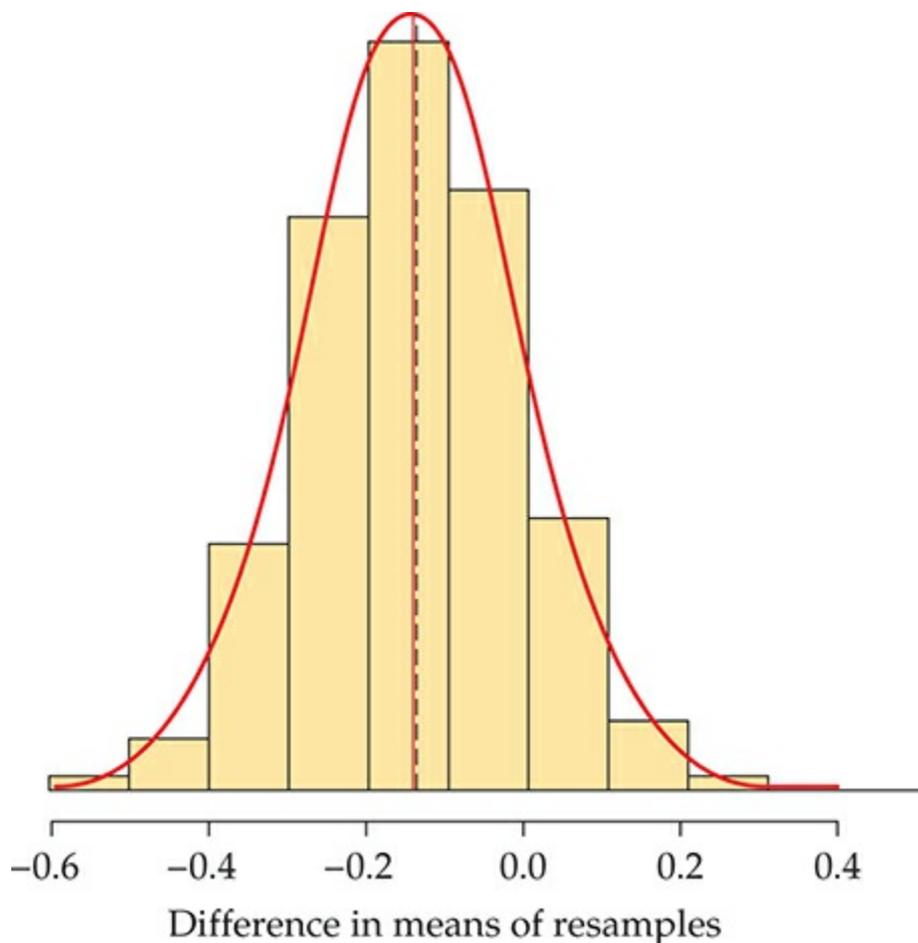
Bootstrap Statistics :

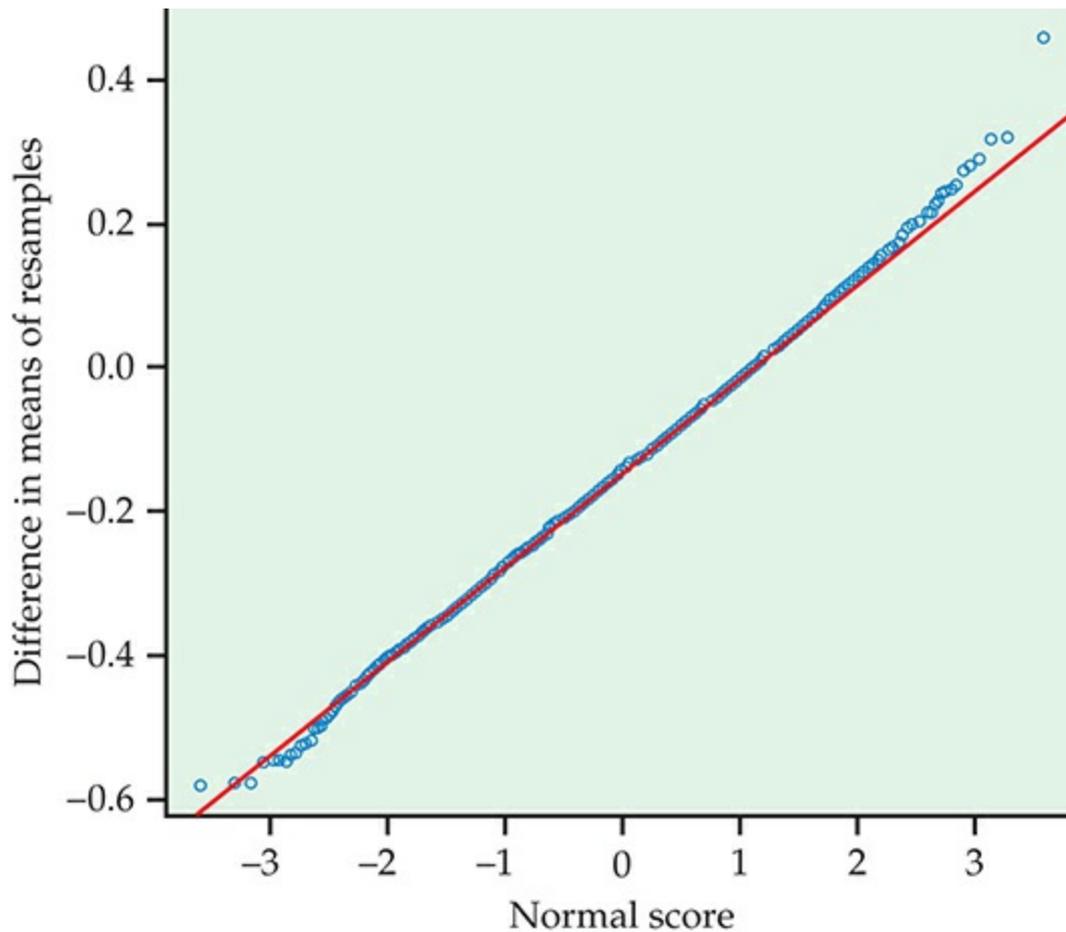
	original	bias	std. error
t1*	-0.1490259	0.003989901	0.1327419

Figure 16.11 shows that the bootstrap distribution is close to Normal. We can trust the bootstrap  $t$  confidence interval for these data. A 95% confidence interval for the difference in mean GPAs (males versus females) is therefore

$$\begin{aligned} \bar{x} \pm 2.5\% \pm t^* SE_{boot} &= -0.149 \pm (2.009)(0.133) \\ &= -0.149 \pm 0.267 \\ &= (-0.416, 0.118) \end{aligned}$$

Because Table D does not have entries for  $(n_1 - 1, n_2 - 1) = 58$  degrees of freedom, we used  $t^* = 2.009$ , the entry for 50 degrees of freedom.





**FIGURE 16.11**

The bootstrap distribution and Normal quantile plot for the differences in means for the GPA data.

We are 95% confident that the difference in the mean GPAs of males and females at this large university after three semesters is between  $-0.416$  and  $0.118$ . Because  $0$  is in this interval, we cannot conclude that the two population means are different. We will discuss hypothesis testing in Section 16.5.



In this example, the bootstrap distribution of the difference is close to Normal. *When the bootstrap distribution is non-Normal, we can't trust the bootstrap  $t$  confidence interval.* Fortunately, there are more general ways of using the bootstrap to get confidence intervals that can be safely applied when the bootstrap distribution is not Normal. These methods, which we discuss in Section 16.4, are the next step in practical use of the bootstrap.

## USE YOUR KNOWLEDGE

### 16.18 Bootstrap comparison of average reading abilities.



Table 7.4 (page 452) gives the scores on a test of reading ability for two groups of third-grade students. The treatment group used “directed reading activities” and the control group followed the same curriculum without the activities.

- Bootstrap the difference in means  $\bar{x}_1 - \bar{x}_2$  and report the bootstrap standard error.
- Inspect the bootstrap distribution. Is a bootstrap  $t$  confidence interval appropriate? If so, give a 95% confidence interval.
- Compare the bootstrap results with the two-sample  $t$  confidence interval reported in Example 7.14 on page 453.

### 16.19 Formula-based versus bootstrap standard error.



We have a formula (page 451) for the standard error of  $\bar{x}_1 - \bar{x}_2$ . This formula does not depend on Normality. How does this formula-based standard error for the data of Example 16.6 compare with the bootstrap standard error?

## BEYOND THE BASICS

### The Bootstrap for a Scatterplot Smoother

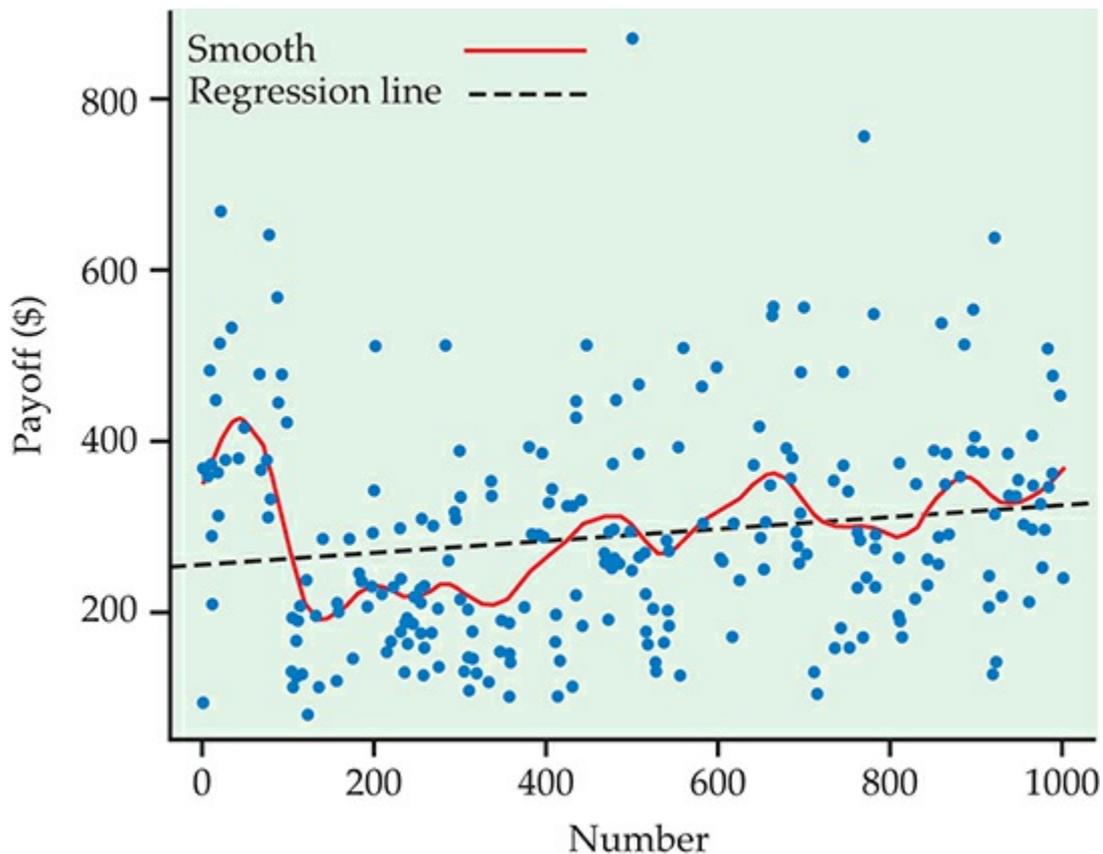
The bootstrap idea can be applied to quite complicated statistical methods, such as the scatterplot smoother illustrated in Chapter 2 (page 96).

## EXAMPLE

### 16.7 Do all daily numbers have an equal payoff?

The New Jersey Pick-It Lottery is a daily numbers game run by the state of New Jersey. We'll analyze the first 254 drawings after the lottery was started in 1975.<sup>4</sup> Buying a ticket entitles a player to pick a number between 000 and 999. Half the money bet each day goes into the prize pool. (The state takes the other half.) The state picks a winning number at random, and the prize pool is shared equally among all winning tickets.

Although all numbers are equally likely to win, numbers chosen by fewer people have bigger payoffs if they win because the prize is shared among fewer tickets. Figure 16.12 is a scatterplot of the first 254 winning numbers and their payoffs. What patterns can we see?

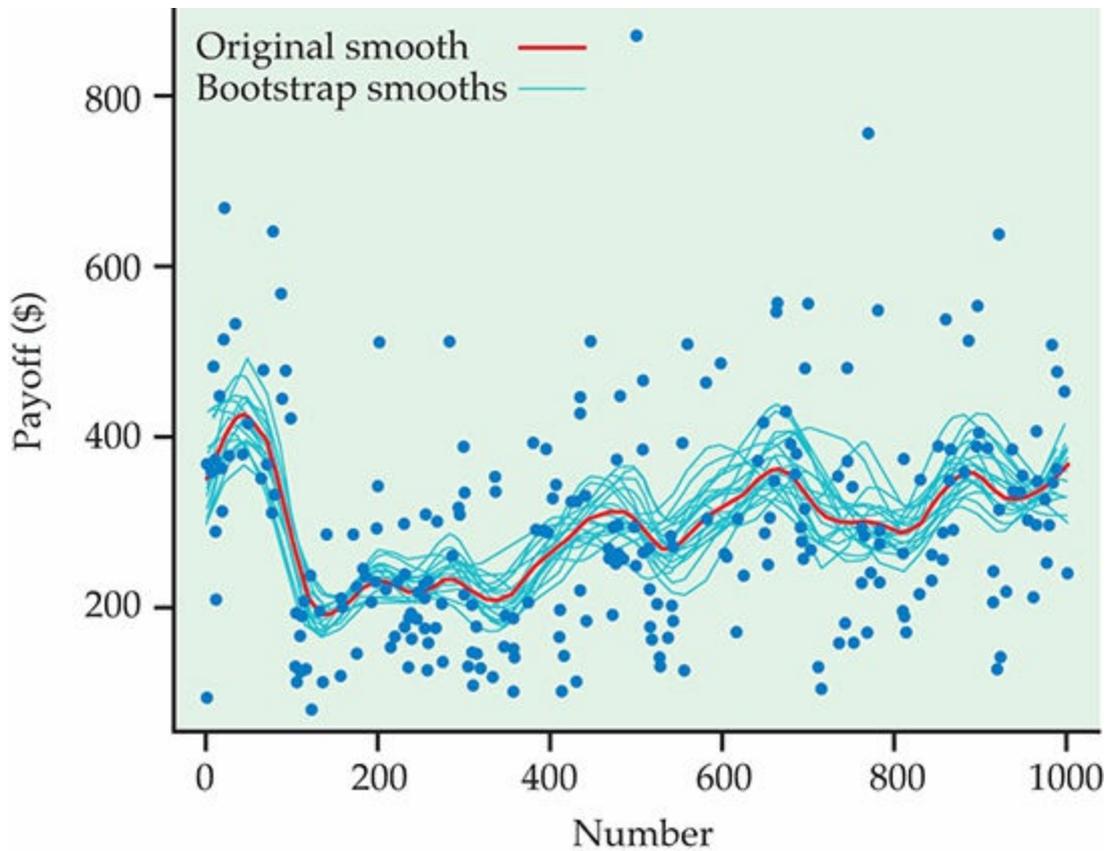


**FIGURE 16.12**

The first 254 winning numbers in the New Jersey Pick-It Lottery and the payoffs for each, for Example 16.7. To see patterns we use least-squares regression (dashed line) and a scatterplot smoother (curve).

The straight line in Figure 16.12 is the least-squares regression line. The line shows a general trend of higher payoffs for larger winning numbers. The curve in

the figure was fitted to the plot by a scatterplot smoother that follows local patterns in the data rather than being constrained to a straight line. The curve suggests that there were larger payoffs for numbers in the intervals 000 to 100, 400 to 500, 600 to 700, and 800 to 999.



**FIGURE 16.13**

The curves produced by the scatterplot smoother for 20 resamples from the data displayed in Figure 16.12. The curve for the original sample is the heavy line.

Are the patterns displayed by the scatterplot smoother just chance? We can use the bootstrap distribution of the smoother’s curve to get an idea of how much random variability there is in the curve. Each resample “statistic” is now a curve rather than a single number. Figure 16.13 shows the curves that result from applying the smoother to 20 resamples from the 254 data points in Figure 16.12. The original curve is the thick line. The spread of the resample curves about the original curve shows the sampling variability of the output of the scatterplot smoother.

Nearly all the bootstrap curves mimic the general pattern of the original smoother curve, showing, for example, the same low average payoffs for numbers in the 200s and 300s. This suggests that these patterns are real, not just chance. In fact, when people pick “random” numbers, they tend to choose numbers starting with 2, 3, 5, or 7, so these numbers have lower payoffs. This pattern disappeared after 1976; it appears that players noticed the pattern and changed their number choices.

## SECTION 16.2 Summary

Bootstrap distributions mimic the shape, spread, and bias of sampling distributions.

The **bootstrap standard error**  $SE_{\text{boot}}$  of a statistic is the standard deviation of its bootstrap distribution. It measures how much the statistic varies under random sampling.

The bootstrap estimate of the **bias** of a statistic is the mean of the bootstrap distribution minus the statistic for the original data. Small bias means that the bootstrap distribution is centered at the statistic of the original sample and suggests that the sampling distribution of the statistic is centered at the population parameter.

The bootstrap can estimate the sampling distribution, bias, and standard error of a wide variety of statistics, such as the **trimmed mean**, whether or not statistical theory tells us about their sampling distributions.

If the bootstrap distribution is approximately Normal and the bias is small, we can give a **bootstrap  $t$  confidence interval**,  $\text{statistic} \pm t^* SE_{\text{boot}}$ , for the parameter. Do not use this  $t$  interval if the bootstrap distribution is not Normal or shows substantial bias.

To use the bootstrap **to compare two populations**, draw separate resamples from each sample and compute a statistic that compares the two groups. Repeat many times and use the bootstrap distribution for inference.

## SECTION 16.2 Exercises

*For Exercises 16.16 and 16.17, see page 16-17; and for Exercises 16.18 and 16.19, see page 16-20.*

### 16.20 Should you use the bootstrap standard error and the $t$ distribution for the confidence interval?

For each of the following situations, explain whether or not you would use the bootstrap standard error and the  $t$  distribution for the confidence interval. Give reasons for your answers.

- (a) The bootstrap distribution of the mean is approximately Normal, and the difference between the mean of the data and the mean of the bootstrap distribution is large relative to the mean of the data.
- (b) The bootstrap distribution of the mean is approximately Normal, and the difference between the mean of the data and the mean of the bootstrap distribution is small relative to the mean of the data.
- (c) The bootstrap distribution of the mean is clearly skewed, and the difference between the mean of the data and the mean of the bootstrap distribution is large relative to the mean of the data.
- (d) The bootstrap distribution of the mean is clearly skewed, and the difference between the mean of the data and the mean of the bootstrap distribution is small relative to the mean of the data.

### 16.21 Use the bootstrap standard error and the $t$ distribution for the confidence interval.

The observed mean is 112.3, the mean of the bootstrap distribution is 109.8, the standard error is 9.4, and  $n$

= 51. Use the  $t$  distribution to find the 95% confidence interval.

### 16.22 Bootstrap $t$ confidence interval for the StubHub! prices.

In Exercise 16.9 (page 16-12) we examined the bootstrap for the prices of tickets to the NCAA Women's Final Four Basketball Championship in New Orleans.  **STUBHUB**

- (a) Find the bootstrap  $t$  95% confidence interval for these data.
- (b) Compare the interval you found in part (a) with the usual  $t$  interval.
- (c) Which interval do you prefer? Give reasons for your answer.

### 16.23 Bootstrap $t$ confidence interval for the ages of the *Titanic* passengers.

In Exercise 16.11 (page 16-12) we examined the bootstrap for the ages of the *Titanic* passengers.  **TITANIC**

- (a) Find the bootstrap  $t$  95% confidence interval for these data.
- (b) Compare the interval you found in part (a) with the usual  $t$  interval.
- (c) Which interval do you prefer? Give reasons for your answer.

### 16.24 Bootstrap $t$ confidence interval for time spent watching videos on a cell phone.

Return to or re-create the bootstrap distribution of the sample mean for the eight times spent watching videos in Exercise 16.10 (page 16-12).

- (a) Although the sample is small, verify using graphs and numerical summaries of the bootstrap distribution that the distribution is reasonably Normal and that the bias is small relative to the observed  $\bar{x}$ .

 **VIDEO**

- (b) The bootstrap  $t$  confidence interval for the population mean  $\mu$  is therefore justified. Give the 95% bootstrap  $t$  confidence interval for  $\mu$ .
- (c) Give the usual  $t$  95% interval and compare it with your interval from part (b).

### 16.25 Bootstrap $t$ confidence interval for service center call lengths.

Return to or re-create the bootstrap distribution of the sample mean for the 80 service center call lengths in Exercise 16.14 (page 16-12).  **CALLS80**

- (a) What is the bootstrap estimate of the bias? Verify from the graphs of the bootstrap distribution that the distribution is reasonably Normal (some right-skew remains) and that the bias is small relative to the observed  $\bar{x}$ . The bootstrap  $t$  confidence interval for the population mean  $\mu$  is therefore justified.

- (b) Give the 95% bootstrap  $t$  confidence interval for  $\mu$ .

- (c) The only difference between the bootstrap  $t$  and usual one-sample  $t$  confidence intervals is that the bootstrap interval uses  $SE_{boot}$  in place of the formula-based standard error  $s/n$ . What are the values of the

two standard errors? Give the usual  $t$  95% interval and compare it with your interval from part (b).

### 16.26 Another bootstrap distribution of the trimmed mean.

Bootstrap distributions and quantities based on them differ randomly when we repeat the resampling process. A key fact is that they do not differ very much if we use a large number of resamples. Figure 16.9 (page 16-15) shows one bootstrap distribution of the trimmed mean of the GPA data. Repeat the resampling of these data to get another bootstrap distribution of the trimmed mean.  GPA

- (a) Plot the bootstrap distribution and compare it with Figure 16.9. Are the two bootstrap distributions similar?
- (b) What are the values of the bias and bootstrap standard error for your new bootstrap distribution? How do they compare with the previous values given on page 16-15?
- (c) Find the 95% bootstrap  $t$  confidence interval based on your bootstrap distribution. Compare it with the previous result in Example 16.5 (page 16-16).

### 16.27 Bootstrap distribution of the standard deviation $s$ .

For Example 16.5 (page 16-16) we bootstrapped the 25% trimmed mean of 150 GPAs. Another statistic whose sampling distribution is unfamiliar to us is the standard deviation  $s$ . Bootstrap  $s$  for these data. Discuss the shape and bias of the bootstrap distribution. Is the bootstrap  $t$  confidence interval for the population standard deviation  $\sigma$  justified? If it is, give a 95% confidence interval.  GPA

### 16.28 Bootstrap comparison of tree diameters.

In Exercise 7.85 (page 471) you were asked to compare the mean diameter at breast height (DBH) for trees from the northern and southern halves of a land tract using a random sample of 30 trees from each region.  NSPINES

- (a) Use a back-to-back stemplot or side-by-side boxplots to examine the data graphically. Does it appear reasonable to use standard  $t$  procedures?
- (b) Bootstrap the difference in means  $\bar{x}^{\text{North}} - \bar{x}^{\text{South}}$  and look at the bootstrap distribution. Does it meet the conditions for a bootstrap  $t$  confidence interval?
- (c) Report the bootstrap standard error and the 95% bootstrap  $t$  confidence interval.
- (d) Compare the bootstrap results with the usual two-sample  $t$  confidence interval.

### 16.29 Bootstrapping a Normal data set.

The following data are “really Normal.” They are an SRS from the standard Normal distribution  $N(0, 1)$ , produced by a software Normal random number generator.  NORMALD

0.01	-0.04	-1.02	-0.13	-0.36	-0.03	-1.88	0.34	-0.00	1.21
-0.02	-1.01	0.58	0.92	-1.38	-0.47	-0.80	0.90	-1.16	0.11
0.23	2.40	0.08	-0.03	0.75	2.29	-1.11	-2.23	1.23	1.56
-0.52	0.42	-0.31	0.56	2.69	1.09	0.10	-0.92	-0.07	-1.76
0.30	-0.53	1.47	0.45	0.41	0.54	0.08	0.32	-1.35	-2.42

0.34	0.51	2.47	2.99	-1.56	1.27	1.55	0.80	-0.59	0.89
-2.36	1.27	-1.11	0.56	-1.12	0.25	0.29	0.99	0.10	0.30
0.05	1.44	-2.46	0.91	0.51	0.48	0.02	-0.54		

- (a) Make a histogram and Normal quantile plot. Do the data appear to be “really Normal”? From the histogram, does the  $N(0, 1)$  distribution appear to describe the data well? Why?
- (b) Bootstrap the mean. Why do your bootstrap results suggest that  $t$  confidence intervals are appropriate?
- (c) Give both the bootstrap and the formula-based standard errors for  $\bar{x}$ . Give both the bootstrap and usual  $t$  95% confidence intervals for the population mean  $\mu$ .

### 16.30 Bootstrap distribution of the median.

We will see in Section 16.3 that bootstrap methods often work poorly for the median. To illustrate this, bootstrap the sample median of the 50 times to start a business that we studied in Example 16.1 (page 16-

3). Why is the bootstrap  $t$  confidence interval not justified?  TIME50

### 16.31 Bootstrap distribution of the mpg standard deviation.

Computers in some vehicles calculate various quantities related to performance. One of these is the fuel efficiency, or gas mileage, usually expressed as miles per gallon (mpg). For one vehicle equipped in this way, the mpg were recorded each time the gas tank was filled, and the computer was then reset. We studied these data in Exercise 7.30 (page 443) using methods based on Normal distributions.<sup>5</sup> Here are the mpg values for a random sample of 20 of these records:  MPG20

41.5	50.7	36.6	37.3	34.2	45.0	48.0	43.2	47.7	42.2
43.2	44.6	48.4	46.4	46.8	39.2	37.3	43.5	44.3	43.3

In addition to the average mpg, the driver is also interested in how much variability there is in the mpg.

- (a) Calculate the sample standard deviation  $s$  for these mpg values.
- (b) We have no formula for the standard error of  $s$ . Find the bootstrap standard error for  $s$ .
- (c) What does the standard error indicate about how accurate the sample standard deviation is as an estimate of the population standard deviation?
- (d) Would it be appropriate to give a bootstrap  $t$  interval for the population standard deviation? Why or why not?

## 16.3 How Accurate Is a Bootstrap Distribution?

When you complete this section, you will be able to

- Describe the effect of the size of the original sample on the variation in bootstrap distributions.
- Describe the effect of the number of resamples on the variation in bootstrap distributions.

We said earlier that “When can I safely bootstrap?” is a somewhat subtle issue. Now we will give some insight into this issue.

We understand that a statistic will vary from sample to sample and that inference about the population must take this random variation into account. The sampling distribution of a statistic displays the variation in the statistic due to selecting samples at random from the population. For example, the margin of error in a confidence interval expresses the uncertainty due to sampling variation. In this chapter we have used the bootstrap distribution as a substitute for the sampling distribution. This introduces a second source of random variation: choosing resamples at random from the original sample.

### SOURCES OF VARIATION IN A BOOTSTRAP DISTRIBUTION

Bootstrap distributions and conclusions based on them include two sources of random variation:

1. Choosing an original sample at random from the population.
2. Choosing bootstrap resamples at random from the original sample.

A statistic in a given setting has only one sampling distribution. It has many bootstrap distributions, formed by the two-step process just described. Bootstrap inference generates one bootstrap distribution and uses it to tell us about the sampling distribution. Can we trust such inference?

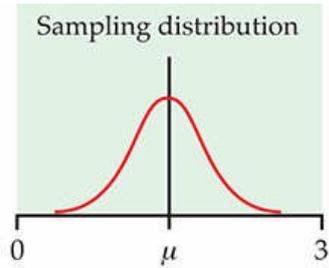
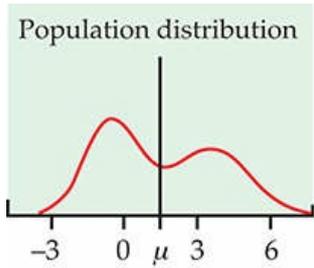
Figure 16.14 displays an example of the entire process. The population distribution (top left) has two peaks and is far from Normal. The histograms in the left column of the figure show five random samples from this population, each of size 50. The line in each histogram marks the mean  $\bar{x}$  of that sample. These vary from sample to sample. The distribution of the  $\bar{x}$ -values from all possible samples

is the sampling distribution. This sampling distribution appears to the right of the population distribution. It is close to Normal, as we expect because of the central limit theorem.

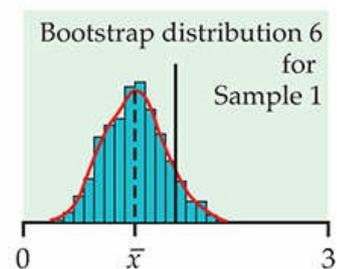
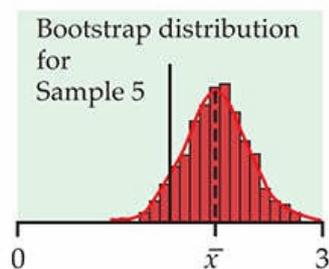
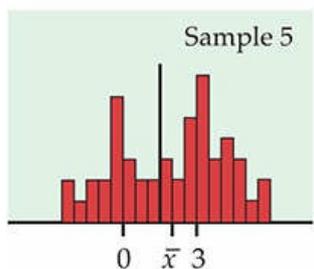
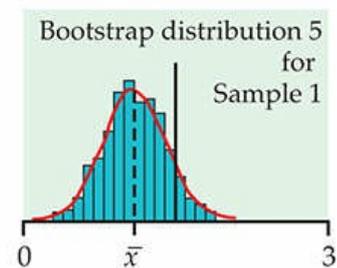
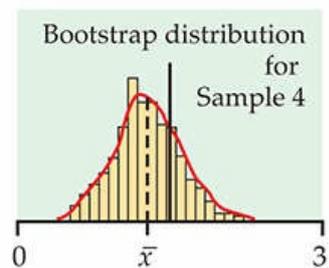
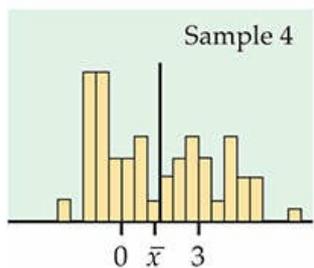
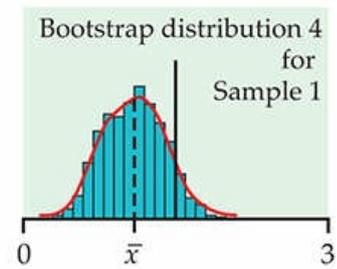
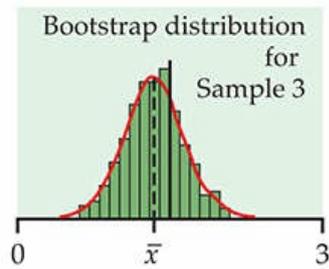
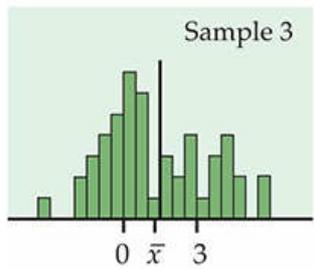
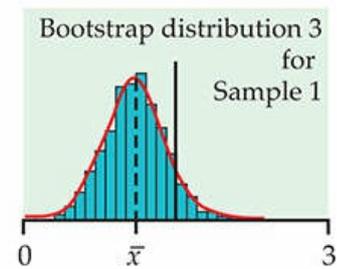
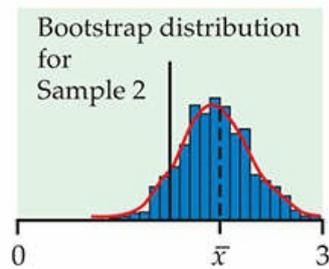
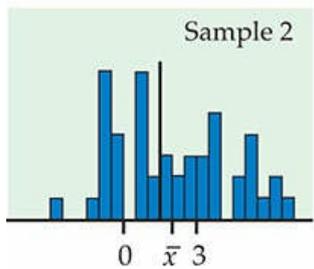
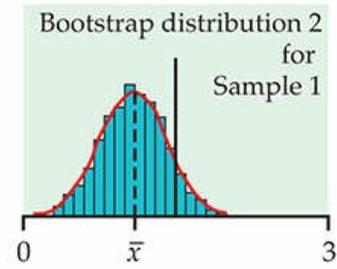
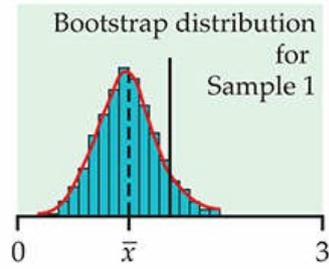
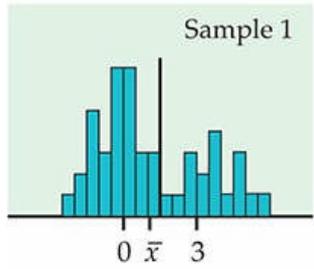
The middle column in Figure 16.14 displays the bootstrap distribution of  $\bar{x}$  for each of the five samples. Each distribution was created by drawing 1000 resamples from the original sample, calculating  $\bar{x}$  for each resample, and presenting the 1000  $\bar{x}$ 's in a histogram. The right column shows the bootstrap distribution of the first sample, repeating the resampling five more times.

Compare the five bootstrap distributions in the middle column to see the effect of the random choice of the original sample. Compare the six bootstrap distributions drawn from the first sample to see the effect of the random resampling. Here's what we see:

- Each bootstrap distribution is centered close to the value of  $\bar{x}$  for its original sample. That is, the bootstrap estimate of bias is small in all five cases. Of course, the five  $\bar{x}$ -values vary, and not all are close to the population mean  $\mu$ .



Population mean =  $\mu$   
 Sample mean =  $\bar{x}$



### FIGURE 16.14

Five random samples of  $n = 50$  from the same population, with a bootstrap distribution of the sample mean formed by resampling from each of the five samples. At the right are five more bootstrap distributions from the first sample.

- The shape and spread of the bootstrap distributions in the middle column vary a bit, but all five resemble the sampling distribution in shape and spread. That is, the shape and spread of a bootstrap distribution depend on the original sample, but the variation from sample to sample is not great.
- The six bootstrap distributions from the same sample are very similar in shape, center, and spread. That is, *random resampling adds very little variation to the variation due to the random choice of the original sample from the population.*

Figure 16.14 reinforces facts that we have already relied on. If a bootstrap distribution is based on a moderately large sample from the population, its shape and spread don't depend heavily on the original sample and do mimic the shape and spread of the sampling distribution. Bootstrap distributions do not have the same center as the sampling distribution; they mimic bias, not the actual center.

The figure also illustrates a fact that is important for practical use of the bootstrap: the bootstrap resampling process (using 1000 or more resamples) introduces very little additional variation. We can rely on a bootstrap distribution to inform us about the shape, bias, and spread of the sampling distribution.

### Bootstrapping small samples

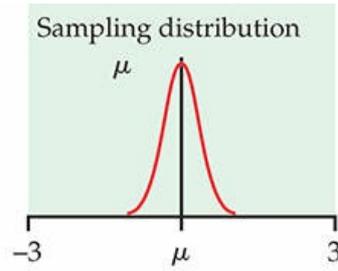
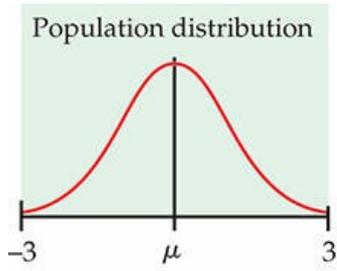
We now know that almost all the variation in bootstrap distributions for a statistic such as the mean comes from the random selection of the original sample from the population. We also know that in general statisticians prefer large samples because small samples give more variable results. This general fact is also true for bootstrap procedures.

Figure 16.15 repeats Figure 16.14, with two important differences. The five original samples are only of size  $n = 9$ , rather than the  $n = 50$  of Figure 16.14. Also, the population distribution (top left) is Normal, so that the sampling distribution of  $\bar{x}$  is Normal despite the small sample size.

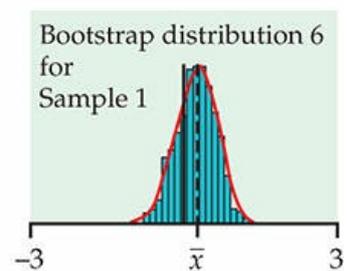
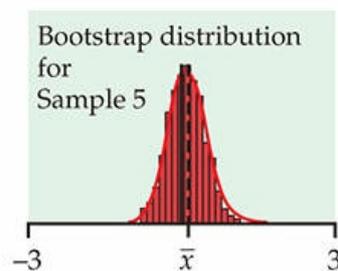
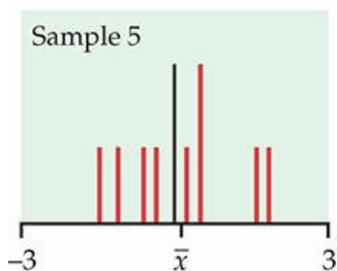
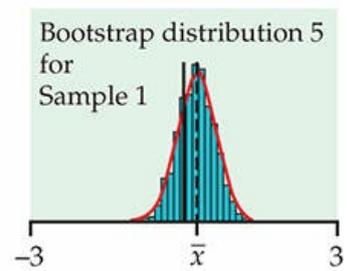
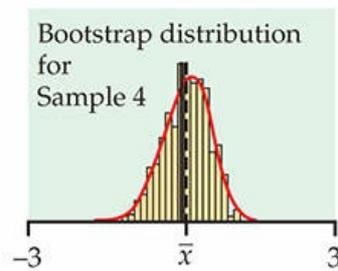
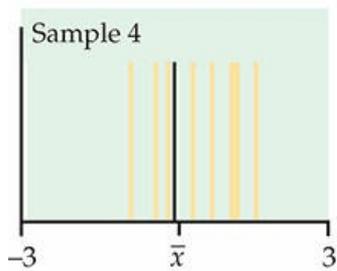
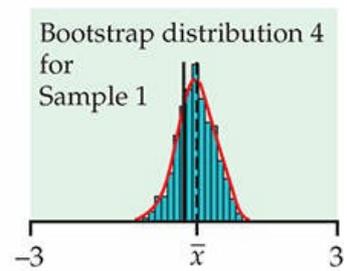
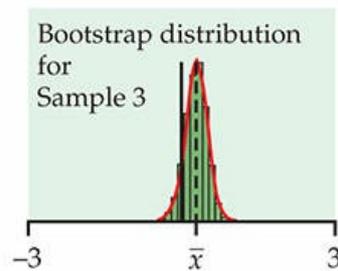
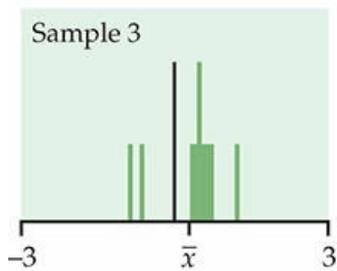
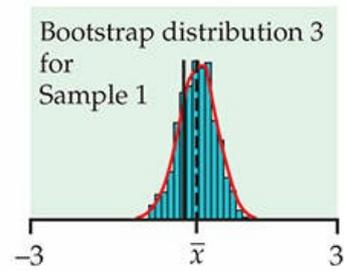
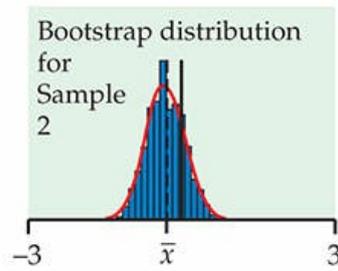
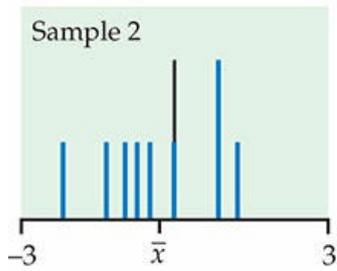
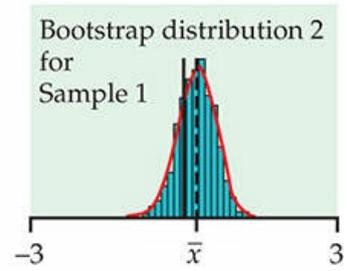
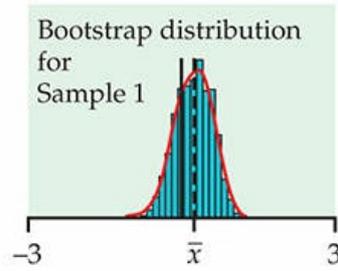
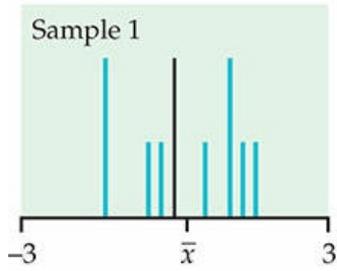
Even with a Normal population distribution, the bootstrap distributions in the middle column show much more variation in shape and spread than those for larger samples in Figure 16.14. Notice, for example, how the skewness of the fourth sample produces a skewed bootstrap distribution. The bootstrap distributions are no longer all similar to the sampling distribution at the top of the column.



*We can't trust a bootstrap distribution from a very small sample to closely mimic the shape and spread of the sampling distribution.* Bootstrap confidence intervals will sometimes be too long or too short, or too long in one direction and too short in the other. The six bootstrap distributions based on the first sample are again very similar. Because we used 1000 resamples, resampling adds very little variation. There are subtle effects that can't be seen from a few pictures, but the main conclusions are clear.



Population mean =  $\mu$   
Sample mean =  $\bar{x}$



## FIGURE 16.15

Five random samples of  $n = 9$  from the same population, with a bootstrap distribution of the sample mean formed by resampling from each of the five samples. At the right are five more bootstrap distributions from the first sample.

### VARIATION IN BOOTSTRAP DISTRIBUTIONS

For most statistics, almost all the variation in bootstrap distributions comes from the selection of the original sample from the population. You can reduce this variation by using a larger original sample.

Bootstrapping does not overcome the weakness of small samples as a basis for inference. We will describe some bootstrap procedures that are usually more accurate than standard methods, but even they may not be accurate for very small samples. Use caution in any inference—including bootstrap inference—from a small sample.

The bootstrap resampling process using 1000 or more resamples introduces very little additional variation.

### Bootstrapping a sample median

In dealing with the grade point averages in Example 16.5, we chose to bootstrap the 25% trimmed mean rather than the median. We did this in part because the usual bootstrapping procedure doesn't work well for the median unless the original sample is quite large. Now we will bootstrap the median in order to understand the difficulties.

Figure 16.16 follows the format of Figures 16.14 and 16.15. The population distribution appears at top left, with the population median  $M$  marked. Below in the left column are five samples of size  $n = 15$  from this population, with their sample medians  $m$  marked. Bootstrap distributions of the median based on resampling from each of the five samples appear in the middle column. The right column again displays five more bootstrap distributions from re-sampling the first sample. The six bootstrap distributions from the same sample are once again very similar to each other—resampling adds little variation—so we concentrate on the middle column in the figure.

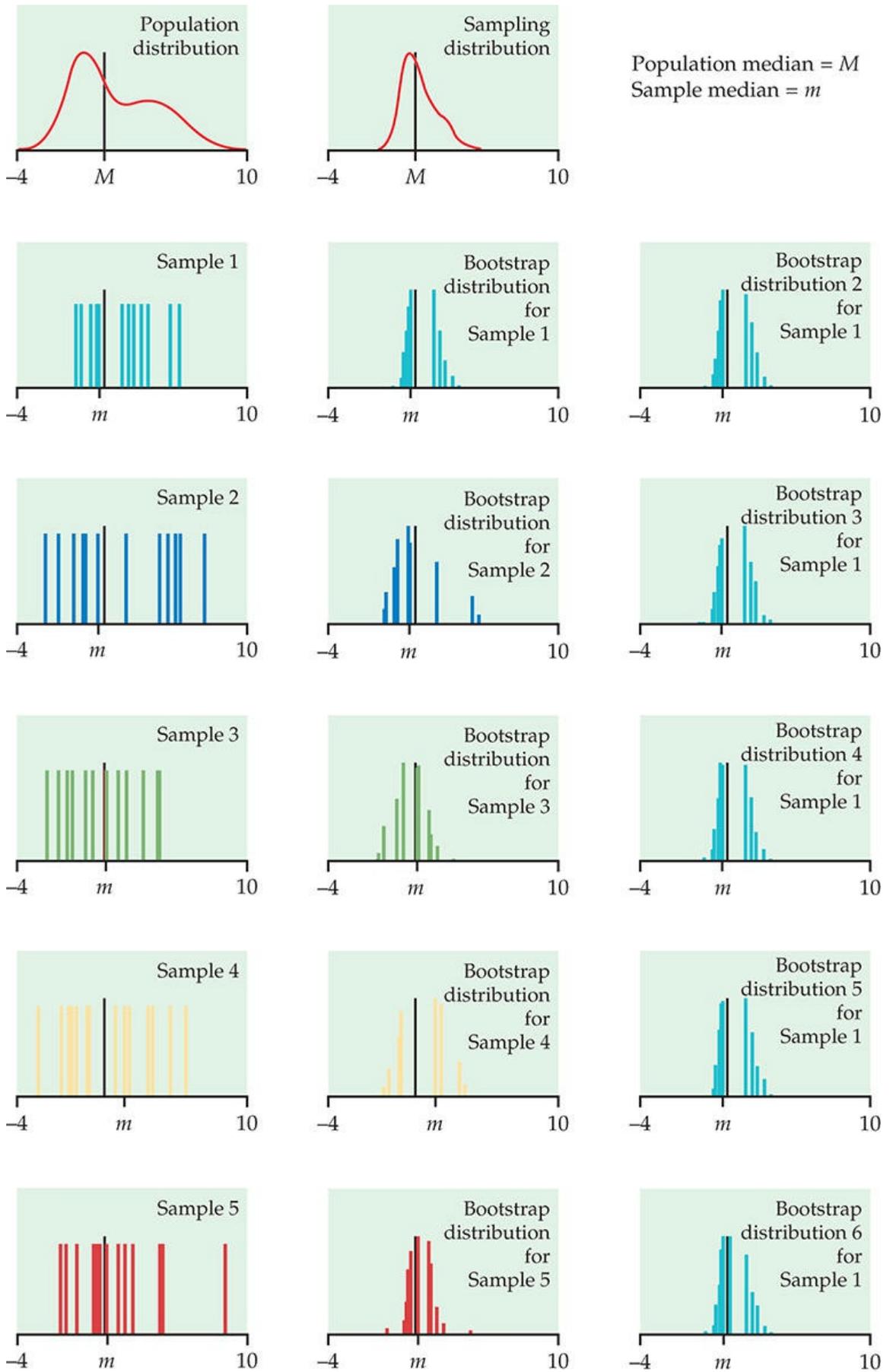
Bootstrap distributions from the five samples differ markedly from each other and from the sampling distribution at the top of the column. Here's why. The median of a resample of size 15 is the eighth-largest observation in the resample. This is always one of the 15 observations in the original sample and is usually one of the middle observations. Each bootstrap distribution repeats the same few values, and these values depend on the original sample. The sampling distribution, on the other hand, contains the medians of all possible samples and is not confined to a few values.

The difficulty is somewhat less when  $n$  is even, because the median is then the

average of two observations. It is much less for moderately large samples, say  $n = 100$  or more. Bootstrap standard errors and confidence intervals from such samples are reasonably accurate, though the shapes of the bootstrap distributions may still appear odd. You can see that the same difficulty will occur for small samples with other statistics, such as the quartiles, that are calculated from just one or two observations from a sample.



There are more advanced variations of the bootstrap idea that improve performance for small samples and for statistics such as the median and quartiles. *Unless you have expert advice or undertake further study, avoid bootstrapping the median and quartiles unless your sample is rather large.*



### FIGURE 16.16

Five random samples of  $n = 15$  from the same population, with a bootstrap distribution of the sample median formed by resampling from each of the five samples. At the right are five more bootstrap distributions from the first sample.

## SECTION 16.3 Summary

Almost all the variation in a bootstrap distribution for a statistic is due to the selection of the original random sample from the population. Resampling introduces little additional variation.

Bootstrap distributions based on small samples can be quite variable. Their shape and spread reflect the characteristics of the sample and may not accurately estimate the shape and spread of the sampling distribution. Bootstrap inference from a small sample may therefore be unreliable.

Bootstrap inference based on samples of moderate size is unreliable for statistics like the median and quartiles that are calculated from just a few of the sample observations.

## SECTION 16.3 Exercises

### 16.32 Variation in the bootstrap distributions.

Consider the variation in the bootstrap for each of the following situations with two scenarios, S1 and S2. In comparing the variation, do you expect, in general, that S1 will have less variation than S2, that S2 will have less variation than S1, or that the variation for S1 and S2 will be approximately the same? Give reasons for your answers. Here, we use  $n$  for the size of the original sample and  $B$  for the number of resamples.

- (a) S1:  $n = 50, B = 2000$ ; S2:  $n = 50, B = 4000$ .
- (b) S1:  $n = 10, B = 2000$ ; S2:  $n = 50, B = 2000$ .
- (c) S1:  $n = 50, B = 200$ ; S2:  $n = 50, B = 2000$ .
- (d) S1:  $n = 10, B = 2000$ ; S2:  $n = 50, B = 4000$ .

### 16.33 Bootstrap versus sampling distribution.

Most statistical software includes a function to generate samples from Normal distributions. Set the mean to 26 and the standard deviation to 27. You can think of all the numbers that would be produced by this function if it ran forever as a population that has the  $N(26, 27)$  distribution. Samples produced by the function are samples from this population.

- (a) What is the exact sampling distribution of the sample mean  $\bar{x}$  for a sample of size  $n$  from this population?
- (b) Draw an SRS of size  $n = 10$  from this population. Bootstrap the sample mean  $\bar{x}$  using 2000 resamples from your sample. Give a histogram of the bootstrap distribution and the bootstrap standard error.
- (c) Repeat the same process for samples of sizes  $n = 40$  and  $n = 160$ .

(d) Write a careful description comparing the three bootstrap distributions and also comparing them with the exact sampling distribution. What are the effects of increasing the sample size?

### 16.34 The effect of increasing the sample size.

The data for Example 16.1 (page 16-3) are the times to start a business for a random sample of 50 countries. The entire survey included 185 countries. The distribution of times is very non-Normal. A histogram with a smooth density curve is given in Figure 1.19(a) (page 54). However, for this histogram we excluded one country, Suriname, where it takes 694 days to start a business. Exclude Suriname from the data set and use the remaining data for the remaining 184 countries.  **TIME184**

(a) Let's think of the 184 countries as the population for this exercise. Find the mean  $\mu$  and the standard deviation  $\sigma$  for this population.

(b) Although we don't know the shape of the sampling distribution of the sample mean  $\bar{x}$  for a sample of size  $n$  from this population, we do know the mean and standard deviation of this distribution. What are they?

(c) Draw an SRS of size  $n = 10$  from this population. Bootstrap the sample mean  $\bar{x}$  using 2000 resamples from your sample. Give a histogram of the bootstrap distribution and the bootstrap standard error.

(d) Repeat the same process for samples of sizes  $n = 40$  and  $n = 160$ .

(e) Write a careful description comparing the three bootstrap distributions. What are the effects of increasing the sample size?

### 16.35 The effect of non-Normality.

The populations in the two previous exercises have the same mean and standard deviation, but one is Normal and the other is strongly non-Normal. Based on your work in these exercises, how does non-Normality of the population affect the bootstrap distribution of  $\bar{x}$ ? How does it affect the bootstrap standard error? Do either of these effects diminish when we start with a larger sample? Explain what you have observed based on what you know about the sampling distribution of  $\bar{x}$  and the way in which bootstrap distributions mimic the sampling distribution.

## 16.4 Bootstrap Confidence Intervals

When you complete this section, you will be able to

- Use the bootstrap distribution to find a bootstrap percentile confidence interval.
- Read software output to find the BCa confidence interval.

Until now, we have met just one type of inference procedure based on resampling, the bootstrap  $t$  confidence intervals. We can calculate a bootstrap  $t$  confidence interval for any parameter by bootstrapping the corresponding statistic. We don't need conditions on the population or special knowledge about the sampling distribution of the statistic.

The flexible and almost automatic nature of bootstrap  $t$  intervals is appealing—but there is a catch. These intervals work well only when the bootstrap distribution tells us that the sampling distribution is approximately Normal and has small bias. How well must these conditions be met? What can we do if we don't trust the bootstrap  $t$  interval? In this section we will see how to quickly check  $t$  confidence intervals for accuracy, and we will learn alternative bootstrap confidence intervals that can be used more generally than the bootstrap  $t$ .

### Bootstrap percentile confidence intervals

Confidence intervals are based on the sampling distribution of a statistic. If a statistic has no bias as an estimator of a parameter, its sampling distribution is centered at the true value of the parameter. We can then get a 95% confidence interval by marking off the central 95% of the sampling distribution. The  $t$  critical values in a  $t$  confidence interval are a shortcut to marking off the central 95%.

This shortcut doesn't work under all conditions—it depends both on lack of bias and on Normality. One way to check whether  $t$  intervals (using either bootstrap or formula-based standard errors) are reasonable is to compare them with the central 95% of the bootstrap distribution. The 2.5 and 97.5 percentiles mark off the central 95%. The interval between the 2.5 and 97.5 percentiles of the bootstrap distribution is often used as a confidence interval in its own right. It is known as a *bootstrap percentile confidence interval*.

### BOOTSTRAP PERCENTILE CONFIDENCE INTERVALS

The interval between the 2.5 and 97.5 percentiles of the bootstrap distribution

of a statistic is a 95% **bootstrap percentile confidence interval** for the corresponding parameter. Use this method when the bootstrap estimate of bias is small.

The conditions for safe use of bootstrap  $t$  and bootstrap percentile intervals are a bit vague. We recommend that you check whether these intervals are reasonable by comparing them with each other. If the bias of the bootstrap distribution is small and the distribution is close to Normal, the bootstrap  $t$  and percentile confidence intervals will agree closely.



Percentile intervals, unlike  $t$  intervals, do not ignore skewness. Percentile intervals are therefore usually more accurate, as long as the bias is small. Because we will soon meet a much more accurate bootstrap interval, our recommendation is that *when bootstrap and bootstrap percentile intervals do not agree closely, neither type of interval should be used.*

## EXAMPLE

### 16.8 Bootstrap percentile confidence interval for the trimmed mean.

In Example 16.5 (page 16-16) we found that a 95% bootstrap  $t$  confidence interval for the 25% trimmed mean of GPA for the population of college students after three semesters at this large university is between 2.794 and 3.106. The bootstrap distribution in Figure 16.9 shows a small bias and, though closely Normal, is a bit skewed. Is the bootstrap  $t$  confidence interval accurate for these data?

We can use the quantile function in R to compute the needed percentiles of our 3000 resamples. For this bootstrap distribution, the 2.5 and 97.5 percentiles are 2.793 and 3.095, respectively. These are the endpoints of the 95% bootstrap percentile confidence interval. This interval is quite close to the bootstrap  $t$  interval. We conclude that both intervals are reasonably accurate.

The bootstrap  $t$  interval for the trimmed mean of GPA in Example 16.8 is

$$\bar{x}^{-25\% \pm t * SE_{boot}} = 2.950 \pm 0.156$$

We can learn something by also writing the percentile interval starting at the statistic  $\bar{x}_{25\%}=2.950$ . In this form, it is

$$2.950 - 0.157, 2.950 + 0.145$$

Unlike the  $t$  interval, the percentile interval is not symmetric—its endpoints are different distances from the statistic. The slightly greater distance to the 2.5 percentile reflects the slight left-skewness of the bootstrap distribution.

## USE YOUR KNOWLEDGE

### 16.36 Determining the percentile endpoints.

What percentiles of the bootstrap distribution are the endpoints of a 99% bootstrap percentile confidence interval? How do they change for a 90% bootstrap percentile confidence interval?

### 16.37 Bootstrap percentile confidence interval for time to start a business.

Consider the random subset of the time to start a business data in Exercise 16.1 (page 16-3). Bootstrap the sample mean using 2000 resamples.

- Make a histogram and a Normal quantile plot. Does the bootstrap distribution appear close to Normal? Is the bias small relative to the observed sample mean?
- Find the 95% bootstrap  $t$  confidence interval.
- Give the 95% confidence percentile interval and compare it with the interval in part (b).

## A more accurate bootstrap confidence interval: BCa

Any method for obtaining confidence intervals requires some conditions in order to produce exactly the intended confidence level. These conditions (for example, Normality) are never exactly met in practice. So a 95% confidence interval in practice will not capture the true parameter value exactly 95% of the time.

In addition to “hitting” the parameter 95% of the time, a good confidence interval should divide its 5% of “misses” equally between high misses and low misses. We will say that a method for obtaining 95% confidence intervals is **accurate** in a particular setting if 95% of the time it produces intervals that capture the parameter and if the 5% of misses are equally shared between high and low

misses. Perfect accuracy isn't available in practice, but some methods are more accurate than others.

### accurate

One advantage of the bootstrap is that we can to some extent check the accuracy of the bootstrap  $t$  and percentile confidence intervals by examining the bootstrap distribution for bias and skewness and by comparing the two intervals with each other. The interval in Example 16.8 reveals a slight left-skewness, but not enough to invalidate inference.



*In general, the  $t$  and percentile intervals may not be sufficiently accurate when*

- the statistic is strongly biased, as indicated by the bootstrap estimate of bias.
- the sampling distribution of the statistic is clearly skewed, as indicated by the bootstrap distribution and by comparing the  $t$  and percentile intervals.

Most confidence interval procedures are more accurate for larger sample sizes. The  $t$  and percentile procedures improve only slowly: they require 100 times more data to improve accuracy by a factor of 10. (Recall the  $n$  in the formula for the usual one-sample  $t$  interval.) These intervals may not be very accurate except for quite large sample sizes. There are more elaborate bootstrap procedures that improve faster, requiring only 10 times more data to improve accuracy by a factor of 10. These procedures are quite accurate unless the sample size is very small.

## BCa CONFIDENCE INTERVALS

The **bootstrap bias-corrected accelerated (BCa) interval** is a modification of the percentile method that adjusts the percentiles to correct for bias and skewness.

This method is accurate in a wide variety of settings, has reasonable computation requirements (by modern standards), and does not produce excessively wide intervals. The BCa intervals are among the most widely used intervals. Since this interval is related to the percentile method, it is still based on the key ideas of resampling and the bootstrap distribution.

Now that you understand these concepts, you should always use this more accurate method (or an alternative like tilting intervals) if your software offers it. The details of producing confidence intervals are quite technical.<sup>6</sup> The BCa method requires more than 1000 resamples for high accuracy. We recommend that you use

5000 or more resamples. *Don't forget that even BCa confidence intervals should be used cautiously when sample sizes are small, because there are not enough data to accurately determine the necessary corrections for bias and skewness.*



## EXAMPLE

### 16.9 The BCa confidence interval for the ratio of variances.



In Example 16.6 (page 16-18), we compared the GPA means of men and women using a 95% bootstrap  $t$  confidence interval. Because 0 was contained in the interval, we concluded that there was not enough evidence to state that the two means were different. Suppose we also want to compare the variances. Figure 16.10 (page 16-18) suggests that the spread among the male GPAs is larger than that of the females. The ratio of the male sample variance to the female sample variance is 1.321. Can we conclude there is a difference?

In Section 7.3, we discussed an  $F$  test for the equality of spread but also warned that this approach was very sensitive to non-Normal data. Because our GPA data are heavily skewed, we cannot trust this test and instead will use the bootstrap. Specifically, we'll form a 95% confidence interval for  $\sigma_1^2/\sigma_2^2$ .

Figure 16.17 shows the bootstrap distribution of the ratio of sample variances  $s_1^2/s_2^2$ . We see strong skewness in the bootstrap distribution and therefore in the sampling distribution. This is not unexpected. Recall that if the data are Normal and the variances are equal, we'd expect this ratio to follow an  $F$  distribution.

The bootstrap  $t$  and percentile intervals aren't reliable when the sampling distribution of the statistic is skewed. Figure 16.18 shows software output that includes the percentile and BCa intervals. The bootstrap  $t$  interval is closely related to the Normal interval that is also supplied. The basic confidence interval is another method based on the percentiles of the bootstrap distribution that we will not discuss here.

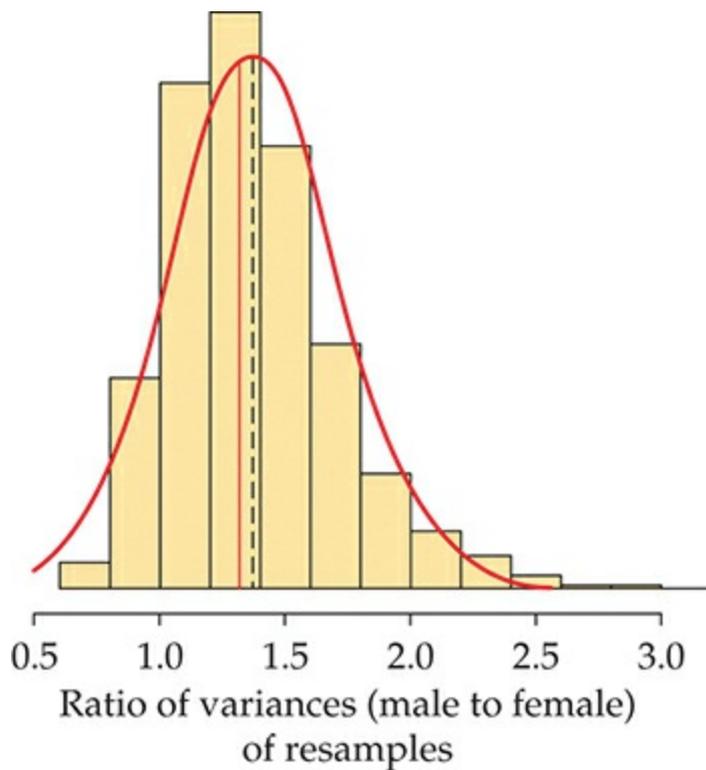
The BCa interval is

$$(1.321 - 0.456, 1.321 + 0.914) = (0.865, 2.235)$$

and the percentile interval is

$$(1.321 - 0.468, 1.321 + 0.880) = (0.853, 2.201)$$

In this case the percentile and BCa intervals are similar, but the BCa is shifted slightly, as it has adjusted for the bias, which was estimated at 0.054. Both intervals are strongly asymmetrical: the upper endpoint is about twice as far from the sample ratio as the lower endpoint. This reflects the strong right-skewness of the bootstrap distribution.



**FIGURE 16.17**

The bootstrap distribution of the ratio of sample variances of 5000 resamples from the data in Example 16.6.

```
R Console

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 5000 bootstrap replicates

CALL :
boot.ci(boot.out = gpa2.boot)

Intervals :
Level      Normal          Basic
95%      (0.608, 1.926)      (0.441, 1.788)

Level      Percentile          BCa
95%      (0.853, 2.201)      (0.865, 2.235)
Calculations and Intervals on Original Scale
```

**FIGURE 16.18**

R output for bootstrapping the ratio of variances for the GPA data.

The output in Figure 16.18 also shows that both endpoints of the less-accurate intervals (bootstrap  $t$  via the Normal interval and the percentile interval) are too low. These intervals miss the population ratio on the low side too often (more than 2.5% of the time) and miss on the high side too seldom. They give a biased picture of where the true ratio is likely to be.

### Confidence intervals for the correlation

The bootstrap allows us to find confidence intervals for a wide variety of statistics. So far, we have looked at the sample mean, trimmed mean, the difference between two means, and the ratio of sample variances using a variety of different bootstrap confidence intervals. The choice of interval depended on the shape of the bootstrap distribution and the desired accuracy.

Now we will bootstrap the correlation coefficient. This is our first use of the bootstrap for a statistic that depends on two related variables. As with the difference between two means, we must pay attention to how we should resample.

---

#### EXAMPLE

##### 16.10 Correlation between price and rating.



## LAUNDRY

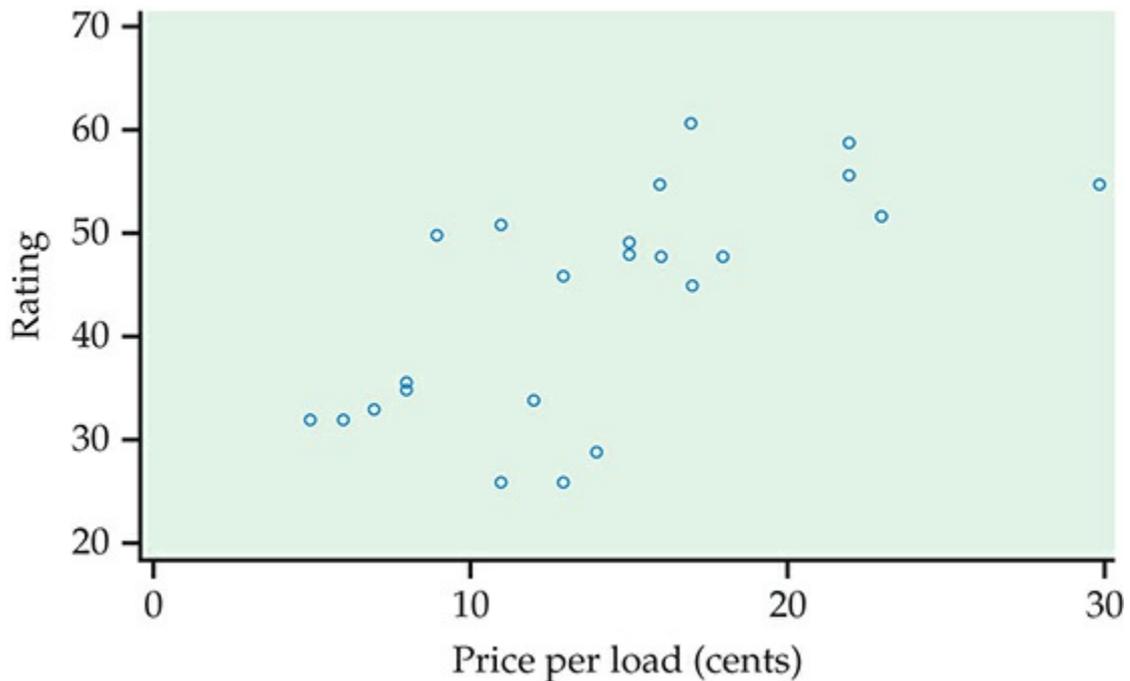
Consumers Union provides ratings on a large variety of consumer products. They use sophisticated testing methods as well as surveys of their members to create these ratings. The ratings are published in their magazine, *Consumer Reports*.

An article in *Consumer Reports* rated laundry detergents on a scale from 1 to 100. Here are the ratings along with the price per load, in cents, for 24 laundry detergents:

Rating	Price(cents)	Rating	Price(cents)	Rating	Price(cents)	Rating	Price(cents)
61	17	59	22	56	22	55	16
55	30	52	23	51	11	50	15
50	9	48	16	48	15	48	18
46	13	46	13	45	17	36	8
35	8	34	12	33	7	32	6
32	5	29	14	26	11	26	13

In Example 2.8 (page 87) we examined the relationship between rating and price per load for these laundry detergents. We expect that the higher-priced detergents will tend to have higher ratings. The scatterplot in Figure 16.19 shows that the higher-priced products do tend to have better ratings, but the relationship is not particularly strong. The correlation is 0.671. Let's use the bootstrap to find a 95% confidence interval for the population correlation.

Our confidence interval will also provide a test of the null hypothesis that the population correlation is zero. If the 95% confidence interval does not include zero, we can reject the null hypothesis in favor of the two-sided alternative. Although we would expect the correlation to be positive, we could be surprised and find that it is negative. It is important to keep in mind that *we cannot use what we learned by looking at the scatterplot to formulate our alternative hypothesis.*



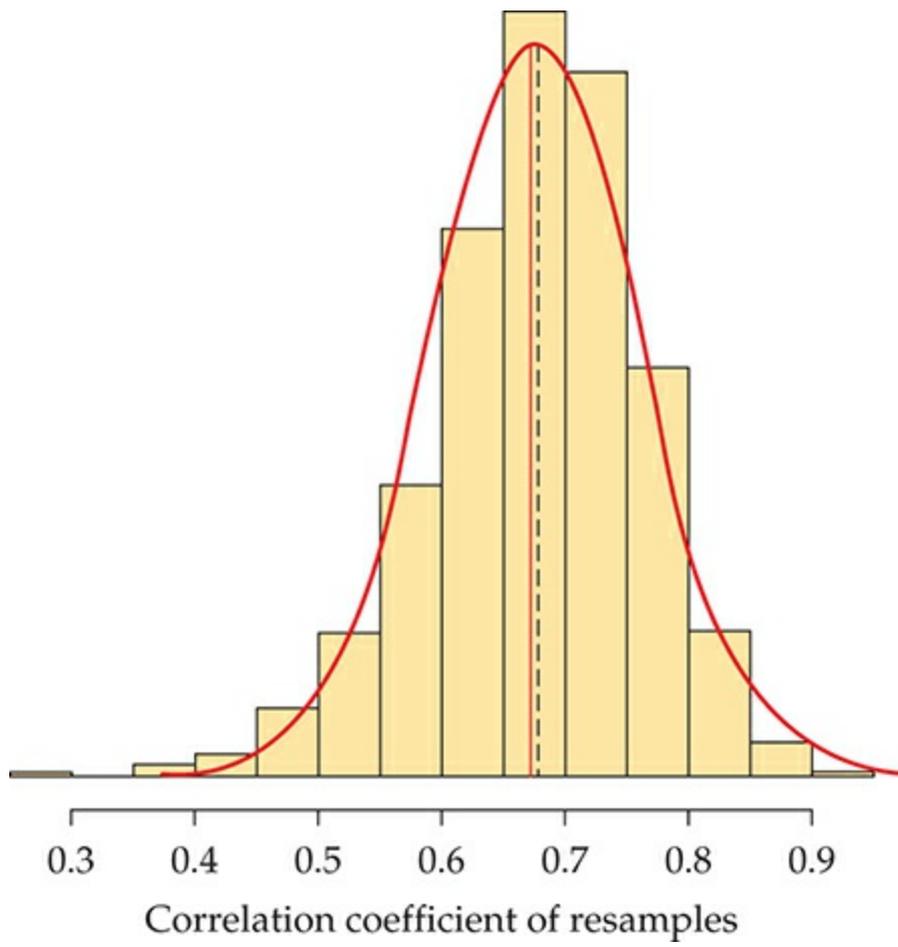
**FIGURE 16.19**

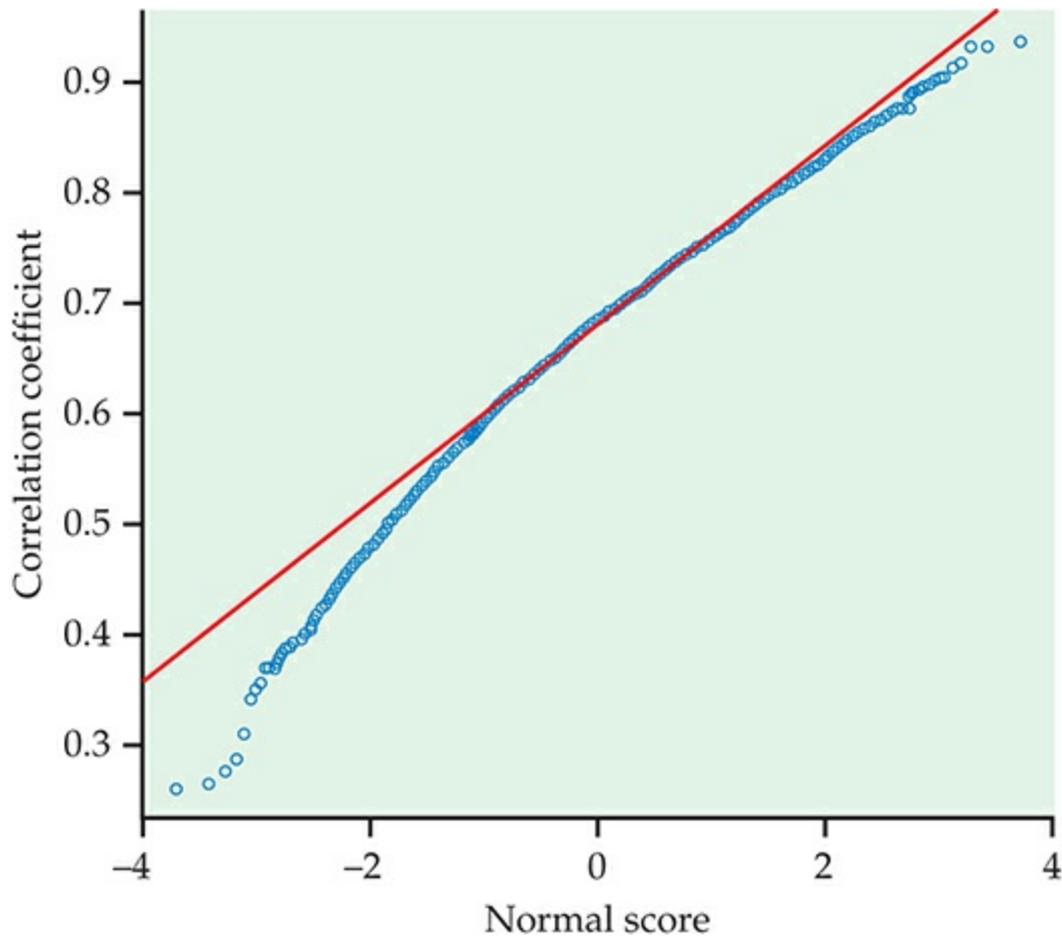
Scatterplot of price per load (in cents) versus rating for 24 laundry detergents, for Example 16.10.



How shall we resample from the laundry detergent data? Because each observation consists of the price and the rating for one product, we resample products. Resampling prices and ratings separately would lose the connection between a product's price and its rating. Software such as R automates proper resampling. Once we have produced a bootstrap distribution by resampling, we can examine the distribution and construct a confidence interval in the usual way. We need no special formulas or procedures to handle the correlation.

Figure 16.20 shows the bootstrap distribution and Normal quantile plot for the sample correlation for 5000 resamples from the 24 laundry detergents in our sample. The bootstrap distribution is skewed to the left with relatively small bias. We'll need to check whether a 95% bootstrap  $t$  confidence interval is reasonable here.





**FIGURE 16.20**

The bootstrap distribution and Normal quantile plot for the correlation  $r$  for 5000 resamples from the laundry detergent data set.

The bootstrap standard error is  $SE_{\text{boot}} = 0.086$ . The  $t$  interval using the bootstrap standard error is

$$\begin{aligned} r \pm t^* SE_{\text{boot}} &= 0.671 \pm (2.074)(0.086) \\ &= 0.671 \pm 0.178 \\ &= (0.493, 0.849) \end{aligned}$$

The 95% bootstrap percentile interval is

$$\begin{aligned} (2.5 \text{ percentile}, 97.5 \text{ percentile}) &= (0.485, 0.827) \\ &= (0.671 - 0.186, 0.671 + 0.156) \end{aligned}$$

The two confidence intervals are not too different. If you feel this discrepancy is acceptable, you might want to use the percentile interval to account for the skewness in the bootstrap distribution.

While the confidence intervals give a wide range for the population correlation, both of them include only positive values. Thus, these data provide significant

evidence that there is a positive relationship between a laundry detergent's rating and its price per load.

## SECTION 16.4 Summary

Both bootstrap  $t$  and (when they exist) traditional  $z$  and  $t$  confidence intervals require statistics with small bias and sampling distributions close to Normal. We can check these conditions by examining the bootstrap distribution for bias and lack of Normality.

The **bootstrap percentile confidence interval** for 95% confidence is the interval from the 2.5 percentile to the 97.5 percentile of the bootstrap distribution. Agreement between the bootstrap  $t$  and percentile intervals is an added check on the conditions needed by the  $t$  interval. Do not use  $t$  or percentile intervals if these conditions are not met.

When bias or skewness is present in the bootstrap distribution, use a **BCa** interval. The  $t$  and percentile intervals are inaccurate under these circumstances unless the sample sizes are very large. The BCa confidence intervals adjust for bias and skewness and are generally accurate except for small samples.

## SECTION 16.4 Exercises

For Exercises 16.36 and 16.37, see page 16-33.

### 16.38 Find the 95% bootstrap percentile confidence interval.

The mean of a sample is  $\bar{x}=218.3$  and the standard deviation is  $s = 55.2$ . The mean of the bootstrap distribution is  $\bar{x}^*=220.2$  and the standard deviation is  $s^* = 11.3$ . A bootstrap distribution has the following percentiles:

Percentile								
<b>0.01</b>	<b>0.025</b>	<b>0.05</b>	<b>0.10</b>	<b>0.50</b>	<b>0.90</b>	<b>0.95</b>	<b>0.975</b>	<b>0.99</b>
193	198	202	206	220	234	238	242	246

Find the 95% bootstrap percentile confidence interval.

### 16.39 Summarize the output.

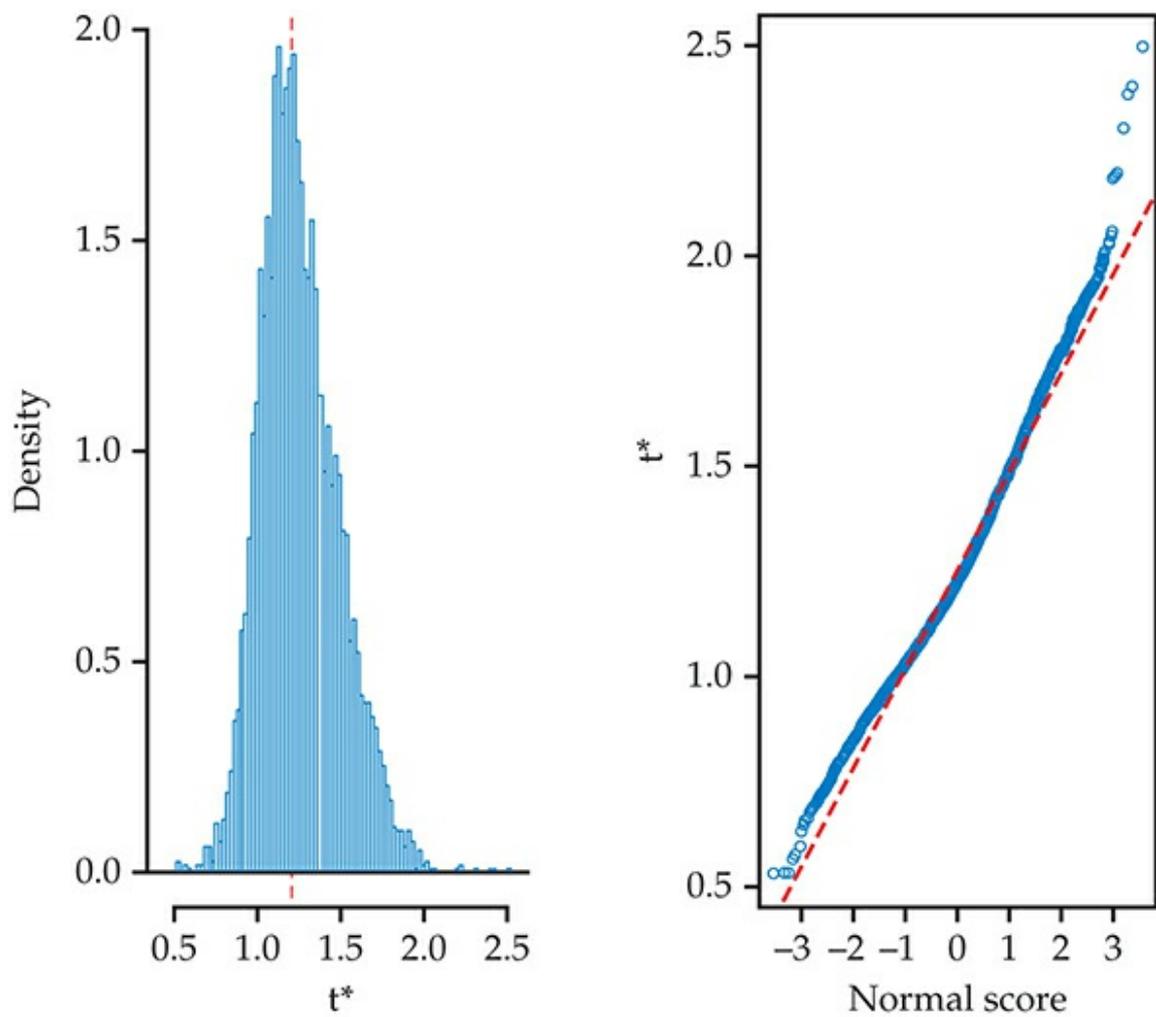
Figures 16.21 and 16.22 show software output from R with information about a bootstrap analysis. Summarize the information in the output. Be sure to include the BCa confidence interval.

### 16.40 Confidence interval for the average IQ score.

The distribution of the 60 IQ test scores in Table 1.1 (page 16) is roughly Normal, and the sample size is large enough that we expect a Normal sampling distribution. We will compare confidence intervals for the population mean IQ  $\mu$  based on this sample.  IQ

(a) Use the formula  $s/n$  to find the standard error of the mean. Give the 95%  $t$  confidence interval based on

this standard error.



**FIGURE 16.21**  
R graphical output for Exercise 16.39.

```

R Console

ORDINARY NONPARAMETRIC BOOTSTRAP

Call :
boot(data = bc, statistic = theta, R = 5000)

Bootstrap Statistics :
      original      bias      std. error
t1*   1.20713    0.04544967    0.2336016

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 5000 bootstrap replicates

CALL :
boot.ci(boot.out = corrl.boot)

Intervals :
Level      Normal          Basic
95%    (0.704, 1.620)    (0.653, 1.554)

Level      Percentile          BCa
95%    (0.860, 1.762)    (0.766, 1.671)

```

**FIGURE 16.22**

Output from R with bootstrap confidence intervals, for Exercise 16.39.

(b) Bootstrap the mean of the IQ scores. Make a histogram and a Normal quantile plot of the bootstrap distribution. Does the bootstrap distribution appear Normal? What is the bootstrap standard error? Give the 95% bootstrap  $t$  confidence interval.

(c) Give the 95% confidence percentile and BCa intervals. Make a graphical comparison by drawing a vertical line at the original sample mean  $\bar{x}$  and displaying the three intervals vertically, one above the other. How well do your four confidence intervals agree? Was bootstrapping needed to find a reasonable confidence interval, or was the formula-based confidence interval good enough?

### 16.41 Confidence interval for a Normal data set.

In Exercise 16.29 (page 16-24) you bootstrapped the mean of a simulated SRS from the standard Normal distribution  $N(0, 1)$  and found the 95% standard  $t$  and bootstrap  $t$  confidence intervals for the mean.  **NORMALD**

(a) Find the 95% bootstrap percentile confidence interval. Does this interval confirm that the  $t$  intervals are acceptable?

(b) We know that the population mean is 0. Do the confidence intervals capture this mean?

## 16.42 Using bootstrapping to check traditional methods.

Bootstrapping is a good way to check if traditional inference methods are accurate for a given sample.

Consider the following data:  DATA30

98	107	113	104	94	100	107	98	112	97
99	95	97	90	109	102	89	101	93	95
95	87	91	101	119	116	91	95	95	104

- Examine the data graphically. Do they appear to violate any of the conditions needed to use the one-sample  $t$  confidence interval for the population mean?
- Calculate the 95% one-sample  $t$  confidence interval for this sample.
- Bootstrap the data, and inspect the bootstrap distribution of the mean. Does it suggest that a  $t$  interval should be reasonably accurate? Calculate the bootstrap  $t$  95% interval.
- Find the 95% bootstrap percentile interval. Does it agree with the two  $t$  intervals? What do you conclude about the accuracy of the one-sample  $t$  interval here?

## 16.43 Comparing bootstrap confidence intervals.

The graphs in Figure 16.9 (page 16-15) do not appear to show any important skewness in the bootstrap distribution of the trimmed mean for Example 16.4. Compare the bootstrap percentile and bootstrap  $t$  intervals for the trimmed mean, given in the discussion of Example 16.4 (page 16-14). Does the

comparison suggest any skewness?  GPA

## 16.44 More on using bootstrapping to check traditional methods.

Continue to work with the data given in Exercise 16.42.  DATA30

- Find the 95% BCa confidence interval.
- Does your opinion of the robustness of the one-sample  $t$  confidence interval change when comparing it with the BCa interval?
- To check the accuracy of the one-sample  $t$  confidence interval, would you generally use the bootstrap percentile or the BCa interval? Explain.

## 16.45 BCa interval for the correlation coefficient.

Find the 95% BCa confidence interval for the correlation between price and rating, from the data in Example 16.10 (page 16-36). Is this more accurate interval in general agreement with the 95% bootstrap  $t$  and percentile intervals? Do you still agree with the judgment in the discussion of Example 16.10 that the simpler intervals are adequate?  LAUNDRY

## 16.46 Bootstrap confidence intervals for the average audio file length.

In Exercise 16.17 (page 16-17), you found a bootstrap  $t$  confidence interval for the population mean  $\mu$ . Careful examination of the bootstrap distribution reveals a slight skewness in the right tail. Is this

something to be concerned about? Bootstrap the mean and give all three 95% bootstrap confidence intervals:  $t$ , percentile, and BCa. Make a graphical comparison by displaying the three intervals vertically, one above the other. Discuss what you see.  **SONGS**

### 16.47 Bootstrap confidence intervals for service center call lengths.

The distribution of the call center lengths that you used in Exercise 16.25 (page 16-23) is strongly skewed. In that exercise you found a bootstrap  $t$  confidence interval for the population mean  $\mu$ , even though some skewness remains in the bootstrap distribution. Bootstrap the mean length and give all three bootstrap 95% confidence intervals:  $t$ , percentile, and BCa. Make a graphical comparison by drawing a vertical line at the original sample mean  $\bar{x}$  and displaying the three intervals horizontally, one above the other. Discuss what you see: Do bootstrap  $t$  and percentile agree? Does the more accurate interval agree with the two simpler methods?  **CALLS80**

### 16.48 Bootstrap confidence intervals for the standard deviation.

We would like a 95% confidence interval for the standard deviation  $\sigma$  of 150 GPAs. In Exercise 16.27 (page 16-23) we considered the bootstrap  $t$  interval. Now we have a more accurate method. Bootstrap  $s$  and report all three 95% bootstrap confidence intervals:  $t$ , percentile, and BCa. Make a graphical comparison by drawing a vertical line at the original  $s$  and displaying the three intervals vertically, one above the other. Discuss what you see: Do bootstrap  $t$  and percentile agree? Does the more accurate interval agree with the two simpler methods? What interval would you use in a report on GPAs at this college?  **GPA**

### 16.49 The effect of decreasing the sample size.

Exercise 16.15 (page 16-13) gives an SRS of 10 of the service center call lengths from Table 1.2. Describe the bootstrap distribution of  $\bar{x}$  from this sample. Give a 95% confidence interval for the population mean  $\mu$  based on these data and a method of your choice. Describe carefully how your result differs from the intervals in Exercise 16.47, which use the larger sample of 80 call lengths.  **CALLS10**

### 16.50 Bootstrap confidence interval for the GPA data.

The GPA data for females from Example 16.6 (page 16-18) are strongly skewed to the left and have a cluster of observations at 4.  **GPA**

- Bootstrap the mean of the data. Based on the bootstrap distribution, which bootstrap confidence intervals would you consider for use? Explain your answer.
- Find all three bootstrap confidence intervals. How do the intervals compare? Briefly explain the reasons for any differences. In particular, what kind of errors would you make in estimating the mean GPA by using a  $t$  interval or a percentile interval instead of a BCa interval?

### 16.51 Bootstrap confidence intervals for the difference in GPAs.

Example 16.6 (page 16-18) considers the difference in mean GPAs of men and women. The bootstrap distribution appeared reasonably Normal. Give the 95% BCa confidence interval for the difference in mean GPAs. Is this interval comparable to the bootstrap  $t$  interval calculated in the example?  **GPA**

## 16.52 The correlation between GPA and high school math grades.

The study described in Example 16.4 (page 16-14) used high school grades to predict GPA. For this exercise, we will look at the correlation between GPA and high school math grades.  GPA

- Describe the distribution of GPAs. Do the same for high school math grades.
- Describe the relationship between GPA and high school math grades.
- Generate 2000 resamples and use these to obtain the bootstrap distribution for the correlation.
- Describe the shape and bias of the bootstrap distribution. Does use of the simpler bootstrap confidence intervals ( $t$  and percentile) appear to be justified?
- Find all three 95% bootstrap confidence intervals:  $t$ , percentile, and BCa. Make a graphical comparison by drawing a vertical line at the original correlation  $r$  and displaying the three intervals vertically, one above the other. Discuss what you see. Does it still appear that the simpler intervals are justified? What confidence interval would you include in a report describing the relationship between GPA and high school math grades?

## 16.53 The correlation between debts.

Figure 2.4 (page 92) shows a strong positive relationship between debt in 2010 and debt in 2009 for 33 countries. Use the bootstrap to perform statistical inference for these data.  DEBT

- Describe the shape and bias of the bootstrap distribution. Do you think that a simple bootstrap inference ( $t$  and percentile confidence intervals) is justified? Explain your answer.
- Give the 95% BCa and bootstrap percentile confidence intervals for the population correlation. Do they (as expected) agree closely? Do these intervals provide significant evidence at the 5% level that the population correlation is not 0?

## 16.54 Bootstrap distribution for the slope $\beta_1$ .

Describe carefully how to resample from data on an explanatory variable  $x$  and a response variable  $y$  to create a bootstrap distribution for the slope  $b_1$  of the least-squares regression line.

## 16.55 Predicting ratings of laundry detergents.

Refer to Example 16.10 (page 16-36).  LAUNDRY

- Find the least-squares regression line for predicting rating from price.
- Bootstrap the regression line and give a 95% confidence interval for the slope of the population regression line.
- Compare the bootstrap results with the usual method for finding a confidence interval for a regression slope.

## 16.56 Predicting GPA.

Continue your study of GPA and high school math grades, begun in Exercise 16.52, by performing a regression to predict GPA using high school math grades as the explanatory variable.  GPA

- (a) Plot the residuals against the math grades and make a Normal quantile plot of the residuals. Do these plots suggest that inference based on the usual simple linear regression model may be inaccurate? Give reasons for your answer.
- (b) Examine the bootstrap distribution of the slope  $b_1$  of the least-squares regression line. Based on what you see, what do you recommend regarding the use of bootstrap  $t$  or bootstrap percentile intervals? Give reasons for your recommendation.
- (c) Give the 95% BCa confidence interval for the slope  $\beta_1$  of the population regression line. Compare this with the standard 95% confidence interval based on Normality, the bootstrap  $t$  interval, and the bootstrap percentile interval. Using the BCa interval as a standard, which of the other intervals are adequately accurate for practical use?

### 16.57 Predicting debt in 2010 from debt in 2009.

Continue your study of the relationship between debt in 2009 and debt in 2010 for 33 countries, begun in Exercise 16.53. Run the regression to predict debt in 2010 using debt in 2009 as the explanatory variable.  DEBT

- (a) Plot the residuals against the explanatory variable and make a Normal quantile plot of the residuals. Do the residuals appear to be Normal? Explain your answer.
- (b) Examine the shape and bias of the bootstrap distribution of the slope  $b_1$  of the least-squares line. Does this distribution suggest that even the bootstrap  $t$  interval will be accurate? Give a reason for your answer.
- (c) Find the standard 95%  $t$  confidence interval for  $\beta_1$  and also the BCa, bootstrap  $t$ , and bootstrap percentile confidence intervals. What do you conclude about the accuracy of the two  $t$  intervals?

### 16.58 The effect of outliers.

We know that outliers can strongly influence statistics such as the mean and the least-squares line. Example 7.7 (page 429) describes a matched pairs study of disruptive behavior by dementia patients. The differences in Table 7.2 show several low values that may be considered outliers.  MOON

- (a) Bootstrap the mean of the differences with and without the three low values. How do these values influence the shape and bias of the bootstrap distribution?
- (b) Give the BCa confidence interval from both bootstrap distributions. Discuss the differences.

## 16.5 Significance Testing Using Permutation Tests

When you complete this section, you will be able to

- Outline the steps needed for a permutation test for comparing two means.
- Outline the steps needed for a permutation test for a matched pairs study.
- Outline the steps needed for a permutation test for the relationship between two quantitative variables.



**tests of significance, p. 372**

Significance tests tell us whether an observed effect, such as a difference between two means or a correlation between two variables, could reasonably occur “just by chance” in selecting a random sample. If not, we have evidence that the effect observed in the sample reflects an effect that is present in the population. The reasoning of tests goes like this:

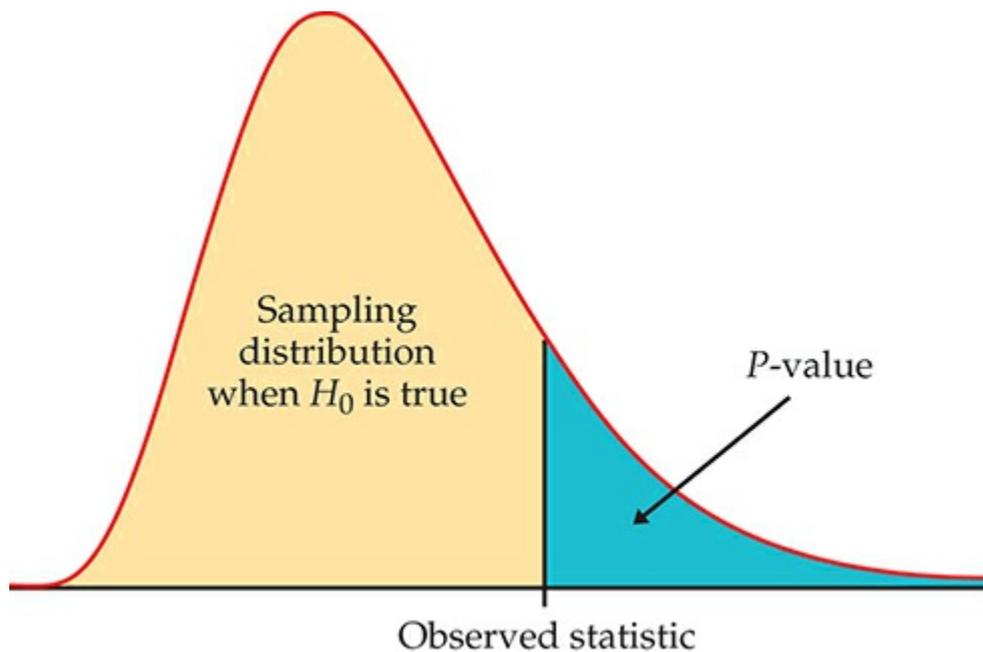
1. Choose a statistic that measures the effect you are looking for.
2. Construct the sampling distribution that this statistic would have if the effect were *not* present in the population.
3. Locate the observed statistic on this distribution. A value in the main body of the distribution could easily occur just by chance. A value in the tail would rarely occur by chance and so is evidence that something other than chance is operating.



**null hypothesis, p. 374**

The statement that the effect we seek is *not* present in the population is the null hypothesis,  $H_0$ . Assuming the null hypothesis is true, the probability that we would observe a statistic value as extreme or more extreme than the one we did observe is the  $P$ -value. Figure 16.23 illustrates the idea of a  $P$ -value. Small  $P$ -values are evidence against the null hypothesis and in favor of a real effect in the population. The reasoning of statistical tests is indirect and a bit subtle but is by now familiar. Tests based on resampling don't change this reasoning. They find  $P$ -values by resampling calculations rather than from formulas and so can be used in settings

where traditional tests don't apply.



**FIGURE 16.23**

The  $P$ -value of a statistical test is found from the sampling distribution the statistic would have if the null hypothesis were true. It is the probability of a result at least as extreme as the value we actually observed.

← **LOOK BACK**

**$P$ -value, p. 377**

Because  $P$ -values are calculated *acting as if the null hypothesis were true*, we cannot resample from the observed sample as we did earlier. In the absence of bias, resampling from the original sample creates a bootstrap distribution centered at the observed value of the statistic. If the null hypothesis is in fact not true, this value may be far from the parameter value stated by the null hypothesis. We must estimate what the sampling distribution of the statistic would be if the null hypothesis were true. That is, we must obey this rule:

**RESAMPLING FOR SIGNIFICANCE TESTS**

To estimate the  $P$ -value for a test of significance, estimate the sampling distribution of the test statistic when the null hypothesis is true by resampling in a manner that is consistent with the null hypothesis.

## EXAMPLE

### 16.11 “Directed reading activities.”



Do new “directed reading activities” improve the reading ability of elementary school students, as measured by their Degree of Reading Power (DRP) scores? A study assigns students at random to either the new method (treatment group, 21 students) or traditional teaching methods (control group, 23 students). The DRP scores at the end of the study appear in Table 16.1.<sup>7</sup> In Example 7.15 (page 454) we applied the two-sample  $t$  test to these data.

To apply resampling, we will start with the difference between the sample means as a measure of the effect of the new activities:

$$\text{statistic} = \bar{x}_{\text{treatment}} - \bar{x}_{\text{control}}$$

The null hypothesis  $H_0$  for the resampling test is that the teaching method has no effect on the distribution of DRP scores. If  $H_0$  is true, the DRP scores in Table 16.1 do not depend on the teaching method. Each student has a DRP score that describes that child and is the same no matter which group the child is assigned to. The observed difference in group means just reflects the accident of random assignment to the two groups.

Now we can see how to resample in a way that is consistent with the null hypothesis: imitate many repetitions of the random assignment of students to treatment and control groups, with each student always keeping his or her DRP score unchanged. Because resampling in this way scrambles the assignment of students to groups, tests based on resampling are called **permutation tests**, from the mathematical name for scrambling a collection of things.

permutation test

TABLE 16.1

Degree of Reading Power Scores for Third-Graders

	Treatment group				Control group			
24	61	59	46	42	33	46	37	
43	44	52	43	43	41	10	42	
58	67	62	57	55	19	17	55	
71	49	54		26	54	60	28	
43	53	57		62	20	53	48	

49	56	33	37	85	42
----	----	----	----	----	----

Here is an outline of the permutation test procedure for comparing the mean DRP scores in Example 16.11:

- Choose 21 of the 44 students at random to be the treatment group; the other 23 are the control group. This is an ordinary SRS, chosen *without replacement*. It is called a **permutation resample**.

**permutation resample**

- Calculate the mean DRP score in each group, using the students' DRP scores in Table 16.1. The difference between these means is our statistic.
- Repeat this resampling and calculation of the statistic hundreds of times. The distribution of the statistic from these resamples estimates the sampling distribution under the condition that  $H_0$  is true. It is called a **permutation distribution**.

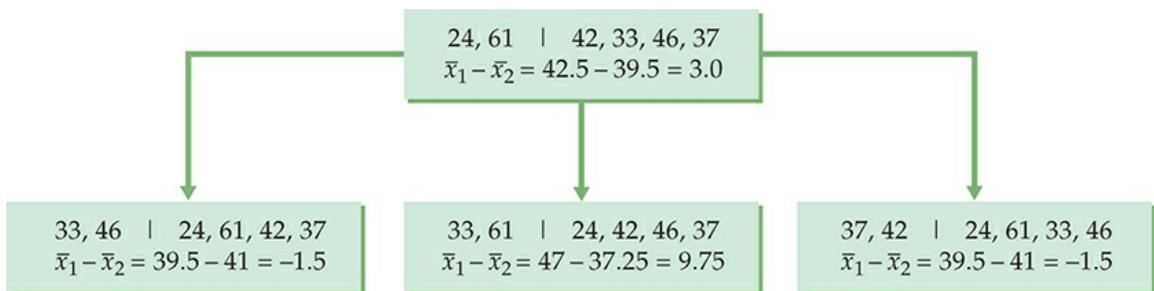
**permutation distribution**

- Consider the value of the statistic actually observed in the study,

$$\bar{x}^{\text{treatment}} - \bar{x}^{\text{control}} = 51.476 - 41.522 = 9.954$$

Locate this value on the permutation distribution to get the  $P$ -value.

Figure 16.24 illustrates permutation resampling on a small scale. The top box shows the results of a study with four subjects in the treatment group and two subjects in the control group. A permutation resample chooses an SRS of four of the six subjects to form the treatment group. The remaining two are the control group. The results of three permutation resamples appear below the original results, along with the statistic (difference in group means) for each.



**FIGURE 16.24**

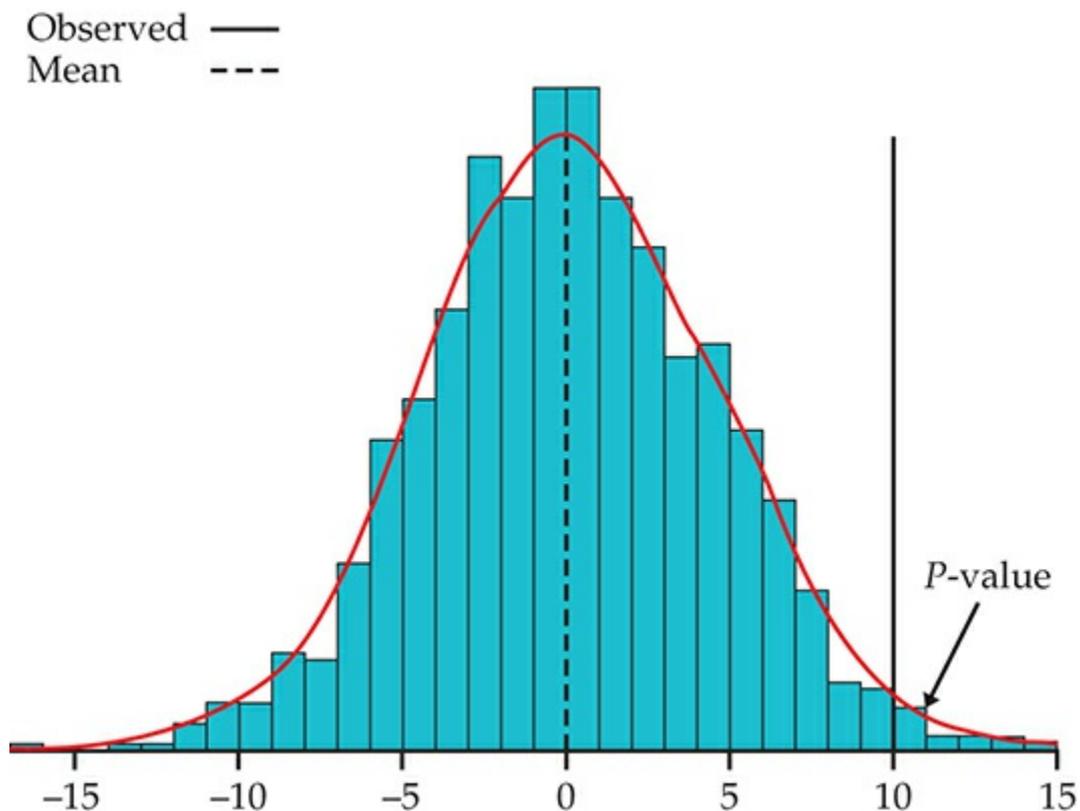
The idea of permutation resampling. The top box shows the outcome of a study with four subjects in one group and two in the other. The boxes below show three permutation resamples. The values of the statistic for many such resamples form the permutation distribution.

## EXAMPLE

### 16.12 Permutation test for the DRP study.



Figure 16.25 shows the permutation distribution of the difference in means based on 1000 permutation resamples from the DRP data in Table 16.1. This is a resampling estimate of the sampling distribution of the statistic when the null hypothesis  $H_0$  is true. As  $H_0$  suggests, the distribution is centered at 0 (no effect). The solid vertical line in the figure marks the location of the statistic for the original sample, 9.954. Use the permutation distribution exactly as if it were the sampling distribution: the  $P$ -value is the probability that the statistic takes a value at least as extreme as 9.954 in the direction given by the alternative hypothesis.



**FIGURE 16.25**

The permutation distribution of the difference between the treatment mean and the control mean based on the DRP scores of 44 students, for Example 16.12. The dashed line marks the mean of the permutation distribution: it is very close to zero, the value specified by the null hypothesis. The solid vertical line marks the observed difference in means, 9.954. Its location in the right tail shows that a value this large is unlikely to occur when the null hypothesis is true.

We seek evidence that the treatment increases DRP scores, so the alternative hypothesis is that the distribution of the statistic  $\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}}$  is centered not at 0 but at some positive value. Large values of the statistic are evidence against the null hypothesis in favor of this one-sided alternative. The permutation test  $P$ -value is the proportion of the 1000 resamples that give a result at least as great as 9.954. A look at the resampling results finds that 14 of the 1000 resamples gave a value of 9.954 or larger, so the estimated  $P$ -value is 14/1000, or 0.014.

Figure 16.25 shows that the permutation distribution has a roughly Normal shape. Because the permutation distribution approximates the sampling distribution, we now know that the sampling distribution is close to Normal. When the sampling distribution is close to Normal, we can safely apply the usual two-sample  $t$  test. The  $t$  test in Example 7.15 gives  $P = 0.013$ , very close to the  $P$ -value from the permutation test.

## Using software

In principle, you can program almost any statistical software to do a permutation test. It is more convenient to use software that automates the process of resampling, calculating the statistic, forming the permutation distribution, and finding the  $P$ -value. The package `perm` in R contains functions that allow you to request permutation tests. The permutation distribution in Figure 16.25 is one output. Another is this summary of the test results:

```
Exact Permutation Test Estimated by Monte Carlo

data: trtgrp and ctrlgrp
p-value = 0.0154
alternative hypothesis: true mean trtgrp - mean ctrlgrp
is greater than 0
sample estimates:
mean trtgrp - mean ctrlgrp
          9.954451

p-value estimated from 5000 Monte Carlo replications
99 percent confidence interval on p-value:
 0.01110640 0.02024333
```

By giving “greater” as the alternative hypothesis, the output makes it clear that 0.015 is the one-sided  $P$ -value. This estimate of the  $P$ -value is more precise than the 0.014 estimate because it is based on 5000 rather than 1000 resamples.

## Permutation tests in practice

**two-sample  $t$  test, page 454**

**Permutation tests versus  $t$  tests.** We have analyzed the data in Table 16.1 both by the two-sample  $t$  test (in Chapter 7) and by a permutation test. Comparing the two approaches brings out some general points about permutation tests versus traditional formula-based tests.

- The hypotheses for the  $t$  test are stated in terms of the two population means,

$$H_0: \mu_{\text{treatment}} - \mu_{\text{control}} = 0$$

$$H_a: \mu_{\text{treatment}} - \mu_{\text{control}} > 0$$

The permutation test hypotheses are more general. The null hypothesis is “same distribution of scores in both groups,” and the one-sided alternative is “scores in the treatment group are systematically higher.” These more general hypotheses imply the  $t$  hypotheses if we are interested in mean scores and the two distributions have the same shape.

- The plug-in principle says that the difference in sample means estimates the difference in population means. The  $t$  statistic starts with this difference. We used the same statistic in the permutation test, but that was a choice: we could use the difference in 25% trimmed means or any other statistic that measures the effect of treatment versus control.
- The  $t$  test statistic is based on standardizing the difference in means in a clever way to get a statistic that has a  $t$  distribution when  $H_0$  is true. The permutation test works directly with the difference in means (or some other statistic) and estimates the sampling distribution by resampling. No formulas are needed.
- The  $t$  test gives accurate  $P$ -values if the sampling distribution of the difference in means is at least roughly Normal. The permutation test gives accurate  $P$ -values even when the sampling distribution is not close to Normal.

The permutation test is useful even if we plan to use the two-sample  $t$  test. Rather than relying on Normal quantile plots of the two samples and the central limit theorem, we can directly check the Normality of the sampling distribution by looking at the permutation distribution. Permutation tests provide a “gold standard” for assessing two-sample  $t$  tests. If the two  $P$ -values differ considerably, it usually indicates that the conditions for the two-sample  $t$  don’t hold for these data. Because permutation tests give accurate  $P$ -values even when the sampling distribution is skewed, they are often used when accuracy is very important. Here is an example.

## EXAMPLE

### 16.13 Permutation test for GPAs.



GPA

In Example 16.6 (page 16-18), we looked at the difference in mean GPAs of male and female students. Figure 16.10 (page 16-18) shows both distributions. Because the distributions are skewed and the sample sizes are somewhat different, a two-sample  $t$  test might be inaccurate.

Based on the summary statistics,

Gender	$n$	$\bar{x}$	$s$
Male	91	2.784	0.859
Female	59	2.933	0.748
Difference		-0.149	

the  $t$  statistic is  $-1.12$  with either 58 or 135.73 degrees of freedom. The  $P$ -value is roughly 0.26 in either case.

We perform permutation tests with 5000 resamples using R. We use the difference in means,  $\bar{x}_1 - \bar{x}_2$ , as our test statistic. This is done by randomly regrouping the total set of GPAs into two groups that are the same sizes as the two original samples. This is consistent with the null hypothesis that gender has no effect on GPA. Each GPA appears once in the data of each resample, but some GPAs move from the male to the female group, and vice versa. We calculate the test statistic for each resample and create its permutation distribution. The  $P$ -value is the proportion of the resamples with statistics that exceed the observed statistic.

A 99% confidence interval for the  $P$ -value based on the 5000 resamples is (0.256, 0.309). This interval contains the  $P$ -value for the  $t$  test. The skewness and differing sample sizes do not have an impact here primarily because the sample sizes are relatively large.

If you read Chapter 15 on nonparametric tests, you will find there more comparison of permutation tests with rank tests as well as tests based on Normal distributions.

### Data from an entire population.

A subtle difference between confidence intervals and significance tests is that

confidence intervals require the distinction between sample and population, but tests do not. If we have data on an entire population—say, all employees of a large corporation—we don't need a confidence interval to estimate the difference between the mean salaries of male and female employees. We can calculate the means for all men and for all women and get an exact answer. But it still makes sense to ask, "Is the difference in means so large that it would rarely occur just by chance?" A test and its  $P$ -value answer that question.

Permutation tests are a convenient way to answer such questions. In carrying out the test we pay no attention to whether the data are a sample or an entire population. The resampling assigns the full set of observed salaries at random to men and women and builds a permutation distribution from repeated random assignments. We can then see if the observed difference in mean salaries is so large that it would rarely occur if gender did not matter.

### When are permutation tests valid?

The two-sample  $t$  test starts from the condition that the sampling distribution of  $\bar{x}_1 - \bar{x}_2$  is Normal. This is the case if both populations have Normal distributions, and it is approximately true for large samples from non-Normal populations because of the central limit theorem. The central limit theorem helps explain the robustness of the two-sample  $t$  test. The test works well when both populations are symmetric, especially when the two sample sizes are similar.



two-sample  $t$  test, page 454



Robustness of two-sample procedures, p. 455



The permutation test completely removes the Normality condition. However, *resampling in a way that moves observations between the two groups requires that the two populations are identical when the null hypothesis is true—that not only their means are the same but also their spreads and shapes.* Our preferred version of the two-sample  $t$  allows different standard deviations in the two groups, so the shapes are both Normal but need not have the same spread.

In Example 16.13, the distributions are skewed but we do not rule out the  $t$  test because of the central limit theorem. The permutation test is valid if the GPA distributions for males and females have the same shape, so that they are identical

under the null hypothesis that the centers (the means) are the same. Based on Figure 16.10 (page 16-18), it appears that the distribution for the males has a little more spread than the distribution for the females. Fortunately, the permutation test is robust. That is, it gives accurate  $P$ -values when the two population distributions have somewhat different shapes, such as when they have slightly different standard deviations.

### **Sources of variation.**

Just as in the case of bootstrap confidence intervals, permutation tests are subject to two sources of random variability: the original sample is chosen at random from the population, and the resamples are chosen at random from the sample. Again as in the case of the bootstrap, the added variation due to resampling is usually small and can be made as small as we like by increasing the number of resamples.

The number of resamples on which a permutation test is based determines the number of decimal places and precision in the resulting  $P$ -value. Tests based on 1000 resamples give  $P$ -values to three places (multiples of 0.001), with a margin of error of  $2P(1-P)/1000$  equal to 0.014 when the true one-sided  $P$ -value is 0.05. If higher precision is needed or your computer is sufficiently fast, you may choose to use 10,000 or more resamples.

## **USE YOUR KNOWLEDGE**

### **16.59 Is a permutation test valid?**

Suppose a professor wants to compare the effectiveness of two different instruction methods. By design, one method is more team oriented, so he expects the variability in individual tests scores for this method to be smaller. Is it valid to use a permutation test to compare the mean scores of the two methods? Explain.

### **16.60 Declaring significance.**

Suppose that a one-sided permutation test based on 250 permutation resamples resulted in a  $P$ -value of 0.04. What is the approximate standard deviation of the distribution? Would you feel comfortable declaring the results significant at the 5% level? Explain.

## Permutation tests in other settings

The bootstrap procedure can replace many different formula-based confidence intervals, provided that we resample in a way that matches the setting. Permutation testing is also a general method that we can adapt to various settings.

### GENERAL PROCEDURE FOR PERMUTATION TESTS

To carry out a permutation test based on a statistic that measures the size of an effect of interest:

1. Compute the statistic for the original data.
2. Choose permutation resamples from the data without replacement in a way that is consistent with the null hypothesis of the test and with the study design. Construct the permutation distribution of the statistic from its values in a large number of resamples.
3. Find the  $P$ -value by locating the original statistic on the permutation distribution.

### Permutation test for matched pairs.

The key step in the general procedure for permutation tests is to form permutation resamples in a way that is consistent with the study design and with the null hypothesis. Our examples to this point have concerned two-sample settings. How must we modify our procedure for a matched pairs design?

### EXAMPLE

#### 16.14 Permutation test for full-moon study.



Can the full moon influence behavior? A study observed 15 nursing-home patients with dementia. The number of incidents of aggressive behavior was recorded each day for 12 weeks. Call a day a “moon day” if it is the day of a full moon or the day before or after a full moon. Table 16.2 gives the average number of aggressive incidents for moon days and other days for each subject.<sup>8</sup> These are matched pairs data. In Example 7.7 (page 429), the matched pairs  $t$  test found evidence that the mean number of aggressive incidents is higher on moon days ( $t = 6.45$ ,  $df = 14$ ,  $P < 0.001$ ). The data show some signs of non-Normality. We want to apply a permutation test.

The null hypothesis says that the full moon has no effect on behavior. If this is true, the two entries for each patient in Table 16.2 are two measurements of aggressive behavior made under the same conditions. There is no distinction between “moon days” and “other days.” Resampling in a way consistent with this null hypothesis randomly assigns one of each patient’s two scores to “moon” and the other to “other.” We don’t mix results for different subjects, because the original data are paired.

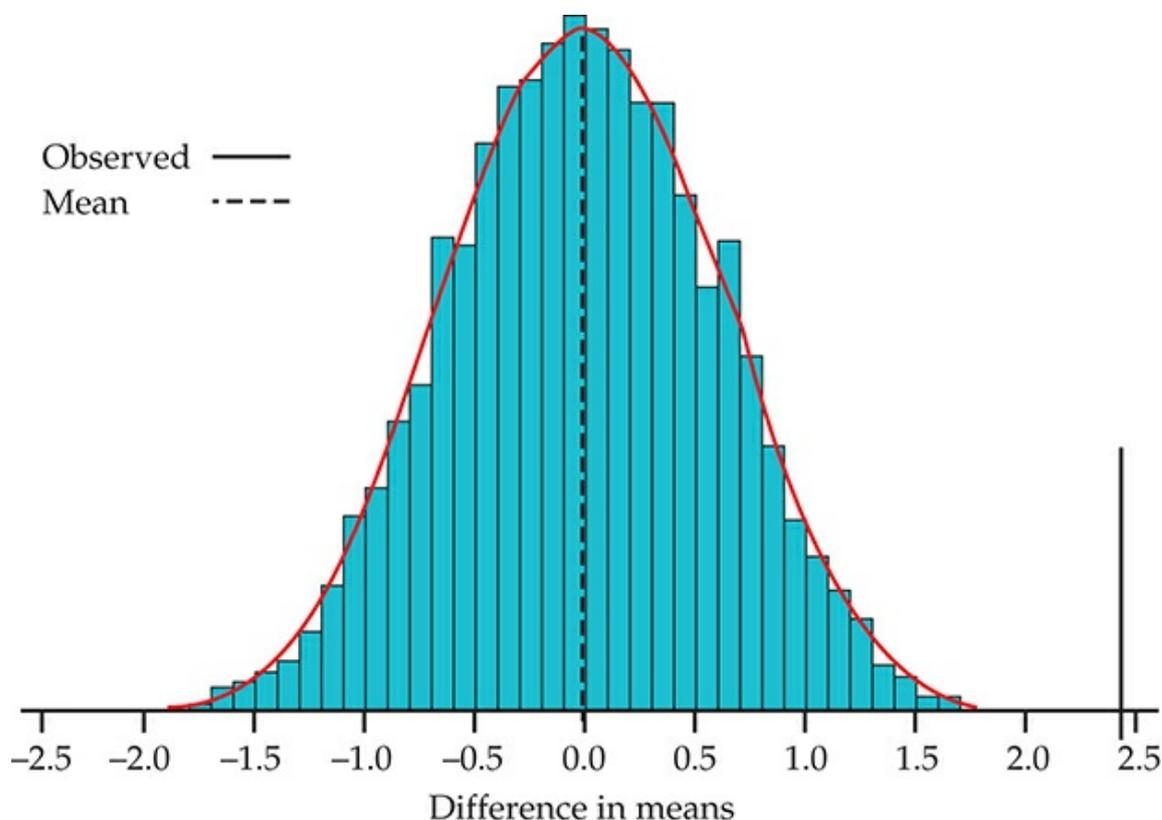
The permutation test (like the matched pairs  $t$  test) uses the difference in means  $\bar{x}_{\text{moon}} - \bar{x}_{\text{other}}$ . Figure 16.26 shows the permutation distribution of this statistic from 10,000 resamples. None of these resamples produces a difference as large as the observed difference,  $\bar{x}_{\text{moon}} - \bar{x}_{\text{other}} = 2.433$ . The estimated one-sided  $P$ -value is less than 1 in a thousand. We report this result as  $P < 0.0001$ . There is strong evidence that aggressive behavior is more common on moon days.

**TABLE 16.2**

**Aggressive Behaviors of Dementia Patients**

Patient	Moon days	Other days	Patient	Moon days	Other days
1	3.33	0.27	9	6.00	1.59
2	3.67	0.59	10	4.33	0.60

3	2.67	0.32	11	3.33	0.65
4	3.33	0.19	12	0.67	0.69
5	3.33	1.26	13	1.33	1.26
6	3.67	0.11	14	0.33	0.23
7	4.67	0.30	15	2.00	0.38
8	2.67	0.40			



**FIGURE 16.26**

The permutation distribution for the mean difference (moon days minus other days) from 10,000 paired resamples from the data in Table 16.2, for Example 16.14.

The permutation distribution in Figure 16.26 is close to Normal, as a Normal quantile plot confirms. The matched pairs  $t$  test is therefore reliable and agrees with the permutation test that the  $P$ -value is very small.

### Permutation test for the significance of a relationship.

Permutation testing can be used to test the significance of a relationship between two variables. For example, in Example 16.10 we looked at the relationship between price and rating of laundry detergents.

The null hypothesis is that there is no relationship. In that case, prices are assigned to detergents for reasons that have nothing to do with rating. We can resample in a way consistent with the null hypothesis by permuting the observed ratings among the detergents at random.

Take the correlation as the test statistic. For every resample, calculate the

correlation between the prices (in their original order) and ratings (in the reshuffled order). The  $P$ -value is the proportion of the resamples with correlation larger than the original correlation.

### When can we use permutation tests?

We can use a permutation test only when we can see how to resample in a way that is consistent with the study design and with the null hypothesis. We now know how to do this for the following types of problems:

- **Two-sample problems** when the null hypothesis says that the two populations are identical. We may wish to compare population means, proportions, standard deviations, or other statistics. You may recall from Section 7.3 that traditional tests for comparing population standard deviations work very poorly. Permutation tests are a much better choice.
- **Matched pairs designs** when the null hypothesis says that there are only random differences within pairs. A variety of comparisons is again possible.
- **Relationships between two quantitative variables** when the null hypothesis says that the variables are not related. The correlation is the most common measure of association, but not the only one.



These settings share the characteristic that the null hypothesis specifies a simple situation such as two identical populations or two unrelated variables. We can see how to resample in a way that matches these situations. *Permutation tests can't be used for testing hypotheses about a single population, comparing populations that differ even under the null hypothesis, or testing general relationships.* In these settings, we don't know how to resample in a way that matches the null hypothesis. Researchers are developing resampling methods for these and other settings, so stay tuned.

When we can't do a permutation test, we can often calculate a bootstrap confidence interval instead. If the confidence interval fails to include the null hypothesis value, then we reject  $H_0$  at the corresponding significance level. This is not as accurate as doing a permutation test, but a confidence interval estimates the size of an effect as well as giving some information about its statistical significance. Even when a test is possible, it is often helpful to report a confidence interval along with the test result. Confidence intervals don't assume that a null hypothesis is true, so we use bootstrap resampling with replacement rather than permutation resampling without replacement.

## SECTION 16.5 Summary

**Permutation tests** are significance tests based on **permutation resamples** drawn at random from the original data. Permutation resamples are drawn **without replacement**, in contrast to bootstrap samples, which are drawn with replacement.

Permutation resamples must be drawn in a way that is consistent with the null hypothesis and with the study design. In a **two-sample design**, the null hypothesis says that the two populations are identical. Resampling randomly reassigns observations to the two groups. In a **matched pairs** design, randomly permute the two observations within each pair separately. To test the hypothesis of **no relationship** between two variables, randomly reassign values of one of the two variables.

The **permutation distribution** of a suitable statistic is formed by the values of the statistic in a large number of resamples. Find the  $P$ -value of the test by locating the original value of the statistic on the permutation distribution.

When they can be used, permutation tests have great advantages. They do not require specific population shapes such as Normality. They apply to a variety of statistics, not just to statistics that have a simple distribution under the null hypothesis. They can give very accurate  $P$ -values, regardless of the shape and size of the population (if enough permutations are used).

It is often useful to give a confidence interval along with a test. To create a confidence interval, we no longer assume that the null hypothesis is true, so we use bootstrap resampling rather than permutation resampling.

## SECTION 16.5 Exercises

*For Exercises 16.59 and 16.60, see page 16-49.*

### 16.61 Marketing cell phones.

You have two prototypes of a new cell phone and designed an experiment to help you decide which one to market. Forty students were randomly assigned to use one of the two phones for two weeks. Their overall satisfaction with the phone is recorded on a subjective scale with a range of 1 to 100. Outline the steps needed to compare the means for the two phones using a permutation test.

### 16.62 Marketing cell phones.

Refer to the previous exercise. Suppose that you had each of the 40 students use both phones. Outline the steps needed to compare the means for the two phones using a permutation test.

### 16.63 Characteristics of cell phones.

Refer to Exercise 16.61. Before asking the students to provide an overall satisfaction rating, they were asked to provide ratings for several characteristics of the cell phone. Two of these were satisfaction with the screen and satisfaction with the keyboard. Outline the steps needed to evaluate the relationship between these two variables for the first phone using a permutation test.

## 16.64 Compare the correlations.

Refer to the previous exercise. Suppose that you calculate the correlation between satisfaction with the screen and satisfaction with the keyboard for each phone. Outline the steps needed to compare these two correlations using a permutation test.

## 16.65 A small-sample permutation test.

To illustrate the process, let's perform a permutation test by hand for a small random subset of the DRP data (Example 16.11, page 16-43). Here are the data:

Treatment group	57	53		
Control group	19	37	41	42

- Calculate the difference in means  $\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}}$  between the two groups. This is the observed value of the statistic.
- Resample: Start with the 6 scores and choose an SRS of 2 scores to form the treatment group for the first resample. You can do this by labeling the scores from 1 to 6 and using consecutive random digits from Table B or by rolling a die. Using either method, be sure to skip repeated digits. A resample is an ordinary SRS, without replacement. The remaining 4 scores are the control group. What is the difference in group means for this resample?
- Repeat Step (b) 20 times to get 20 resamples and 20 values of the statistic. Make a histogram of the distribution of these 20 values. This is the permutation distribution for your resamples.
- What proportion of the 20 statistic values were equal to or greater than the original value in part (a)? You have just estimated the one-sided  $P$ -value for the original 6 observations.
- For this small data set, there are only 15 possible permutations of the data. As a result, we can calculate the exact  $P$ -value by counting the number of permutations with a statistic value greater than or equal to the original value and then dividing by 15. What is the exact  $P$ -value here? How close was your estimate?

## 16.66 Product labels with animals?

Participants in a study were asked to indicate their attitude toward a product on a seven-point scale (from 1 = dislike very much to 7 = like very much). A bottle of MagicCoat pet shampoo, with a picture of a collie on the label, was the product. Prior to indicating this preference, subjects were randomly assigned to two groups and were asked to do a word find. Four of the words were common to both groups and four were either related to the product image or conflicted with the image. The group with words related to the product image were considered primed. In Exercise 7.72 (page 469) the mean scores were compared using the two-sample  $t$  procedures. Let's use a permutation test for the comparison. Here are the data: 

### BRANDPR

Group	Brand Attitude
Primed	2 2 3 3 3 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5
Nonprimed	1 1 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 5

- Examine the scores of each group graphically. Is it appropriate to use the two-sample  $t$  procedures? Explain your answer.
- Perform the two-sample  $t$  test to compare the group means. Use a two-sided alternative hypothesis and a significance level of 5%.

- (c) Perform a permutation test to compare the group means. Summarize your results and conclusions.
- (d) Write a short summary comparing your results in parts (b) and (c). Which method do you recommend for these data? Give reasons for your answer.

### 16.67 Timing of food intake.

Examples 7.16 and 7.17 (pages 456 and 457) examine data on an experiment to compare weight loss in subjects who were classified as early eaters or late eaters, based on the timing of their main meal. In

Example 7.17, the following data were analyzed:  **FOOD10**

Group	Weight loss (kg)				
Early eater	6.3	15.1	9.4	16.8	10.2
Late eater	7.8	0.2	1.5	11.5	4.6

- (a) State appropriate null and alternative hypotheses for these data.
- (b) Report the result of the pooled two-sample  $t$  test.
- (c) Perform a permutation test to compare the two means and report the results. Compare the  $P$ -value for this test with the  $P$ -value for the  $t$  test in part (b).
- (d) Find a BCa confidence interval for the difference in means. How is this interval related to your results in part (c)?

### 16.68 Standard deviation of the estimated $P$ -value.

The estimated  $P$ -value for the DRP study (Example 16.12, page 16-45) based on 1000 resamples is  $P = 0.015$ . Suppose that we obtained the same  $P$ -value based on 4000 resamples. What is the approximate standard deviation of each of these  $P$ -values?

### 16.69 When is a permutation test valid?

You want to test the equality of the means of two populations. Sketch density curves for two populations for which

- (a) a permutation test is valid but a  $t$  test is not.
- (b) both permutation and  $t$  tests are valid.
- (c) a  $t$  test is valid but a permutation test is not.

### 16.70 Testing the correlation between debts.

In Exercise 16.53 (page 16-41), we assessed the significance of the *correlation* between debt in 2009 and debt in 2010 for 33 countries by creating bootstrap confidence intervals. If a 95% confidence interval does not cover 0, the observed correlation is significantly different from 0 at the  $\alpha$  level. Let's do a test that provides a  $P$ -value. Carry out a permutation test and give the  $P$ -value. What do you conclude? Is your conclusion consistent with your work in Exercise 16.53 (page 16-41)?  **DEBT**

### 16.71 Assessing a summer language institute.

Exercise 7.45 (page 446) gives data on a study of the effect of a summer language institute on the ability of high school language teachers to understand spoken French. This is a matched pairs study, with scores for 20 teachers at the beginning (pretest) and end (posttest) of the institute. We conjecture that the posttest scores are higher on the average.  **FRENCH**

- (a) Carry out the matched pairs  $t$  test. That is, state hypotheses, calculate the test statistic, and give its  $P$ -value.
- (b) Make a Normal quantile plot of the gains: posttest score—pretest score. The data have a number of ties and a low outlier. A permutation test can help check the  $t$  test result.
- (c) Carry out the permutation test for the *difference in means in a matched pairs setting*, using 9999 resamples. The Normal quantile plot shows that the permutation distribution is reasonably Normal. What is the  $P$ -value for the permutation test? Do your tests in parts (a) and (c) lead to the same practical conclusion?

### 16.72 Compare the medians.

Refer to the previous exercise. Use a permutation test to compare the medians. Write a short summary of your results and conclusions. Include a comparison of what you found here with what you found in the previous exercise.  **FRENCH**

### 16.73 Testing the correlation between price and rating.

Example 16.10 (page 16-36) uses the bootstrap to find a confidence interval for the correlation between price and rating for 24 laundry detergents. Let's use a permutation test to examine this correlation.  **LAUNDRY**

- (a) State the null and alternative hypotheses.
- (b) Perform a permutation test based on the sample correlation. Report the  $P$ -value and draw a conclusion.

### 16.74 Comparing mpg calculations.

Exercise 7.39 (page 445) gives data on a comparison of driver and computer mpg calculations. This is a matched pairs study, with mpg values for 20 fill-ups.  **MPG20**

- (a) Carry out the matched pairs  $t$  test. That is, state hypotheses, calculate the test statistic, and give its  $P$ -value.
- (b) A permutation test can help check the  $t$  test result. Carry out the permutation test for the *difference in means in a matched pairs setting*, using 10,000 resamples. What is the  $P$ -value for the permutation test? Does this test and the test in part (a) lead to the same practical conclusion?

### 16.75 Comparing the average northern and southern tree diameter.

In Exercise 7.107 (page 480), the standard deviations of tree diameters for the northern and southern regions of the tract were compared. This test is unreliable because it is sensitive to non-Normality of the data. Perform a permutation test using the  $F$  statistic (ratio of sample variances) as your statistic. What do you conclude? Are the two tests comparable?  **NSPINES**

## 16.76 Comparing serum retinol levels.

The formal medical term for vitamin A in the blood is serum retinol. Serum retinol has various beneficial effects, such as protecting against fractures. Medical researchers working with children in Papua New Guinea asked whether recent infections reduce the level of serum retinol. They classified children as recently infected or not on the basis of other blood tests and then measured serum retinol. Of the 90 children in the sample, 55 had been recently infected. Table 16.3 gives the serum retinol levels for both groups, in micromoles per liter.<sup>9</sup>  **RETINOL**

(a) The researchers are interested in the proportional reduction in serum retinol. Verify that the mean for infected children is 0.620 and that the mean for uninfected children is 0.778.

Not infected						Infected					
0.59	1.08	0.88	0.62	0.46	0.39	0.68	0.56	1.19	0.41	0.84	0.37
1.44	1.04	0.67	0.86	0.90	0.70	0.38	0.34	0.97	1.20	0.35	0.87
0.35	0.99	1.22	1.15	1.13	0.67	0.30	1.15	0.38	0.34	0.33	0.26
0.99	0.35	0.94	1.00	1.02	1.11	0.82	0.81	0.56	1.13	1.90	0.42
0.83	0.35	0.67	0.31	0.58	1.36	0.78	0.68	0.69	1.09	1.06	1.23
1.17	0.35	0.23	0.34	0.49		0.69	0.57	0.82	0.59	0.24	0.41
						0.36	0.36	0.39	0.97	0.40	0.40
						0.24	0.67	0.40	0.55	0.67	0.52
						0.23	0.33	0.38	0.33	0.31	0.35
						0.82					

(b) There is no standard test for the null hypothesis that the *ratio of the population means* is 1. We can do a permutation test on the ratio of sample means. Carry out a one-sided test and report the  $P$ -value. Briefly describe the center and shape of the permutation distribution. Why do you expect the center to be close to 1?

## 16.77 Methods of resampling.

In Exercise 16.76, we did a permutation test for the hypothesis “no difference between infected and uninfected children” using the ratio of mean serum retinol levels to measure “difference.” We might also want a bootstrap confidence interval for the ratio of population means for infected and uninfected children. Describe carefully how resampling is done for the permutation test and for the bootstrap, paying attention to the difference between the two resampling methods.  **RETINOL**

## 16.78 Podcast downloads.

A 2006 Pew survey of Internet users asked whether or not they had downloaded a podcast at least once. The survey was repeated with different users in 2008. For the 2006 survey, 198 of the 2822 Internet users reported that they had downloaded at least one podcast. In the 2008 survey, the results were 295 of 1553 users. We want to use these sample data to test equality of the population proportions of successes. Carry

out a permutation test. Describe the permutation distribution. Give the  $P$ -value and report your conclusion.

### 16.79 Gender and GPA.

In Exercise 16.51 (page 16-41) we used the bootstrap to compare the mean GPA scores for men and women.  GPA

- Use permutation methods to compare the means for men and women.
- Use permutation methods to compare the standard deviations for men and women.
- Write a short paragraph summarizing your results and conclusions.

### 16.80 Sadness and spending.

A study of sadness and spending randomized subjects to watch videos designed to produce sad or neutral moods. Each subject was given \$10, and after watching the video, he or she was asked to trade \$0.50 increments of their \$10 for an insulated bottle of water. Here are the data:  SADNESS

Group	Purchase price (\$)										
Neutral	0.00	2.00	0.00	1.00	0.50	0.00	0.50	2.00	1.00	0.00	0.00
	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
Sad	3.00	4.00	0.50	1.00	2.50	2.00	1.50	0.00	1.00	1.50	1.50
	2.50	4.00	3.00	3.50	1.00	3.50	1.00	3.50	1.00	3.50	

- Use the two-sample  $t$  significance test (page 454) to compare the means of the two groups. Summarize your results.
- Use the pooled two-sample  $t$  significance test (page 462) to compare the means of the two groups. Summarize your results.
- Use a permutation test to compare the two groups. Summarize your results.
- Discuss the differences among the results you found for parts (a), (b), and (c). Which method do you prefer? Give reasons for your answer.

### 16.81 Comparing the variances for sadness and spending.

Refer to the previous example. Some treatments in randomized experiments such as this can cause variances to be different. Are the variances of the neutral and sad subjects equal?  SADNESS

- Use the  $F$  test for equality of variances (page 474) to answer this question. Summarize your results.
- Compare the variances using a permutation test. Summarize your results.
- Write a short paragraph comparing the  $F$  test with the permutation test for these data.

### 16.82 Comparing two operators.

Exercise 7.43 (page 445) gives these data on a delicate measurement of total body bone mineral content made by two operators on the same eight subjects:<sup>10</sup>  OPERAT

Operator	Subject							
	1	2	3	4	5	6	7	8
1	1.328	1.342	1.075	1.228	0.939	1.004	1.178	1.286
2	1.323	1.322	1.073	1.233	0.934	1.019	1.184	1.304

Do permutation tests give good evidence that measurements made by the two operators differ systematically? If so, in what way do they differ? Do two tests, one that compares centers and one that compares spreads.

## CHAPTER 16 Exercises

### 16.83 Gender and GPA.

In Example 16.5 (page 16-16) you used the bootstrap to find a 95% confidence interval for the 25% trimmed mean of GPA. Let's change the statistic of interest to the 5% trimmed mean. Using Example 16.5 as a guide, find the corresponding 95% confidence interval. Compare this interval with the one in Example 16.5.  GPA

### 16.84 Change the trim.

Refer to the previous exercise. Change the statistic of interest to the 10% trimmed mean. Answer the questions in the previous exercise and also compare your new interval with the one you found there.  GPA

### 16.85 Compare the correlations.

In Exercise 16.51 (page 16-41) we compared the mean GPA for men and women using the bootstrap. In Exercise 16.52 we used the bootstrap to examine the correlation between GPA and high school math grades. Let's find the correlations for men and women separately and ask whether there is evidence that they differ.  GPA

- Find the correlation between GPA and high school math grades for the men. Use the bootstrap to find a 95% confidence interval for the population correlation.
- Repeat part (a) for the women.
- Use the bootstrap to test the null hypothesis that the population correlations for men and women are the same,  $\rho_{\text{Men}} = \rho_{\text{Women}}$ .
- Summarize your findings.

### 16.86 Use the regression slope.

Refer to the previous exercise, where we used correlations to address the question of whether or not the relationship between GPA and high school math grades is the same for men and women. In Exercise 16.56 (page 16-42) we used the bootstrap to examine the slope of the least-squares regression line for predicting GPA using high school math grades. Let's compute the slope separately for men and women and ask whether or not they differ. This is another way to ask the question about whether or not the relationship between GPA and high school math grades is the same for men and women. Answer the questions from the previous exercise using the slope.

Compare the results that you find here with those you found in the previous exercise.  GPA

### 16.87 Bootstrap confidence interval for the difference in proportions.

Refer to Exercise 16.78 (page 16-55). We want a 95% confidence interval for the change from 2006 to 2008 in the proportions of Internet users who report that they have downloaded a podcast at least once. Bootstrap the sample data. Give all three bootstrap confidence intervals ( $t$ , percentile, and BCa). Compare the three intervals and summarize the results. Which intervals would you recommend? Give reasons for your answer.

### **16.88 Bootstrap confidence interval for the ratio.**

Here is one conclusion from the data in Table 16.3, described in Exercise 16.76: “The mean serum retinol level in uninfected children was 1.255 times the mean level in the infected children. A 95% confidence interval for the ratio of means in the population of all children in Papua New Guinea is . . .”

 **RETINOL**

- (a) Bootstrap the data and use the BCa method to complete this conclusion.
- (b) Briefly describe the shape and bias of the bootstrap distribution. Does the bootstrap percentile interval agree closely with the BCa interval for these data?

### **16.89 Poetry: an occupational hazard.**

According to William Butler Yeats, “She is the Gaelic muse, for she gives inspiration to those she persecutes. The Gaelic poets die young, for she is restless, and will not let them remain long on earth.” One study designed to investigate this issue examined the age at death for writers from different cultures and genders.<sup>11</sup>

In Example 1.32 (page 41) we examined the distributions of the age at death for female novelists, poets, and nonfiction writers. Figure 1.17 shows modified side-by-side boxplots for the three categories of writers. The poets do appear to die young! Note that there is an outlier among the nonfiction writers. This writer died at the age of 40, young for a nonfiction writer, but not for a novelist or a poet! Let’s use the methods of this chapter to compare the ages at death for poets and nonfiction writers.  **POETS**

- (a) Use numerical and graphical summaries to describe the distribution of age at death for the poets. Do the same for the nonfiction writers.
- (b) Use the methods of Chapter 7 (page 454) to compare the means of the two distributions. Summarize your findings.
- (c) Use the bootstrap methods of this chapter to compare the means of the two distributions. Summarize your findings.

### **16.90 Medians for the poets.**

Refer to the previous exercise. Use the bootstrap methods of this chapter to compare the medians of the two distributions. Summarize your findings and compare them with what you found in part (c) of the previous exercise.  **POETS**

### **16.91 Permutation test for the poets.**

Refer to Exercise 16.89. Answer part (c) of that exercise using the permutation test. Summarize your findings and compare them with what you found in Exercise 16.89.  **POETS**

## 16.92 Variance for poets.

Refer to Exercises 16.89 and 16.91.

- (a) Instead of comparing means, compare variances. Summarize your findings.
- (b) Explain how questions about the equality of standard deviations are related to questions about the equality of variances.
- (c) Use the results of this exercise and the previous three exercises to address the question of whether or not the distributions of the poets and nonfiction writers are the same.  **POETS**

## 16.93 Bootstrap confidence interval for the median.

Your software can generate random numbers that have the uniform distribution on 0 to 1. Figure 4.9 (page 258) shows the density curve. Generate a sample of 50 observations from this distribution.

- (a) What is the population median? Bootstrap the sample median and describe the bootstrap distribution.
- (b) What is the bootstrap standard error? Compute a 95% bootstrap  $t$  confidence interval.
- (c) Find the 95% BCa confidence interval. Compare with the interval in (b). Is the bootstrap  $t$  interval reliable here?

## 16.94 Are female personal trainers, on average, younger?

A fitness center employs 20 personal trainers. Here are the ages in years of the female and male personal trainers working at this center:  **TRAIN**

Male	25	26	23	32	35	29	30	28	31	32	29
Female	21	23	22	23	20	29	24	19	22		

- (a) Make a back-to-back stemplot. Do you think the difference in mean ages will be significant?
- (b) A two-sample  $t$  test gives  $P < 0.001$  for the null hypothesis that the mean age of female personal trainers is equal to the mean age of male personal trainers. Do a two-sided permutation test to check the answer.
- (c) What do you conclude about using the  $t$  test? What do you conclude about the mean ages of the trainers?

## 16.95 Adult gamers versus teen gamers.

A Pew survey compared adult and teen gamers on where they played games. For the adults, 54% of 1063 survey participants played on game consoles such as Xbox, PlayStation, and Wii. For teens, 89% of 1064 survey participants played on game consoles. Use the bootstrap to find a 95% confidence interval for the difference between the teen proportion who play on consoles and the adult proportion.

## 16.96 Use a ratio for adult gamers versus teen gamers.

Refer to the previous exercise. In many settings, researchers prefer to communicate the comparison of two proportions with a ratio. For gamers who play on consoles, they would report that teens are 1.65 (89/54) times more likely to play on consoles. Use the bootstrap to give a 95% confidence interval for this ratio.

### 16.97 Another way to communicate the result.

Refer to the previous two exercises. Here is another way to communicate the result: teen gamers are 65% more likely to play on consoles than adult gamers.

- Explain how the 65% is computed.
- Use the bootstrap to give a 95% confidence interval for this estimate.
- Based on this exercise and the previous two, which of the three ways is most effective for communicating the results? Give reasons for your answer.

### 16.98 Insurance fraud?

Jocko's Garage has been accused of insurance fraud. Data on estimates (in dollars) made by Jocko and another garage were obtained for 10 damaged vehicles. Here is what the investigators found:

 GARAGE

Car	1	2	3	4	5
Jocko's	1375	1550	1250	1300	900
Other	1250	1300	1250	1200	950
Car	6	7	8	9	10
Jocko's	1500	1750	3600	2250	2800
Other	1575	1600	3300	2125	2600

- Compute the mean estimate for Jocko and the mean estimate for the other garage. Report the difference in the means and the 95% standard  $t$  confidence interval. Be sure to choose the appropriate  $t$  procedure for your analysis and explain why you made this choice.
- Use the bootstrap to find the confidence interval. Be sure to give details about how you used the bootstrap, which options you chose, and why.
- Compare the  $t$  interval with the bootstrap interval.

### 16.99 Other ways to look at Jocko's estimates.

Refer to the previous exercise. Let's consider some other ways to analyze these data.  GARAGE

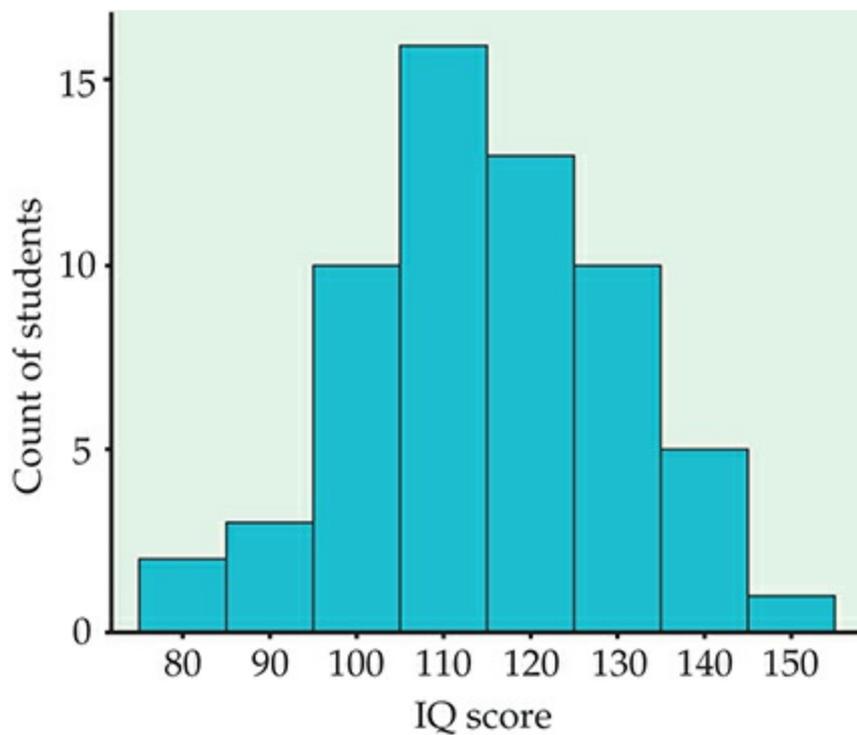
- For each damaged vehicle, divide Jocko's estimate by the estimate from the other garage. Perform your analysis on these data. Write a short report that includes numerical and graphical summaries, your estimate, the 95%  $t$  confidence interval, the 95% bootstrap confidence interval, and an explanation for all choices (such as whether you chose to examine the mean or the median, bootstrap options, etc.).
- Compute the mean of Jocko's estimates and the mean of the estimates made by the other garage. Divide Jocko's mean by the mean for the other garage. Report this ratio and find a 95% confidence interval for this quantity. Be sure to justify choices that you made for the bootstrap.

(c) Using what you have learned in this exercise and the previous one, how would you summarize the comparison of Jocko's estimates with those made by the other garage? Assume that your audience knows very little about statistics but a lot about insurance.

## CHAPTER 16 Notes and Data Sources

1. Information about this free software is available at [r-project.org](http://r-project.org).
2. The origin of this quaint phrase is Rudolph Raspe, *The Singular Adventures of Baron Munchausen*, 1786. Here is the passage, from the edition by John Carswell, Heritage Press, 1952: “I was still a couple of miles above the clouds when it broke, and with such violence I fell to the ground that I found myself stunned, and in a hole nine fathoms under the grass, when I recovered, hardly knowing how to get out again. Looking down, I observed that I had on a pair of boots with exceptionally sturdy straps. Grasping them firmly, I pulled with all my might. Soon I had hoist myself to the top and stepped out on terra firma without further ado.”
3. In fact, the bootstrap standard error underestimates the true standard error. Bootstrap standard errors are generally too small by a factor of roughly  $1-1/n$ . This factor is about 0.95 for  $n = 10$  and 0.98 for  $n = 25$ , so we ignore it in this elementary exposition.
4. The 254 winning numbers and their payoffs are republished here by permission of the New Jersey State Lottery Commission.
5. The vehicle is a 2002 Toyota Prius owned by the third author.
6. The standard advanced introduction to bootstrap methods is B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall, 1993. For tilting intervals, see B. Efron, “Nonparametric standard errors and confidence intervals” (with discussion), *Canadian Journal of Statistics*, 36 (1981), pp. 369–401; and T. J. DiCiccio and J. P. Romano, “Nonparametric confidence limits by resampling methods and least favourable families,” *International Statistical Review*, 58 (1990), pp. 59–76.
7. This example is adapted from Maribeth C. Schmitt, “The effects of an elaborated directed reading activity on the metacomprehension skills of third graders,” PhD dissertation, Purdue University, 1987.
8. These data were collected as part of a larger study of dementia patients conducted by Nancy Edwards, School of Nursing, and Alan Beck, School of Veterinary Medicine, Purdue University.
9. Data provided by Francisco Rosales of the Department of Nutritional Sciences, Pennsylvania State University. See Francisco Rosales et al., “Relation of serum retinol to acute phase proteins and malarial morbidity in Papua New Guinea children,” *American Journal of Clinical Nutrition*, 71 (2000), pp. 1580–1588.
10. These data were collected in connection with a bone health study at Purdue University and were provided by Linda McCabe.
11. The data were provided by James Kaufman. The study is described in James C. Kaufman, “The cost of the muse: poets die young,” *Death Studies*, 27 (2003), pp. 813–821. The quote from Yeats appears in this article.

# 17 Statistics for Quality: Control CHAPTER and Capability



**17.1 Processes and Statistical Process Control**

**17.2 Using Control Charts**

**17.3 Process Capability Indexes**

**17.4 Control Charts for Sample Proportions**

## Introduction

Quality is a broad concept. Often it refers to a degree or grade of excellence. For example, you may feel that a restaurant serving filet mignon is a higher-quality establishment than a fast-food outlet that primarily serves hamburgers. You may also consider a name-brand sweater of higher quality than one sold at a discount store.

In this chapter, we consider a narrower concept of quality: *consistently meeting standards appropriate for a specific product or service*. The fast-food outlet, for example, may serve high-quality hamburgers. The hamburgers are freshly grilled and served promptly at the right temperature every time you visit. Similarly, the discount store sweaters may be high quality because they are consistently free of defects and the tight knit helps them keep their shape wash after wash.

Statistically minded management can assess this concept of quality through sampling. For example, the fast-food outlet could sample hamburgers and measure the time from order to being served as well as the temperature and tenderness of the burgers. This chapter discusses the methods used to monitor the quality of a product or service and effectively detect changes in the process that may affect its quality.

### Use of data to assess quality

Organizations are (or ought to be) concerned about the quality of the products and services they offer. What they don't know about quality can hurt them: rather than make complaints that an alert organization could use as warnings, customers often simply leave when they feel they are receiving poor quality. A key to maintaining and improving quality is systematic use of *data* in place of intuition or anecdotes. Here are two examples.

#### EXAMPLE

##### 17.1 Membership renewal process.

Sometimes data that are routinely produced make a quality problem obvious. The internal financial statements of a professional society showed that hiring temporary employees to enter membership data was causing expenditures above budgeted levels each year during the several months when memberships

were renewed. Investigation led to two actions. Membership renewal dates were staggered across the year to spread the workload more evenly. More important, outdated and inflexible data entry software was replaced by a modern system that was much easier to use. Result: permanent employees could now process renewals quickly, eliminating the need for temps and also reducing member complaints.

## EXAMPLE

### 17.2 Response time process.

Systematic collection of data helps an organization to move beyond dealing with obvious problems. Motorola measures the performance of its services and manufactured products. They track, for example, the average time from a customer's call until the problem is fixed, month by month. The trend should be steadily downward as ways are found to speed response.

 **LOOK BACK**

**time plot, p. 23**

 **LOOK BACK**

**regression line, p. 110**

 **LOOK BACK**

**comparative experiments, p. 178**

Because using data is a key to improving quality, statistical methods have much to contribute. Simple tools are often the most effective. Motorola's service centers calculate mean response times each month and make a time plot. A scatterplot and perhaps a regression line can show how the time to answer telephone calls to a corporate call center influences the percent of callers who hang up before their calls are answered. The design of a new product such as a smartphone may involve interviewing samples of consumers to learn what features they want included and using randomized comparative experiments to determine the best interface.

 **LOOK BACK**

**sampling distributions, p. 208**

This chapter focuses on just one aspect of statistics for improving quality: *statistical process control*. The techniques are simple and are based on sampling distributions, but the underlying ideas are important and a bit subtle.

## 17.1 Processes and Statistical Process Control

When you complete this section, you will be able to

- Describe a process using a flowchart and a cause-and-effect diagram.
- Explain what is meant by a process being in control by distinguishing common and special cause variation.
- Compute the center line and control limits for an  $\bar{x}$  chart and utilize the chart for process monitoring.
- Compute the center line and control limits for an  $s$  chart and utilize the chart for process monitoring.
- Contrast the  $\bar{x}$  and  $s$  charts in terms of what they monitor and which should be interpreted first.

In thinking about statistical inference, we distinguish between the *sample* data we have in hand and the wider *population* that the data represent. We hope to use the sample to draw conclusions about the population. In thinking about quality improvement, it is often more natural to speak of *processes* rather than populations. This is because work is organized in processes. Here are some examples:

- Processing an application for admission to a university and deciding whether or not to admit the student.
- Reviewing an employee's expense report for a business trip and issuing a reimbursement check.
- Hot forging to shape a billet of titanium into a blank that, after machining, will become part of a medical implant for hip, knee, or shoulder replacement.

Each of these processes is made up of several successive operations that eventually produce the output—an admission decision, a reimbursement check, or a metal component.

### PROCESS

A **process** is a chain of activities that turns inputs into outputs.

We can accommodate processes in our sample-versus-population framework: think of the population as containing all the outputs that would be produced by the process if it ran forever in its present state. The outputs produced today or this

week are a sample from this population. Because the population doesn't actually exist now, it is simpler to speak of a process and of recent output as a sample from the process in its present state.

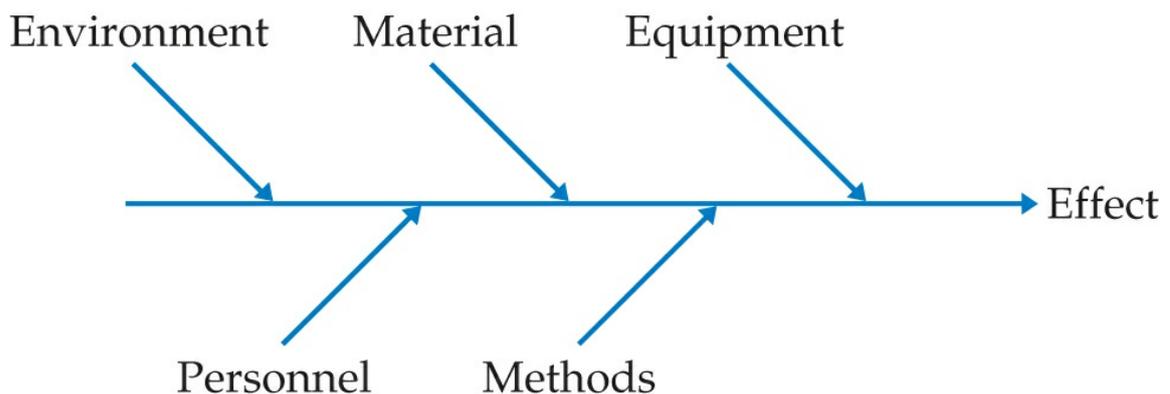
## Describing processes

The first step in improving a process is to understand it. If the process is at all complex, even the people involved with it may not have a full picture of how the activities interact in ways that influence quality. A brainstorming session is in order: bring people together to gain an understanding of the process.

This understanding is often presented graphically using two simple tools: flowcharts and cause-and-effect diagrams. A **flowchart** is a picture of the stages of a process. Many organizations have formal standards for making flowcharts. Because flowcharts are not statistical graphics, we will informally illustrate their use in an example and not insist on a specific format. A **cause-and-effect diagram** organizes the logical relationships between the inputs and stages of a process and an output. Sometimes the output is successful completion of the process task; sometimes it is a quality problem that we hope to solve. A good starting outline for a cause-and-effect diagram appears in Figure 17.1. The main branches organize the causes and serve as a skeleton for detailed entries. You can see why these are sometimes called "fishbone diagrams." Once again we will illustrate the diagram by example rather than insist on a specific format.<sup>1</sup>

flowchart

cause-and-effect diagram



**FIGURE 17.1**

An outline for a cause-and-effect diagram. Group causes under these main headings in the form of branches.

## EXAMPLE

### 17.3 Flowchart and cause-and-effect diagram of a hot-forging process.



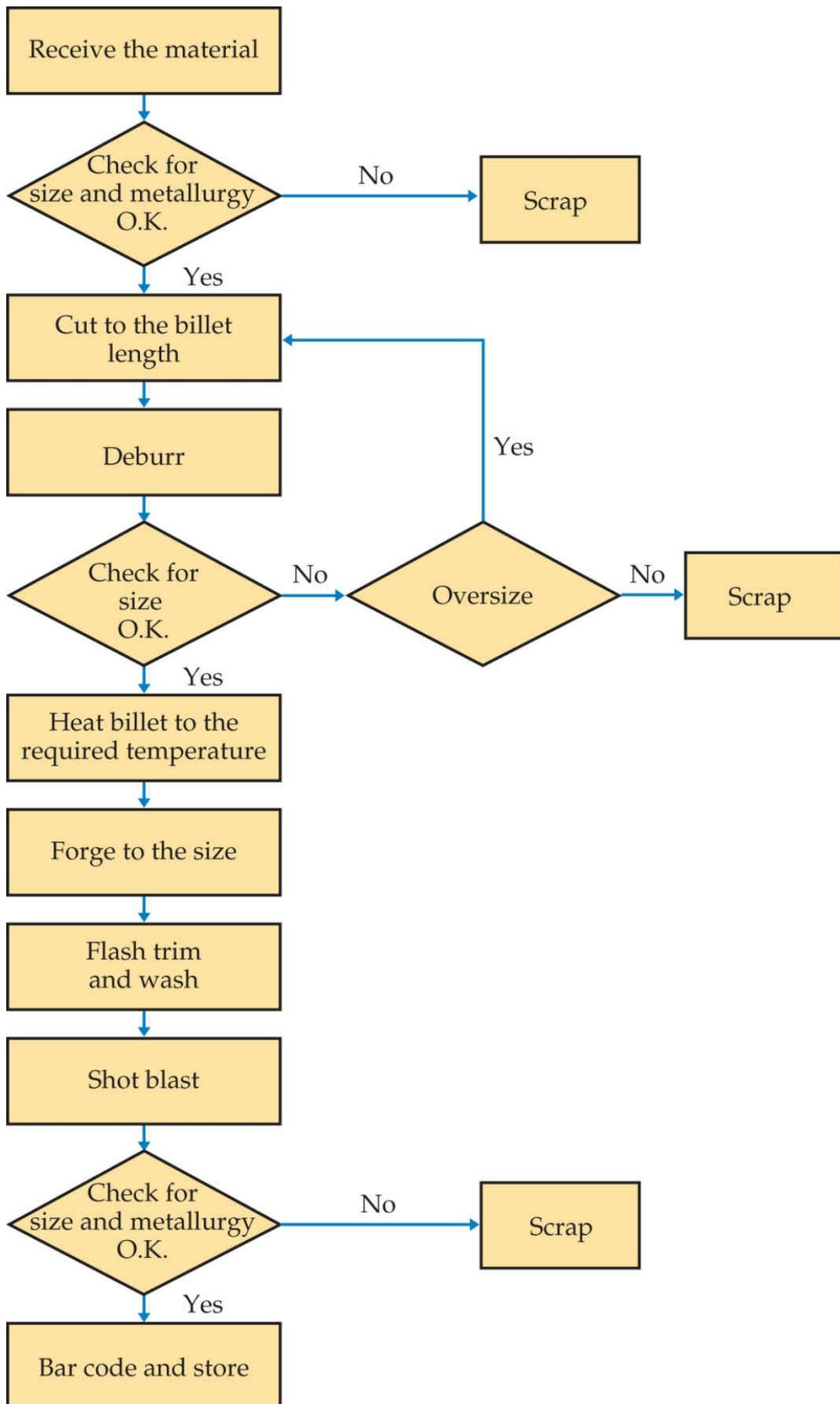
Hot forging involves heating metal to a plastic state and then shaping it by applying thousands of pounds of pressure to force the metal into a die (a kind of mold). Figure 17.2 is a flowchart of a typical hot-forging process.<sup>2</sup>

A process improvement team, after making and discussing this flowchart, came to several conclusions:

- Inspecting the billets of metal received from the supplier adds no value. Insist that the supplier be responsible for the quality of the material. This then eliminates the inspection step.
- If possible, buy the metal billets already cut to rough length and deburred by the supplier. This would eliminate the cost of preparing the raw material.
- Heating the metal billet and forging (pressing the hot metal into the die) are the heart of the process. The company should concentrate attention here.

The team then prepared a cause-and-effect diagram (Figure 17.3) for the heating and forging part of the process. The team members shared their specialist knowledge of the causes in their area, resulting in a more complete picture than any one person could produce. Figure 17.3 is a simplified version of the actual diagram. We have given some added detail for the “hammer stroke” branch under “equipment” to illustrate the next level of branches. Even this requires some knowledge of hot forging to understand. Based on detailed

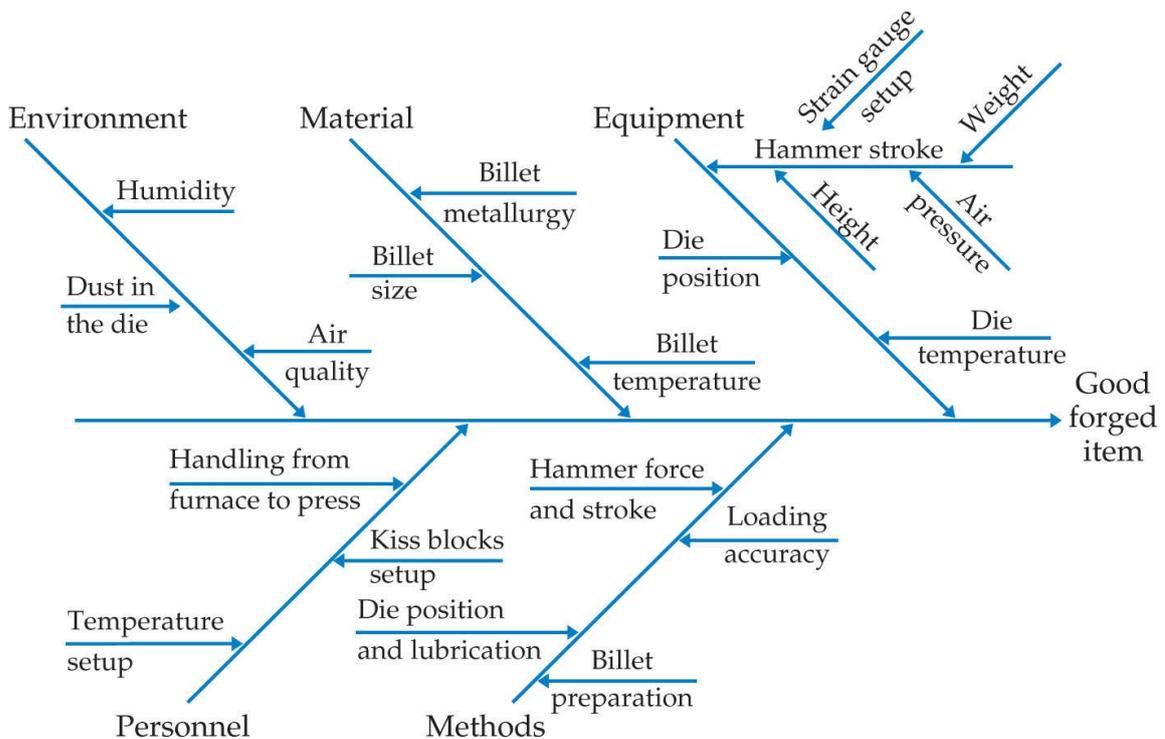
discussion of the diagram, the team decided what variables to measure and at what stages of the process to measure them. Producing well-chosen data is the key to improving the process.



**FIGURE 17.2**

Flowchart of the hot-forging process in Example 17.3. Use this as a model for flowcharts: decision points appear as diamonds, and other steps in the process appear as rectangles. Arrows represent flow from step to step.

We will apply statistical methods to a series of measurements made on a process. Deciding what specific variables to measure is an important step in quality improvement. Often we use a “performance measure” that describes an output of a process. A company’s financial office might record the percent of errors that outside auditors find in expense account reports or the number of data entry errors per week. The personnel department may measure the time to process employee insurance claims or the percent of job offers that are accepted. In the case of complex processes, it is wise to measure key steps within the process rather than just final outputs. The process team in Example 17.3 might recommend that the temperature of the die and of the billet be measured just before forging.



**FIGURE 17.3**

Simplified cause-and-effect diagram of the hot-forging process in Example 17.3. Good cause-and-effect diagram require detailed knowledge of the specific process.

**USE YOUR KNOWLEDGE**

**17.1 Describing your process.**

Choose a process that you know well, preferably from a job you have held. If you lack experience with actual business processes, choose a personal process such as making macaroni and cheese or brushing your teeth. Make a flowchart of the process. Make a cause-and-effect diagram that presents the factors that lead to successful completion of the process.

### 17.2 What variables to measure?

Based on your description of the process in Exercise 17.1, suggest specific variables that you might measure in order to

- (a) assess the overall quality of the process.
- (b) gather information on a key step within the process.

## Statistical process control

The goal of statistical process control is to make a process stable over time and then keep it stable unless planned changes are made. You might want, for example, to keep your weight constant over time. A manufacturer of machine parts wants the critical dimensions to be the same for all parts. “Constant over time” and “the same for all” are not realistic requirements. They ignore the fact that *all processes have variation*. Your weight fluctuates from day to day; the critical dimension of a machined part varies a bit from item to item; the time to process a college admission application is not the same for all applications. Variation occurs in even the most precisely made product due to small changes in the raw material, the behavior of the machine or operator, and even the temperature in the plant.

Because variation is always present, we can’t expect to hold a variable exactly constant over time. The statistical description of stability over time requires that the *pattern of variation* remain stable, not that there be no variation in the variable measured.

In the language of statistical quality control, a process that is in control has only **common cause** variation. Common cause variation is the inherent variability of the process, due to many small causes that are always present. When the normal functioning of the process is disturbed by some unpredictable event, **special cause** variation is added to the common cause variation. We hope to be able to discover what lies behind special cause variation and eliminate that cause to restore the stable functioning of the process.

common cause

special cause

## EXAMPLE

### 17.4 Common and special cause variation.

Imagine yourself doing the same task repeatedly, say folding a circular, stuffing it into a stamped envelope, and sealing the envelope. The time to complete this task will vary a bit, and it is hard to point to any one reason for the variation. Your completion time shows only common cause variation.

Now the telephone rings. You answer, and though you continue folding and stuffing while talking, your completion time rises beyond the level expected from common causes alone. Answering the telephone adds special cause variation to the common cause variation that is always present. The process has been disturbed and is no longer in its normal and stable state.

Control charts work by distinguishing the always-present common cause variation in a process from the additional variation that suggests that the process has been disturbed by a special cause. A control chart sounds an alarm when it sees too much variation. This is accomplished through a combination of graphical and numerical descriptions of data with use of sampling distributions.



**sampling distributions, p. 302**

Control charts were invented in the 1920s by Walter Shewhart at the Bell Telephone Laboratories.<sup>3</sup> The most common application of control charts is to monitor the performance of industrial and business processes. The same methods, however, can be used to check the stability of quantities as varied as the ratings of a television show, the level of ozone in the atmosphere, and the gas mileage of your car.

## STATISTICAL CONTROL

A variable that continues to be described by the same distribution when observed over time is said to be in statistical control, or simply **in control**.

**Control charts** are statistical tools that monitor a process and alert us when the process has been disturbed so that it is now **out of control**. This is a signal to find and correct the cause of the disturbance.

## USE YOUR KNOWLEDGE

### 17.3 Considering common and special cause variation.



In Exercise 17.1 (page 17-6), you described a process that you know well. What are some sources of common cause variation in this process? What are some special causes that might, at times, drive the process out of control?

### 17.4 Examples of special cause variation in arrival times.

Lex takes a 7:45 A.M. shuttle to campus each morning. Her apartment complex is near a major road and is two miles from campus. Her arrival time to campus varies a bit from day to day but is generally stable. Give several examples of special causes that might raise Lex's arrival time on a particular day.

## $\bar{x}$ charts for process monitoring

When you first apply control charts to a process, the process may not be in control. Even if it is in control, you don't yet understand its behavior. You will have to collect data from the process, establish control by uncovering and removing special causes, and then set up control charts to maintain control. We call this the **chart**

setup stage.

### chart setup

Later, when the process has been operating in control for some time, you understand its usual behavior and have a long run of data from the process. You keep control charts to monitor the process because a special cause could erupt at any time. We will call this **process monitoring**.<sup>4</sup>

### process monitoring

Although in practice chart setup precedes process monitoring, the big ideas of control charts are more easily understood in the process-monitoring setting. We will start there and then discuss the more complex process improvement setting.

Consider a quantitative variable  $x$  that is an important measure of quality. The variable might be the diameter of a part, the number of envelopes stuffed in an hour, or the time to respond to a customer call. If this process is in control, the variable  $x$  is described by the same distribution over time. For now, we'll assume this distribution is Normal.

## PROCESS-MONITORING CONDITIONS

The measured quantitative variable  $x$  has a **Normal distribution**. The process has been operating in control for a long period, so that we know the **process mean  $\mu$**  and the **process standard deviation  $\sigma$**  that describe the distribution of  $x$  as long as the process remains in control.

In practice, we must estimate the process mean and standard deviation from past data on the process. Under the process-monitoring conditions, we have numerous observations and the process has remained in control. The law of large numbers tells us that estimates from past data will be very close to the truth about the process. That is, at the process-monitoring stage we can act as if we know the true values of  $\mu$  and  $\sigma$ .



### law of large numbers, p. 268

Note carefully that  $\mu$  and  $\sigma$  describe the center and spread of our variable  $x$  *only as long as the process remains in control*. A special cause may at any time disturb the process and change the mean, the standard deviation, or both.

To make control charts, begin by taking small samples from the process at regular intervals. For example, we might measure 4 or 5 consecutive parts or the

response times to 4 or 5 consecutive customer calls. There is an important idea here: *the observations in a sample are so close together in time that we can assume that the process is stable during this short period.* Variation within a single sample gives us a benchmark for the common cause variation in the process.

*The process standard deviation  $\sigma$  refers to the standard deviation within the time period spanned by one sample.* If the process remains in control, the same  $\sigma$  describes the standard deviation of observations across any time period. Control charts help us decide whether this is the case.

We start with the  $\bar{x}$  **chart**, which is based on plotting the means of the successive samples. Here is the outline:

### $\bar{x}$ chart

1. **Take samples** of size  $n$  from the process at regular intervals. **Plot the means  $\bar{x}$**  of these samples against the order in which the samples were taken.



**sampling distribution of  $\bar{x}$ , p. 307.**

2. We know that the sampling distribution of  $\bar{x}$  under the process-monitoring conditions is Normal with mean  $\mu$  and standard deviation  $\sigma/n$ . Draw a solid **center line** on the chart at height  $\mu$ .

center line



**68–95–99.7 rule, p. 59**

3. The 99.7 part of the 68–95–99.7 rule for Normal distributions says that, as long as the process remains in control, 99.7% of the values of  $\bar{x}$  will fall between  $\mu - 3\sigma/n$  and  $\mu + 3\sigma/n$ . Draw dashed **control limits** on the chart at these heights. The control limits mark off the range of variation in sample means that we expect to see when the process remains in control.

control limits

If the process remains in control and the process mean and standard deviation do not change, we will rarely observe an  $\bar{x}$  outside the control limits. Such an  $\bar{x}$  would be a signal that the process has been disturbed.

**EXAMPLE**

**17.5 Monitoring the water resistance of fabric.**



A manufacturer of outdoor sportswear must control the water resistance and breathability of their jackets. Water resistance is measured by the amount of water (depth in millimeters) that can be suspended above the fabric before water seeps through. For their jackets, this test is done along the seams and zipper, where the resistance is likely the weakest. For one particular style of jacket, the manufacturing process has been stable with mean resistance  $\mu = 2750$  mm and process standard deviation  $\sigma = 430$  mm.

Each four-hour shift, an operator measures the resistance on a sample of 4 jackets. Table 17.1 gives the last 20 samples. The table also gives the mean  $\bar{x}$  and the standard deviation  $s$  for each sample. The operator did not have to calculate these—modern measuring equipment often comes equipped with software that automatically records  $\bar{x}$  and  $s$  and even produces control charts.

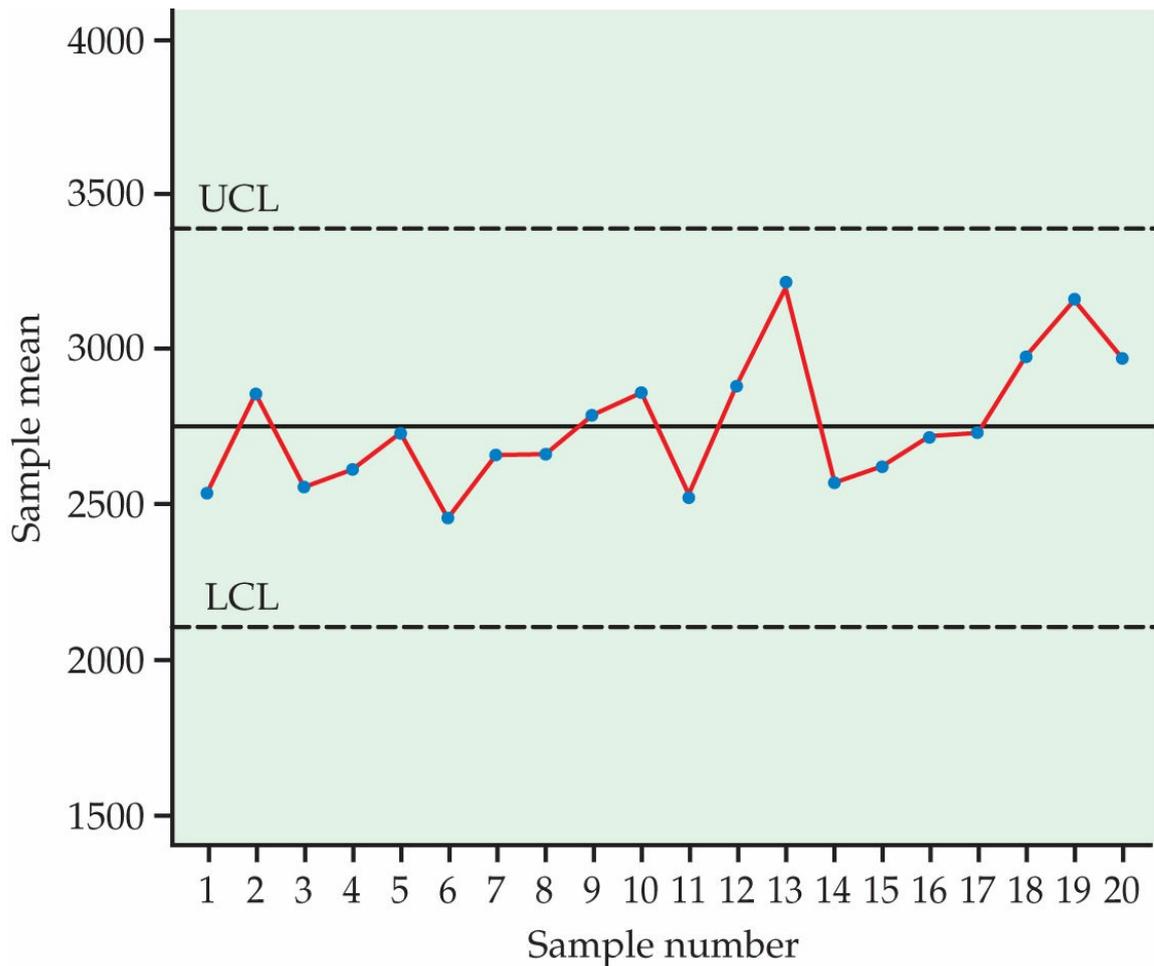
Figure 17.4 is an  $\bar{x}$  control chart for the 20 water resistance samples in Table

17.1. We have plotted each sample mean from the table against its sample number. For example, the mean of the first sample is 2534 mm, and this is the value plotted for Sample 1. The center line is at  $\mu = 2750$  mm. The upper and lower control limits are

**TABLE 17.1**

**Twenty Control Chart Samples of Water Resistance (depth in mm)**

Sample	Depth measurements				Sample mean	Standard deviation
1	2345	2723	2345	2723	2534	218
2	3111	3058	2385	2862	2854	330
3	2471	2053	2526	3161	2553	457
4	2154	2968	2742	2568	2608	344
5	3279	2472	2833	2326	2728	425
6	3043	2363	2018	2385	2452	428
7	2689	2762	2756	2402	2652	170
8	2821	2477	2598	2728	2656	150
9	2608	2599	2479	3453	2785	449
10	3293	2318	3072	2734	2854	425
11	2664	2497	2315	2652	2532	163
12	1688	3309	3336	3183	2879	797
13	3499	3342	2923	3015	3195	271
14	2352	2831	2459	2631	2568	210
15	2573	2184	2962	2752	2618	330
16	2351	2527	3006	2976	2715	327
17	2863	2938	2362	2753	2729	256
18	3281	2726	3297	2601	2976	365
19	3164	2874	3730	2860	3157	407
20	2968	3505	2806	2598	2969	388



**FIGURE 17.4**

The  $\bar{x}$  chart for the water resistance data of Table 17.1. No points lie outside the control limits.

$$\mu + 3\sigma = 2750 + 3(645) = 2750 + 1935 = 4685 \text{ mm (UCL)}$$

$$\mu - 3\sigma = 2750 - 3(645) = 2750 - 1935 = 815 \text{ mm (LCL)}$$

As is common, we have labeled the control limits UCL for upper control limit and LCL for lower control limit.

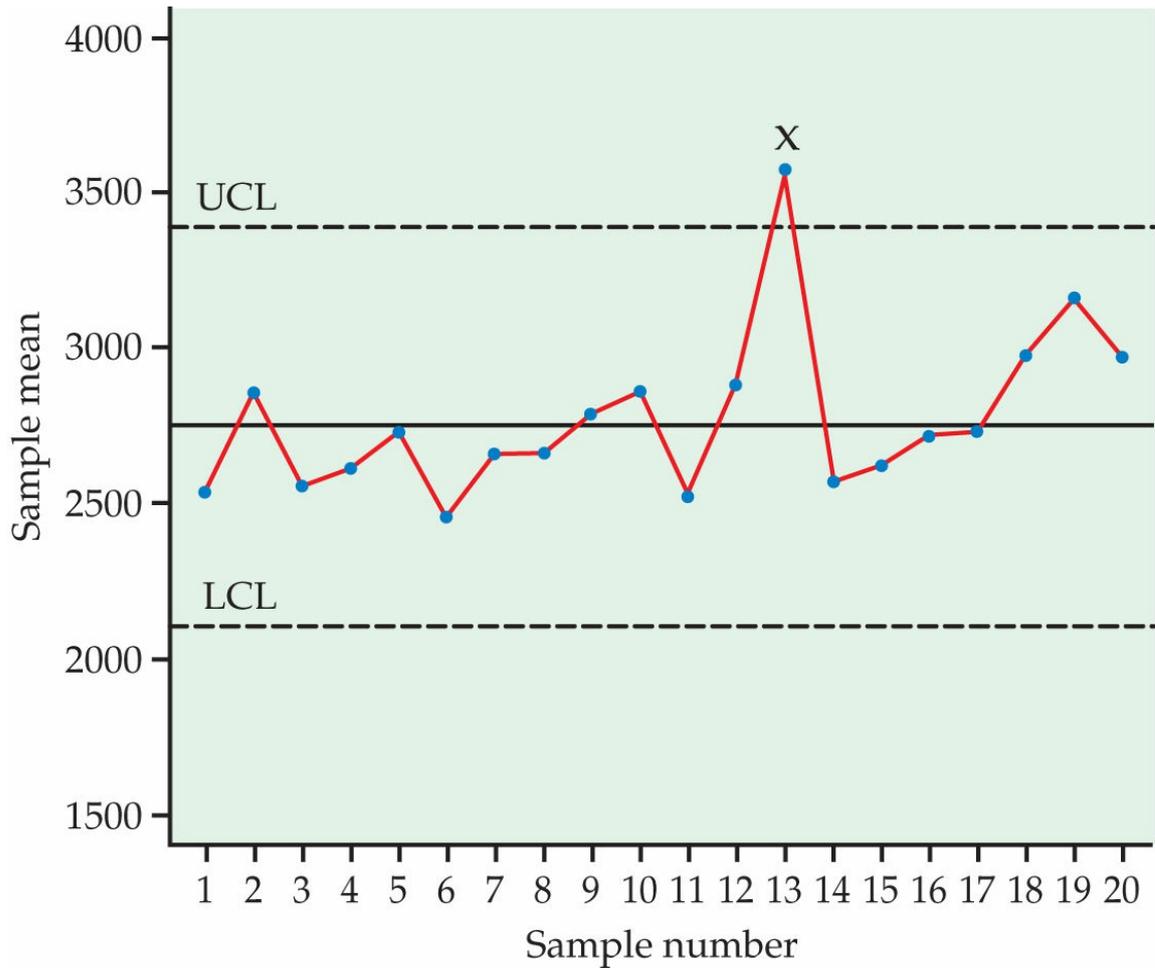
## EXAMPLE

### 17.6 Reading an $\bar{x}$ control chart.

Figure 17.4 is a typical  $\bar{x}$  chart for a process in control. The means of the 20 samples do vary, but all lie within the range of variation marked out by the control limits. We are seeing the common cause variation of a stable process.

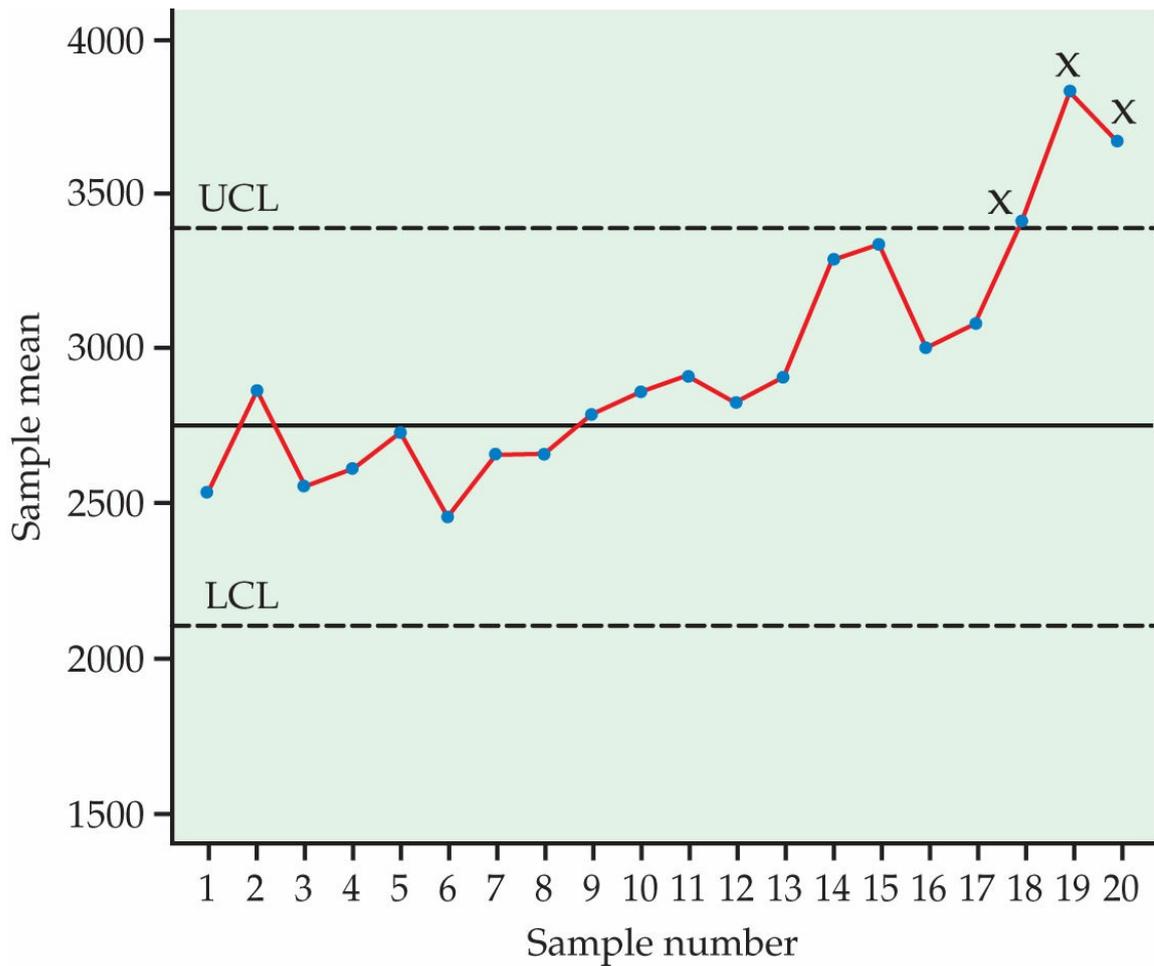
Figures 17.5 and 17.6 illustrate two ways in which the process can go out of control. In Figure 17.5, the process was disturbed by a special cause sometime between Sample 12 and Sample 13. As a result, the mean resistance for Sample 13 falls above the upper control limit. It is common practice to mark all out-of-control points with an “x” to call attention to them. A search for the cause begins as soon as we see a point out of control. Investigation finds that the seam sealer device has slipped, resulting in more sealer being applied. This is good for water resistance but harms the jacket’s breathability. When the problem is corrected, Samples 14 to 20 are again in control.

Figure 17.6 shows the effect of a steady upward drift in the process center, starting at Sample 11. You see that some time elapses before  $\bar{x}$  is out of control (Sample 18). The one-point-out rule works better for detecting sudden large disturbances than for detecting slow drifts in a process.



**FIGURE 17.5**

The  $\bar{x}$  chart is identical to that in Figure 17.4 except that a special cause has driven  $\bar{x}$  for Sample 13 above the upper control limit. The out-of-control point is marked with an x.



**FIGURE 17.6**

The first 10 points on this  $\bar{x}$  chart are as in Figure 17.4. The process mean drifts upward after Sample 10, and the sample means  $\bar{x}$  reflect this drift. The points for Samples 18, 19, and 20 are out of control.

## USE YOUR KNOWLEDGE

**17.5** An  $\bar{x}$  control chart for sandwich orders.



A sandwich shop owner takes a daily sample of five consecutive sandwich orders at a random time during the lunch rush and records the time it takes to complete each order. Past experience indicates that the process mean should be  $\mu = 90$  seconds and the process standard deviation should be  $\sigma = 24$  seconds. Calculate the center line and control limits for an  $\bar{x}$  control chart.

### 17.6 Changing the sample size $n$ or the unit of measure.

Refer to Exercise 17.5. What happens to the center line and control limits if

- the owner samples four consecutive sandwich orders?
- the owner samples six consecutive sandwich orders?
- the owner uses minutes rather than seconds as the units?

### **s** charts for process monitoring

The  $\bar{x}$  charts in Figures 17.4, 17.5, and 17.6 were easy to interpret because the process standard deviation remained fixed at 430 mm. The effects of moving the process mean away from its in-control value (2750 mm) are then clear to see. We know that even the simplest description of a distribution should give both a measure of center and a measure of spread. So it is with control charts. We must monitor both the process center, using an  $\bar{x}$  chart, and the process spread, using a control chart for the sample standard deviation  $s$ .

The standard deviation  $s$  does not have a Normal distribution, even approximately. Under the process-monitoring conditions, the sampling distribution of  $s$  is skewed to the right. Nonetheless, control charts for any statistic are based on the “plus or minus three standard deviations” idea motivated by the 68–95–99.7 rule for Normal distributions.

Control charts are intended to be practical tools that are easy to use. Standard practice in process control therefore ignores such details as the effect of non-Normal sampling distributions. Here is the general control chart setup for a sample statistic  $Q$  (short for “quality characteristic”).

### THREE-SIGMA CONTROL CHARTS

To make a **three-sigma ( $3\sigma$ ) control chart** for any statistic  $Q$ :

1. Take samples from the process at regular intervals and plot the values of the statistic  $Q$  against the order in which the samples were taken.
2. Draw a **center line** on the chart at height  $\mu_Q$ , the mean of the statistic when the process is in control.
3. Draw upper and lower **control limits** on the chart three standard deviations of  $Q$  above and below the mean. That is,  

$$\text{UCL} = \mu_Q + 3\sigma_Q$$

$$\text{LCL} = \mu_Q - 3\sigma_Q$$
 Here  $\sigma_Q$  is the standard deviation of the sampling distribution of the statistic  $Q$  when the process is in control.
4. The chart produces an **out-of-control signal** when a plotted point lies outside the control limits.

We have applied this general idea to  $\bar{x}$  charts. If  $\mu$  and  $\sigma$  are the process mean and standard deviation, the statistic  $\bar{x}$  has mean  $\mu_{\bar{x}} = \mu$  and standard deviation  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ . The center line and control limits for  $\bar{x}$  charts follow from these facts.

What are the corresponding facts for the sample standard deviation  $s$ ? Study of the sampling distribution of  $s$  for samples from a Normally distributed process characteristic gives these facts:

1. The *mean* of  $s$  is a constant times the process standard deviation  $\sigma$ , that is,  $\mu_s = c_4\sigma$ .
2. The *standard deviation* of  $s$  is also a constant times the process standard deviation,  $\sigma_s = c_5\sigma$ .

The constants are called  $c_4$  and  $c_5$  for historical reasons. Their values depend on the size of the samples. For large samples,  $c_4$  is close to 1. That is, the sample standard deviation  $s$  has little bias as an estimator of the process standard deviation  $\sigma$ . Because statistical process control often uses small samples, we pay attention to the value of  $c_4$ . Following the general pattern for three-sigma control charts,

1. The *center line* of an  $s$  chart is at  $c_4\sigma$ .
2. The *control limits* for an  $s$  chart are at

$$UCL = \mu_s + 3\sigma_s = c_4\sigma + 3c_5\sigma = (c_4 + 3c_5)\sigma = B_6\sigma$$

$$LCL = \mu_s - 3\sigma_s = c_4\sigma - 3c_5\sigma = (c_4 - 3c_5)\sigma = B_5\sigma$$

That is, the control limits UCL and LCL are also constants times the process standard deviation. These constants are called (again for historical reasons)  $B_6$  and  $B_5$ . We don't need to remember that  $B_6 = c_4 + 3c_5$  and  $B_5 = c_4 - 3c_5$ , because tables give us the numerical values of  $B_6$  and  $B_5$ .

## $\bar{x}$ AND $s$ CONTROL CHARTS FOR PROCESS MONITORING<sup>5</sup>

Take regular samples of size  $n$  from a process that has been in control with process mean  $\mu$  and process standard deviation  $\sigma$ . The center line and control limits for an  $\bar{x}$  **chart** are

$$UCL = \mu + 3\sigma$$

$$CL = \mu$$

$$LCL = \mu - 3\sigma$$

The center line and control limits for an  $s$  **chart** are

$$UCL = B_6\sigma$$

$$CL = c_4\sigma$$

$$LCL = B_5\sigma$$

The **control chart constants**  $c_4$ ,  $B_5$ , and  $B_6$  depend on the sample size  $n$ .

Table 17.2 gives the values of the control chart constants  $c_4$ ,  $c_5$ ,  $B_5$ , and  $B_6$  for samples of sizes 2 to 10. This table makes it easy to draw  $s$  charts. The table has no

$B_5$  entries for samples smaller than  $n = 6$ . The lower control limit for an  $s$  chart is zero for samples of sizes 2 to 5. This is a consequence of the fact that  $s$  has a right-skewed distribution and takes only values greater than zero. The point three standard deviations above the mean (UCL) lies on the long right side of the distribution. The point three standard deviations below the mean (LCL) on the short left side is below zero, so we say that  $LCL = 0$ .

**TABLE 17.2** Control Chart Constants

Sample size $n$	$c_4$	$c_5$	$B_5$	$B_6$
2	0.7979	0.6028		2.606
3	0.8862	0.4633		2.276
4	0.9213	0.3889		2.088
5	0.9400	0.3412		1.964
6	0.9515	0.3076	0.029	1.874
7	0.9594	0.2820	0.113	1.806
8	0.9650	0.2622	0.179	1.751
9	0.9693	0.2459	0.232	1.707
10	0.9727	0.2321	0.276	1.669

## EXAMPLE

### 17.7 Interpreting an $s$ chart for the waterproofing process.



H2ORES

Figure 17.7 is the  $s$  chart for the water resistance data in Table 17.1. The samples are of size  $n = 4$  and the process standard deviation in control is  $\sigma = 430$  mm. The center line is therefore

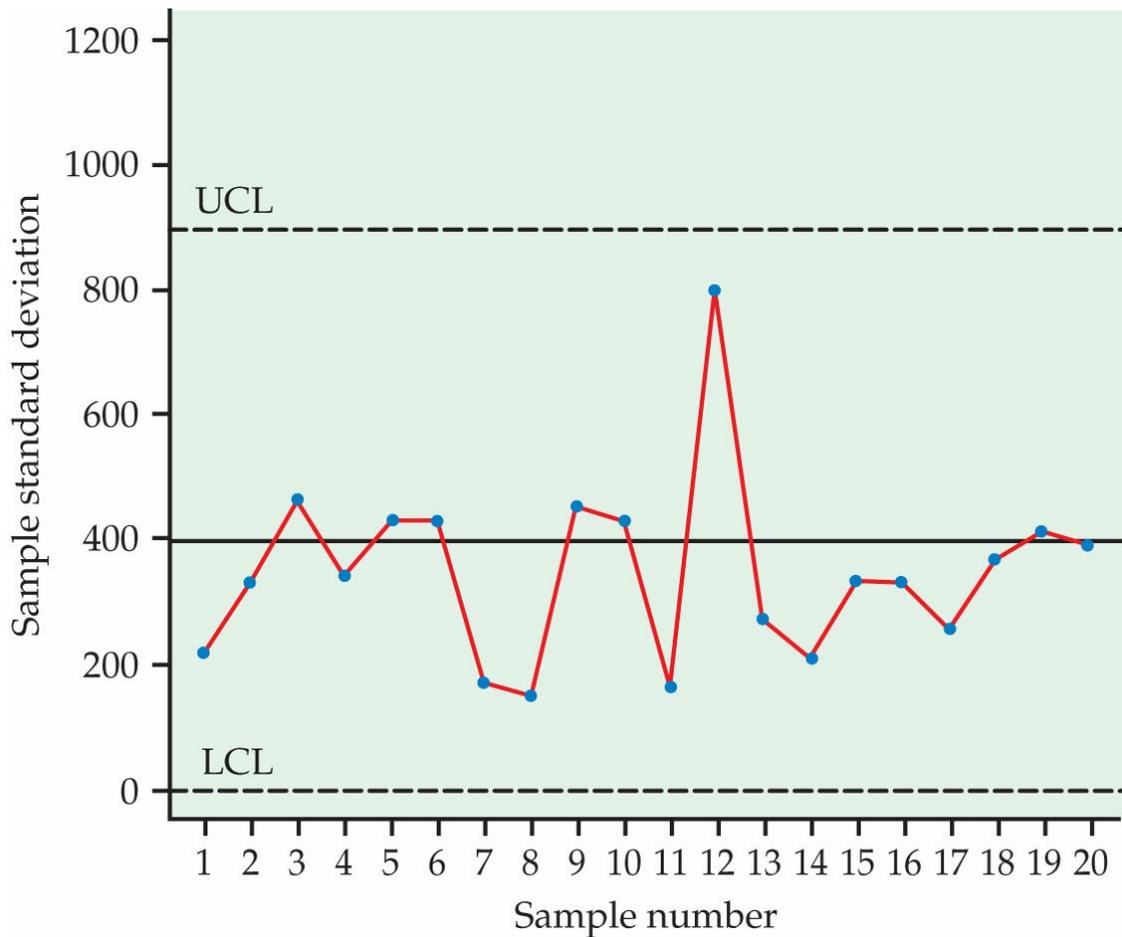
$$CL = c_4\sigma = (0.9213)(430) = 396 \text{ mm}$$

The control limits are

$$UCL = B_6\sigma = (2.088)(430) = 898$$

$$LCL = B_5\sigma = (0)(430) = 0$$

Figures 17.4 and 17.7 go together: they are the  $\bar{x}$  and  $s$  charts for monitoring the waterproofing process. Both charts are in control, showing only common cause variation within the bounds set by the control limits.



**FIGURE 17.7**

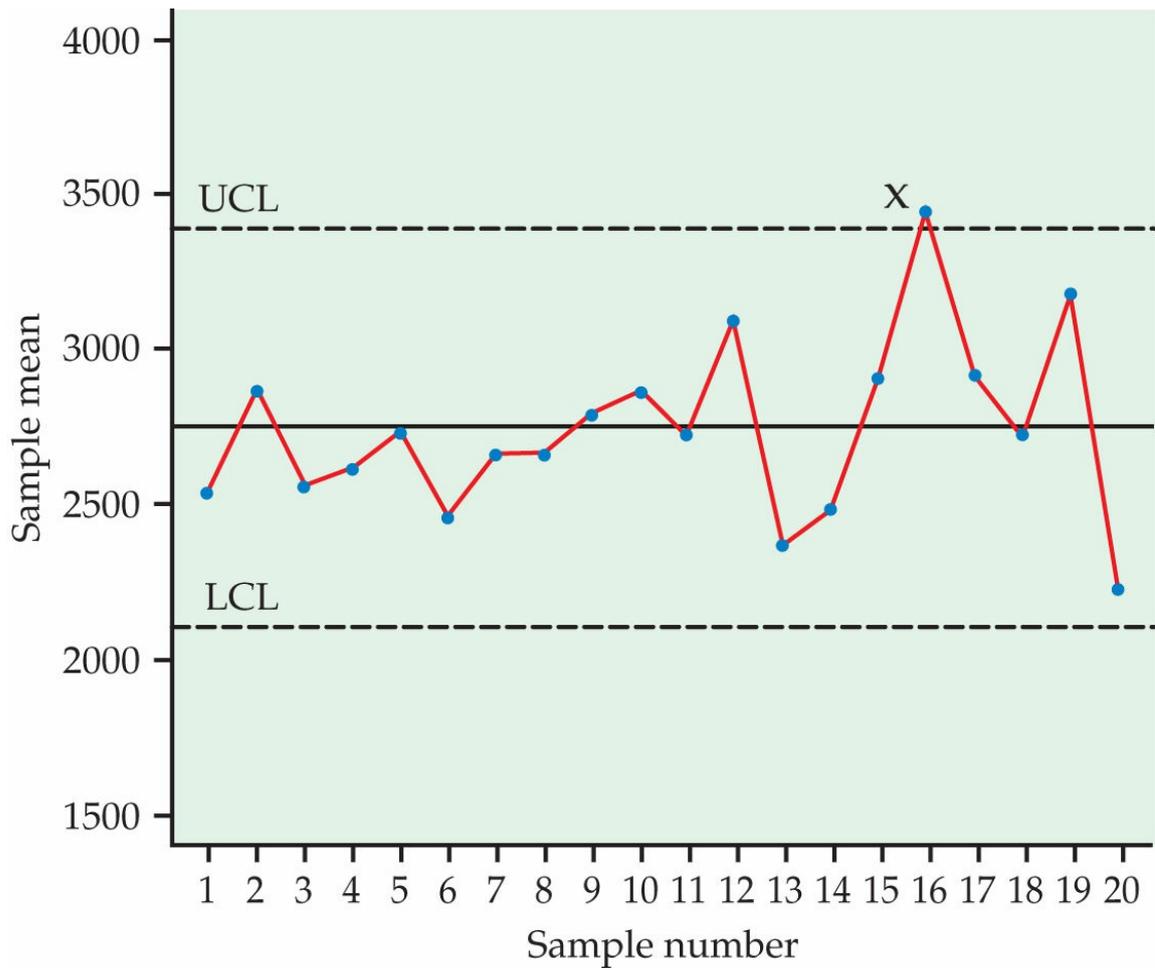
The  $s$  chart for the water resistance data of Table 17.1. Both the  $s$  chart and the  $\bar{x}$  chart (Figure 17.4) are in control.

Figures 17.8 and 17.9 are  $\bar{x}$  and  $s$  charts for the water resistance process when a new and poorly trained operator takes over the seam application between Samples 10 and 11. The new operator introduces added variation into the process, increasing the process standard deviation from its in-control value of 430 mm to 600 mm. The  $\bar{x}$  chart in Figure 17.8 shows one point out of control. Only on closer inspection do we see that the spread of the  $\bar{x}$ 's increases after Sample 10. In fact, the process mean has remained unchanged at 2750 mm. The apparent lack of control in the  $\bar{x}$  chart is entirely due to the larger process variation. There is a lesson here: *it is difficult to interpret an  $\bar{x}$  chart unless  $s$  is in control. When you look at  $\bar{x}$  and  $s$  charts, always start with the  $s$  chart.*



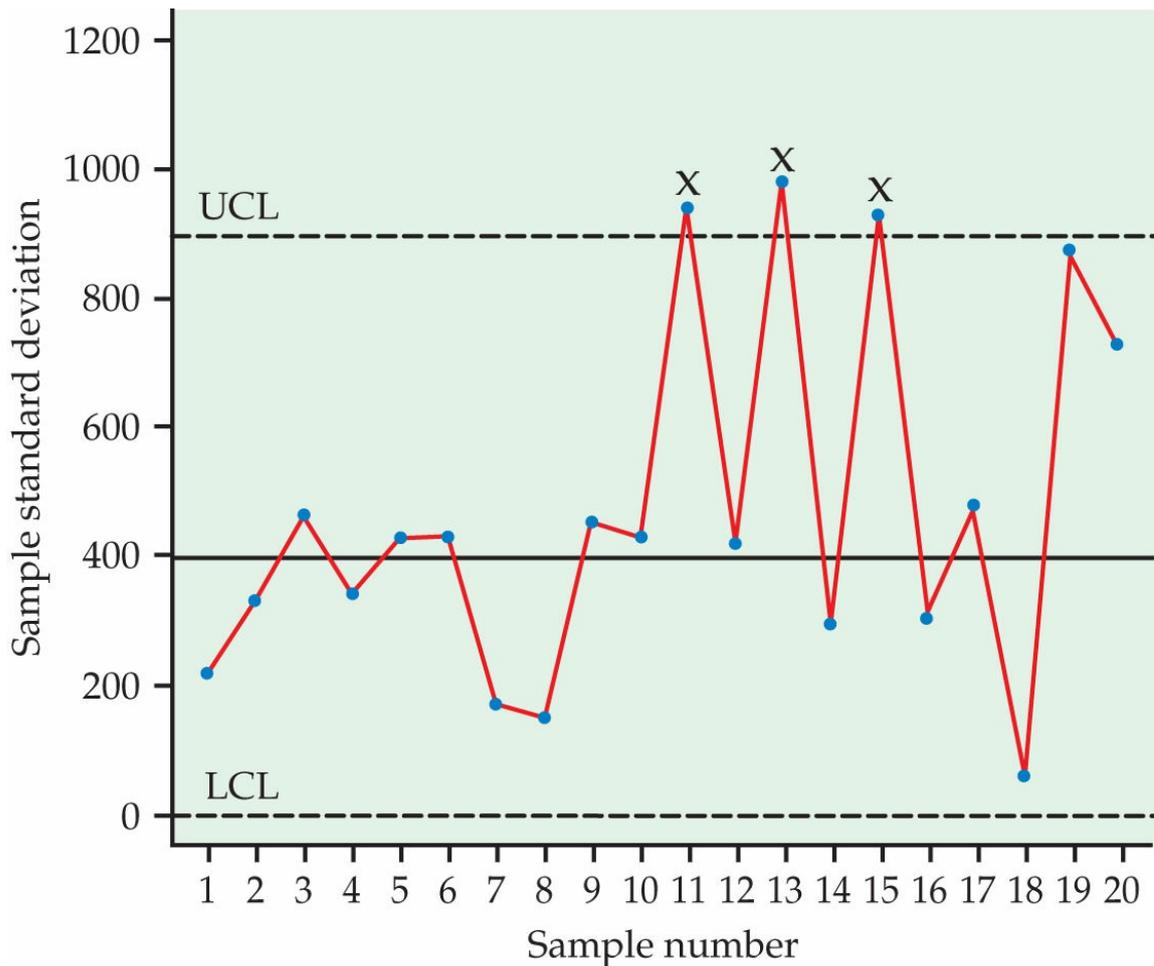
The  $s$  chart in Figure 17.9 shows lack of control starting at Sample 11. As usual, we mark the out-of-control points by an “x.” The points for Samples 13

and 15 also lie above the UCL, and the overall spread of the sample points is much greater than for the first 10 samples. In practice, the  $s$  chart would call for action after Sample 11. We would ignore the  $\bar{x}$  chart until the special cause (the new operator) for the lack of control in the  $s$  chart has been found and removed by training the operator.



**FIGURE 17.8**

The  $\bar{x}$  chart for water resistance when the process variability increases after Sample 10. The  $\bar{x}$  chart does show the increased variability, but the  $s$  chart is clearer and should be read first.



**FIGURE 17.9**

The  $s$  chart for water resistance when the process variability increases after Sample 10. Increased within-sample variability is clearly visible. Find and remove the  $s$ -type special cause before reading the  $\bar{x}$  chart.

Example 17.7 suggests a strategy for using  $\bar{x}$  and  $s$  charts in practice. First examine the  $s$  chart. Lack of control on an  $s$  chart is due to special causes that affect the observations *within a sample* differently. New and nonuniform raw material, a new and poorly trained operator, and mixing results from several machines or several operators are typical “ $s$ -type” special causes.

Once the  $s$  chart is in control, the stable value of the process standard deviation  $\sigma$  means that the variation within samples serves as a benchmark for detecting variation in the level of the process over the longer time periods between samples. The  $\bar{x}$  chart, with control limits that depend on  $\sigma$ , does this. The  $\bar{x}$  chart, as we saw in Example 17.7, responds to  $s$ -type causes as well as to longer-range changes in the process, so it is important to eliminate  $s$ -type special causes first. Then the  $\bar{x}$  chart will alert us to, for example, a change in process level caused by new raw material that differs from that used in the past or a gradual drift in the process level caused by wear in a cutting tool.

## EXAMPLE

### 17.8 Special causes and their effect on control charts.

A large health maintenance organization (HMO) uses control charts to monitor the process of directing patient calls to the proper department or doctor's receptionist. Each day at a random time, 5 consecutive calls are recorded electronically. The first call today is handled quickly by an experienced operator, but the next goes to a newly hired operator who must ask a supervisor for help. The sample has a large  $s$ , and lack of control signals the need to train new hires more thoroughly.

The same HMO monitors the time required to receive orders from its main supplier of pharmaceutical products. After a long period in control, the  $\bar{x}$  chart shows a systematic shift downward in the mean time because the supplier has changed to a more efficient delivery service. This is a desirable special cause, but it is nonetheless a systematic change in the process. The HMO will have to establish new control limits that describe the new state of the process, with smaller process mean  $\mu$ .

The second setting in Example 17.8 reminds us that a major change in the process returns us to the chart setup stage. In the absence of deliberate changes in the process, process monitoring uses the same values of  $\mu$  and  $\sigma$  for long periods of time. One exception is common: careful monitoring and removal of special causes as they occur can permanently reduce the process  $\sigma$ . If the points on the  $\sigma$  chart remain near the center line for a long period, it is wise to update the value of  $\sigma$  to the new, smaller value.

## SECTION 17.1 Summary

Work is organized in **processes**, chains of activities that lead to some result. We use **flowcharts** and **cause-and-effect diagrams** to describe processes.

All processes have variation. If the pattern of variation is stable over time, the process is **in statistical control**. **Control charts** are statistical plots intended to warn when a process is **out of control**.

Standard **3 $\sigma$  control charts** plot the values of some statistic  $Q$  for regular samples from the process against the time order of the samples. The **center line** is at the mean of  $Q$ . The **control limits** lie three standard deviations of  $Q$  above and below the center line. A point outside the control limits is an **out-of-control signal**. For **process monitoring** of a process that has been in control, the mean and standard deviation are based on past data from the process and are updated regularly.

When we measure some quantitative characteristic of the process, we use  $\bar{x}$

**and  $s$  charts** for process control. The  $s$  chart monitors variation within individual samples. If the  $s$  chart is in control, the  $\bar{x}$  chart monitors variation from sample to sample. To interpret the charts, always look first at the  $s$  chart.

## SECTION 17.1 Exercises

*For Exercises 17.1 and 17.2, see page 17-6; for Exercises 17.3 and 17.4, see page 17-8; and for Exercises 17.5 and 17.6, see page 17-12.*

### 17.7 Constructing a flowchart.

Consider the process of calling in a sandwich order for delivery to your apartment. Make a flowchart of this process, making sure to include steps that involve Yes/No decisions.

### 17.8 Determining sources of common and special cause variation.

Refer to the previous exercise. The time it takes from deciding to order a sandwich to receiving the sandwich will vary. List several common causes of variation in this time. Then list several special causes that might result in unusual variation.

### 17.9 Constructing a Pareto chart.

Comparisons are easier if you order the bars in a bar graph by height. A bar graph ordered from tallest to shortest bar is sometimes called a **Pareto chart**, after the Italian economist who recommended this procedure. Pareto charts are often used in quality studies to isolate the “vital few” categories on which we should focus our attention. Here is an example. Painting new auto bodies is a multistep process. There is an “electrocoat” that resists corrosion, a primer, a color coat, and a gloss coat. A quality study for one paint shop produced this breakdown of the primary problem type for those autos whose paint did not meet the manufacturer’s standards:

<b>Problem</b>	<b>Percent</b>
Electrocoat uneven—redone	4
Poor adherence of color to primer	5
Lack of clarity in color	2
“Orange peel” texture in color	32
“Orange peel” texture in gloss	1
Ripples in color coat	28
Ripples in gloss coat	4
Uneven color thickness	19
Uneven gloss thickness	5
<b>Total</b>	<b>100</b>

Make a Pareto chart. Which stage of the painting process should we look at first?

### 17.10 Constructing another Pareto chart.

A large hospital finds that it is losing money on surgery due to inadequate reimbursement by insurance companies and government programs. An initial study looks at losses broken down by diagnosis.

Government standards place cases into Diagnostic Related Groups (DRGs). For example, major joint replacements are DRG 209. Here is what the hospital finds:

DRG	Percent of losses
104	5.2
107	10.1
109	7.7
116	13.7
148	6.8
209	15.2
403	5.6
430	6.8
462	9.4

What percent of total losses do these 9 DRGs account for? Make a Pareto chart of losses by DRG. Which DRGs should the hospital study first when attempting to reduce its losses?

### 17.11 Making a Pareto chart.

Continue the study of the process of calling in a sandwich order (Exercise 17.7). If you kept good records, you could make a Pareto chart of the reasons (special causes) for unusually long order times. Make a Pareto chart of these reasons. That is, list the reasons based on your experience and chart your estimates of the percent each reason explains.

### 17.12 Control limits for label placement.

A rum producer monitors the position of its label on the bottle by sampling 4 bottles from each batch. One quantity measured is the distance from the bottom of the bottle neck to the top of the label. The process mean should be  $\mu = 2$  inches. Past experience indicates that the distance varies with  $\sigma = 0.1$  inches.

- The mean distance  $\bar{x}$  for each batch sample is plotted on an  $\bar{x}$  control chart. Calculate the center line and control limits for this chart.
- The sample standard deviation  $s$  for each batch's sample is plotted on an  $s$  control chart. What are the center line and control limits for this chart?

### 17.13 More on control limits for label placement.

Refer to the previous exercise. What happens to the center line and control limits for the  $\bar{x}$  and  $s$  control charts if

- the distributor samples 10 bottles from each batch?
- the distributor samples 2 bottles from each batch?
- the distributor uses centimeters rather than inches as the units?

### 17.14 Control limits for air conditioner thermostats.

A maker of auto air conditioners checks a sample of 6 thermostatic controls from each hour's production. The thermostats are set at 72°F and then placed in a chamber where the temperature is raised gradually.

The temperature at which the thermostat turns on the air conditioner is recorded. The process mean should be  $\mu = 72^\circ\text{F}$ . Past experience indicates that the response temperature of properly adjusted thermostats varies with  $\sigma = 0.6^\circ\text{F}$ .

- (a) The mean response temperature  $\bar{x}$  for each hour's sample is plotted on an  $\bar{x}$  control chart. Calculate the center line and control limits for this chart.
- (b) The sample standard deviation  $s$  for each hour's sample is plotted on an  $s$  control chart. What are the center line and control limits for this chart?

### 17.15 Control limits for a meat-packaging process.

A meat-packaging company produces 1-pound packages of ground beef by having a machine slice a long circular cylinder of ground beef as it passes through the machine. The timing between consecutive cuts will alter the weight of each section. Table 17.3 gives the weight of three consecutive sections of ground beef taken each hour over two 10-hour days. Past experience indicates that the process mean is 1.014 lb and the weight varies with  $\sigma = 0.019$  lb.  **MEATWGT**

- (a) Calculate the center line and control limits for an  $\bar{x}$  chart.
- (b) What are the center line and control limits for an  $s$  chart for this process?
- (c) Create the  $\bar{x}$  and  $s$  charts for these 20 consecutive samples.
- (d) Does the process appear to be in control? Explain.

### 17.16 Causes of variation in the time to respond to an application.

The personnel department of a large company records a number of performance measures. Among them is the time required to respond to an application for employment, measured from the time the application arrives. Suggest some plausible examples of each of the following.

- (a) Reasons for common cause variation in response time.
- (b)  $s$ -type special causes.
- (c)  $\bar{x}$ -type special causes.

**TABLE 17.3** Twenty Samples of Size 3, with  $\bar{x}$  and  $s$

Sample	Weight (pounds)			$\bar{x}$	$s$
1	0.999	1.071	1.019	1.030	0.0373
2	1.030	1.057	1.040	1.043	0.0137
3	1.024	1.020	1.041	1.028	0.0108
4	1.005	1.026	1.039	1.023	0.0172
5	1.031	0.995	1.005	1.010	0.0185
6	1.020	1.009	1.059	1.029	0.0263
7	1.019	1.048	1.050	1.039	0.0176
8	1.005	1.003	1.047	1.018	0.0247
9	1.019	1.034	1.051	1.035	0.0159
10	1.045	1.060	1.041	1.049	0.0098

11	1.007	1.046	1.014	1.022	0.0207
12	1.058	1.038	1.057	1.051	0.0112
13	1.006	1.056	1.056	1.039	0.0289
14	1.036	1.026	1.028	1.030	0.0056
15	1.044	0.986	1.058	1.029	0.0382
16	1.019	1.003	1.057	1.026	0.0279
17	1.023	0.998	1.054	1.025	0.0281
18	0.992	1.000	1.067	1.020	0.0414
19	1.029	1.064	0.995	1.029	0.0344
20	1.008	1.040	1.021	1.023	0.0159

### 17.17 Control charts for a tablet compression process.

A pharmaceutical manufacturer forms tablets by compressing a granular material that contains the active ingredient and various fillers. The hardness of a sample from each lot of tablets is measured in order to control the compression process. The process has been operating in control with mean at the target value  $\mu = 11.5$  kiloponds (kp) and estimated standard deviation  $\sigma = 0.2$  kp. Table 17.4 gives three sets of data, each representing  $\bar{x}$  for 20 successive samples of  $n = 4$  tablets. One set of data remains in control at the target value. In a second set, the process mean  $\mu$  shifts suddenly to a new value. In a third, the process mean drifts gradually.  **PILL**

**TABLE 17.4**

**Three Sets of  $\bar{x}$ 's from 20 Samples of Size 4**

Sample	Data set A	Data set B	Data set C
1	11.602	11.627	11.495
2	11.547	11.613	11.475
3	11.312	11.493	11.465
4	11.449	11.602	11.497
5	11.401	11.360	11.573
6	11.608	11.374	11.563
7	11.471	11.592	11.321
8	11.453	11.458	11.533
9	11.446	11.552	11.486
10	11.522	11.463	11.502
11	11.664	11.383	11.534
12	11.823	11.715	11.624
13	11.629	11.485	11.629
14	11.602	11.509	11.575
15	11.756	11.429	11.730
16	11.707	11.477	11.680
17	11.612	11.570	11.729
18	11.628	11.623	11.704
19	11.603	11.472	12.052
20	11.816	11.531	11.905

(a) What are the center line and control limits for an  $\bar{x}$  chart for this process?

(b) Draw a separate  $\bar{x}$  chart for each of the three data sets. Mark any points that are beyond the control limits.

(c) Based on your work in part (b) and the appearance of the control charts, which set of data comes from a process that is in control? In which case does the process mean shift suddenly, and at about which sample do you think that the mean changed? Finally, in which case does the mean drift gradually?

### 17.18 More on the tablet compression process.

Exercise 17.17 concerns process control data on the hardness of tablets for a pharmaceutical product. Table 17.5 gives data for 20 new samples of size 4, with the  $\bar{x}$  and  $s$  for each sample. The process has been in control with mean at the target value  $\mu = 11.5$  kp and standard deviation  $\sigma = 0.2$  kp.  **PILL1**

(a) Make both  $\bar{x}$  and  $s$  charts for these data based on the information given about the process.

(b) At some point, the within-sample process variation increased from  $\sigma = 0.2$  to  $\sigma = 0.4$ . About where in the 20 samples did this happen? What is the effect on the  $s$  chart? On the  $\bar{x}$  chart?

(c) At that same point, the process mean changed from  $\mu = 11.5$  to  $\mu = 11.7$ . What is the effect of this change on the  $s$  chart? On the  $\bar{x}$  chart?

### 17.19 Control limits for a milling process.

The width of a slot cut by a milling machine is important to the proper functioning of a hydraulic system for large tractors. The manufacturer checks the control of the milling process by measuring a sample of six consecutive items during each hour's production. The target width for the slot is  $\mu = 0.850$  inch. The process has been operating in control with center close to the target and  $\sigma = 0.002$  inch. What center line and control limits should be drawn on the  $s$  chart? On the  $\bar{x}$  chart?

**TABLE 17.5**

**Twenty Samples of Size 4, with  $\bar{x}$  and  $s$**

Sample	Hardness (kp)				$\bar{x}$	$s$
1	11.193	11.915	11.391	11.500	11.500	0.3047
2	11.772	11.604	11.442	11.403	11.555	0.1688
3	11.606	11.253	11.458	11.594	11.478	0.1642
4	11.509	11.151	11.249	11.398	11.326	0.1585
5	11.289	11.789	11.385	11.677	11.535	0.2362
6	11.703	11.251	11.231	11.669	11.463	0.2573
7	11.085	12.530	11.482	11.699	11.699	0.6094
8	12.244	11.908	11.584	11.505	11.810	0.3376
9	11.912	11.206	11.615	11.887	11.655	0.3284
10	11.717	11.001	11.197	11.496	11.353	0.3170
11	11.279	12.278	11.471	12.055	11.771	0.4725
12	12.106	11.203	11.162	12.037	11.627	0.5145
13	11.490	11.783	12.125	12.010	11.852	0.2801
14	12.299	11.924	11.235	12.014	11.868	0.4513
15	11.380	12.253	11.861	12.242	11.934	0.4118
16	11.220	12.226	12.216	11.824	11.872	0.4726
17	11.611	11.658	11.977	10.813	11.515	0.4952

18	12.251	11.481	11.156	12.243	11.783	0.5522
19	11.559	11.065	12.186	10.933	11.435	0.5681
20	11.106	12.444	11.682	12.378	11.902	0.6331

### 17.20 Control limits for a dyeing process.

The unique colors of the cashmere sweaters your firm makes result from heating undyed yarn in a kettle with a dye liquor. The pH (acidity) of the liquor is critical for regulating dye uptake and hence the final color. There are five kettles, all of which receive dye liquor from a common source. Twice each day, the pH of the liquor in each kettle is measured, giving a sample of size 5. The process has been operating in control with  $\mu = 4.24$  and  $\sigma = 0.137$ .

- Give the center line and control limits for the  $s$  chart.
- Give the center line and control limits for the  $\bar{x}$  chart.

### 17.21 Control charts for a mounting-hole process.

Figure 17.10 reproduces a data sheet from a factory that makes electrical meters.<sup>6</sup> The sheet shows measurements of the distance between two mounting holes for 18 samples of size 5. The heading informs us that the measurements are in multiples of 0.0001 inch above 0.6000 inch. That is, the first measurement, 44, stands for 0.6044 inch. All the measurements end in 4. Although we don't know why this is true, it is clear that in effect the measurements were made to the nearest 0.001 inch, not to the nearest 0.0001 inch. Based on long experience with this process, you are keeping control charts based on  $\mu = 43$  and  $\sigma = 12.74$ .

Make  $s$  and  $\bar{x}$  charts for the data in Figure 17.10 and describe the state of the process. 

### 17.22 Identifying special causes on control charts.

The process described in Exercise 17.20 goes out of control. Investigation finds that a new type of yarn was recently introduced. The pH in the kettles is influenced by both the dye liquor and the yarn. Moreover, on a few occasions a faulty valve on one of the kettles had allowed water to enter that kettle; as a result, the yarn in that kettle had to be discarded. Which of these special causes appears on the  $s$  chart and which on the  $\bar{x}$  chart? Explain your answer.

### 17.23 Determining the probability of detection.

An  $\bar{x}$  chart plots the means of samples of size 4 against center line  $CL = 715$  and control limits  $LCL = 680$  and  $UCL = 750$ . The process has been in control.

- What are the process mean and standard deviation?
- The process is disrupted in a way that changes the mean to  $\mu = 700$ . What is the probability that the first sample after the disruption gives a point beyond the control limits of the  $\bar{x}$  chart?

		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1																								
2		VARIABLES CONTROL CHART ( $\bar{X}$ & R)																		Part No.	Chart No.			
3																				32506	1			
4	Part name (project)	Operation (process)										Specification limits												
5	Metal frame	Distance between mounting holes										0.6054" $\pm$ 0.0010"												
6	Operator	Machine			Gage			Unit of measure			Zero equals													
7		R-5						0.0001"			0.6000"													
8	Date	3/7			3/8			3/9																
9	Time	8:30	10:30	11:45	1:30	8:15	10:15	11:45	2:00	3:00	4:00	8:30	10:00	11:45	1:30	2:30	3:30	4:30	5:30					
10	Sample measurements	1	44	64	34	44	34	34	54	64	24	34	34	54	44	24	54	54	54	54				
11		2	44	44	44	54	14	64	64	34	54	44	44	44	24	24	24	34	34	24				
12		3	44	34	54	54	84	34	34	54	44	44	34	24	34	54	54	24	74	64				
13		4	44	34	44	34	54	44	44	44	34	34	64	54	34	44	44	44	44	34				
14		5	64	54	54	44	44	44	34	44	34	34	34	24	44	44	44	54	54	44				
15																								
16	Average, $\bar{X}$																							
17	Range, R	20	30	20	20	70	30	30	30	10	30	30	20	30	30	30	40	40						
18																								
19																								

**FIGURE 17.10**

A process control record sheet kept by operators, for Exercise 17.21. This is typical of records kept by hand when measurements are not automated. We will see in the next section why such records mention  $\bar{x}$  and  $R$  control charts rather than  $\bar{x}$  and  $s$  charts.

(c) The process is disrupted in a way that changes the mean to  $\mu = 700$  and the standard deviation to  $\sigma = 10$ . What is the probability that the first sample after the disruption gives a point beyond the control limits of the  $\bar{x}$  chart?

### 17.24 Alternative control limits.

American and Japanese practice uses  $3\sigma$  control charts. That is, the control limits are three standard deviations on either side of the mean. When the statistic being plotted has a Normal distribution, the probability of a point outside the limits is about 0.003 (or about 0.0015 in each direction) by the 68–95–99.7 rule (page 59). European practice uses control limits placed so that the probability of a point outside the limits when in control is 0.001 in each direction. For a Normally distributed statistic, how many standard deviations on either side of the mean do these alternative control limits lie?

### 17.25 $2\sigma$ control charts.

Some special situations call for  $2\sigma$  control charts. That is, the control limits for a statistic  $Q$  will be  $\mu_Q \pm 2\sigma_Q$ . Suppose that you know the process mean  $\mu$  and standard deviation  $\sigma$  and will plot  $\bar{x}$  and  $s$  from samples of size  $n$ .

- (a) What are the  $2\sigma$  control limits for an  $\bar{x}$  chart?
- (b) Find expressions for the upper and lower  $2\sigma$  control limits for an  $s$  chart in terms of the control chart constants  $c_4$  and  $c_5$  introduced on page 17-13.

## 17.2 Using Control Charts

When you complete this section, you will be able to

- Implement various out-of-control rules when interpreting control charts.
- Set up a control chart (that is, tentative control limits and center line) based on past data.
- Identify rational subgroups when deciding how to choose samples.
- Distinguish between the natural tolerances for a product and the control limits for a process, as well as between capability and control.

We are now familiar with the ideas behind all control charts as well as the details of making  $\bar{x}$  and  $s$  charts. This section discusses a variety of topics related to using control charts in practice.

### $\bar{x}$ and $R$ charts

We have seen that it is essential to monitor both the center and the spread of a process. Control charts were originally intended to be used by factory workers with limited knowledge of statistics in the era before even calculators, let alone software, were common. In that environment, the standard deviation is too difficult to calculate. The  $\bar{x}$  chart for center was therefore used with a control chart for spread based on the **sample range** rather than the sample standard deviation.

#### sample range

The range  $R$  of a sample is just the difference between the largest and smallest observations. It is easy to find  $R$  without a calculator. Using  $R$  rather than  $s$  to measure the spread of samples replaces the  $s$  chart with an  **$R$  chart**. It also changes the  $\bar{x}$  chart because the control limits for  $\bar{x}$  use the estimated process spread.

#### $R$ chart

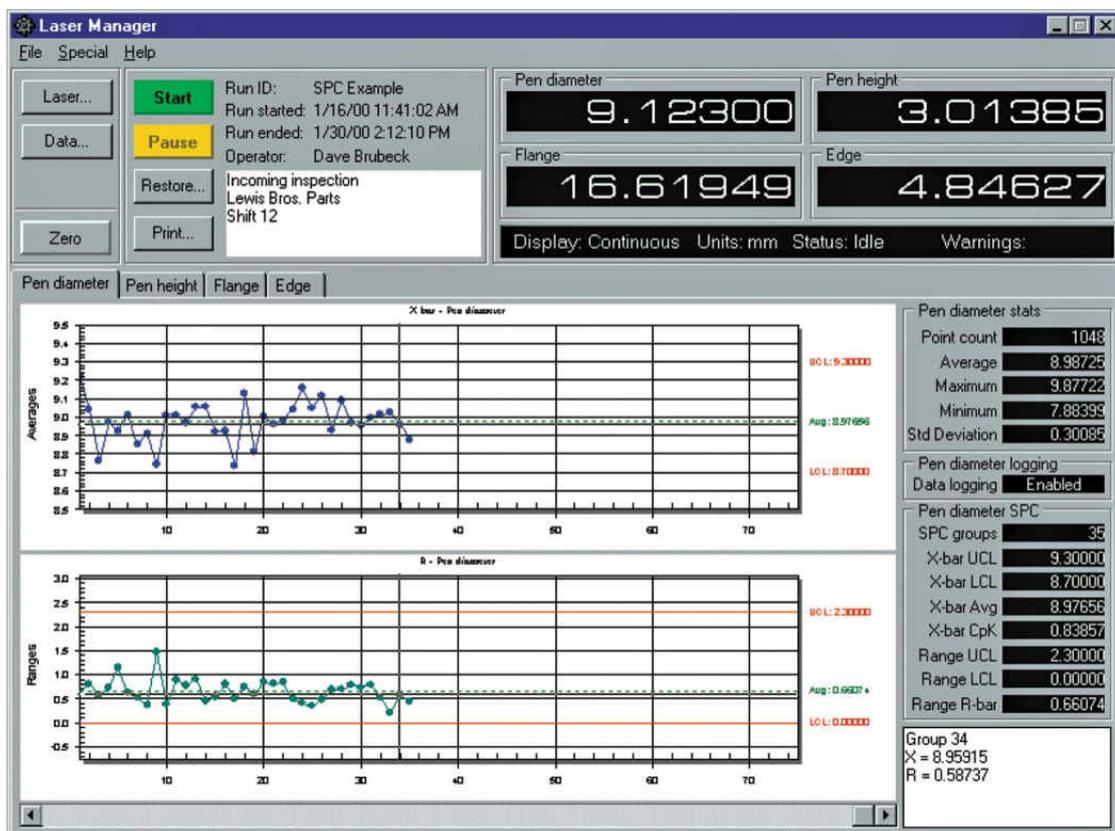
Because the range  $R$  uses only the largest and smallest observations in a sample, it is less informative than the standard deviation  $s$  calculated from all the observations. For this reason,  $\bar{x}$  and  $s$  charts are now preferred to  $\bar{x}$  and  $R$  charts.  $R$  charts, however, remain common because it is easier for workers to understand  $R$  than  $s$ .

In this short introduction, we concentrate on the principles of control charts, so we won't give the details of constructing  $\bar{x}$  and  $R$  charts. These details appear in

any text on quality control.<sup>7</sup> If you meet a set of  $\bar{x}$  and  $R$  charts, remember that the interpretation of these charts is just like the interpretation of  $\bar{x}$  and  $s$  charts.

## EXAMPLE

### 17.9 Example of a typical process control technology.



**FIGURE 17.11**

Output for operators, from the Laser Manager software by System Dynamics, Inc. The software prepares control charts directly from measurements made by a laser micrometer. Compare the hand record sheet in Figure 17.10. (Image provided by Gordon A. Feingold, System Dynamics, Inc. Used by permission.)

Figure 17.11 is a display produced by custom process control software attached to a laser micrometer. In this demonstration prepared by the software maker, the micrometer is measuring the diameter in millimeters of samples of pens shipped by an office supply company. The software controls the laser, records measurements, makes the control charts, and sounds an alarm when a point is out of control. This is typical of process control technology in modern manufacturing settings.

The software presents  $\bar{x}$  and  $R$  charts rather than  $\bar{x}$  and  $s$  charts. The  $R$  chart monitors within-sample variation (just like an  $s$  chart), so we look at it first. We see that the process spread is stable and well within the control limits. Just as in the case of  $s$ , the LCL for  $R$  is 0 for the samples of size  $n = 5$  used here. The  $\bar{x}$  chart is also in control, so process monitoring will continue. The software will sound an alarm if either chart goes out of control.

## USE YOUR KNOWLEDGE

### 17.26 What's wrong?

For each of the following, explain what is wrong and why.

- (a) The  $R$  chart monitors the center of the process.
- (b) The  $R$  chart is commonly used because the range  $R$  is more informative than the standard deviation  $s$ .
- (c) Use of the range  $R$  to monitor process spread does not alter the construction of the control limits for the  $\bar{x}$  chart.

## Additional out-of-control rules

So far, we have used only the basic “one point beyond the control limits” criterion to signal that a process may have gone out of control. We would like a quick signal when the process moves out of control, but we also want to avoid “false alarms,” signals that occur just by chance when the process is really in control.

The standard  $3\sigma$  control limits are chosen to prevent too many false alarms, because an out-of-control signal calls for an effort to find and remove a special cause. As a result,  $\bar{x}$  charts are often slow to respond to a gradual drift in the process center.

We can speed the response of a control chart to lack of control—at the cost of also enduring more false alarms—by adding patterns other than “one-point-out” as rules. The most common step in this direction is to add a *runs rule* to the  $\bar{x}$  chart.

## OUT-OF-CONTROL SIGNALS

$\bar{x}$  and  $s$  or  $\bar{x}$  and  $R$  control charts produce an out-of-control signal if

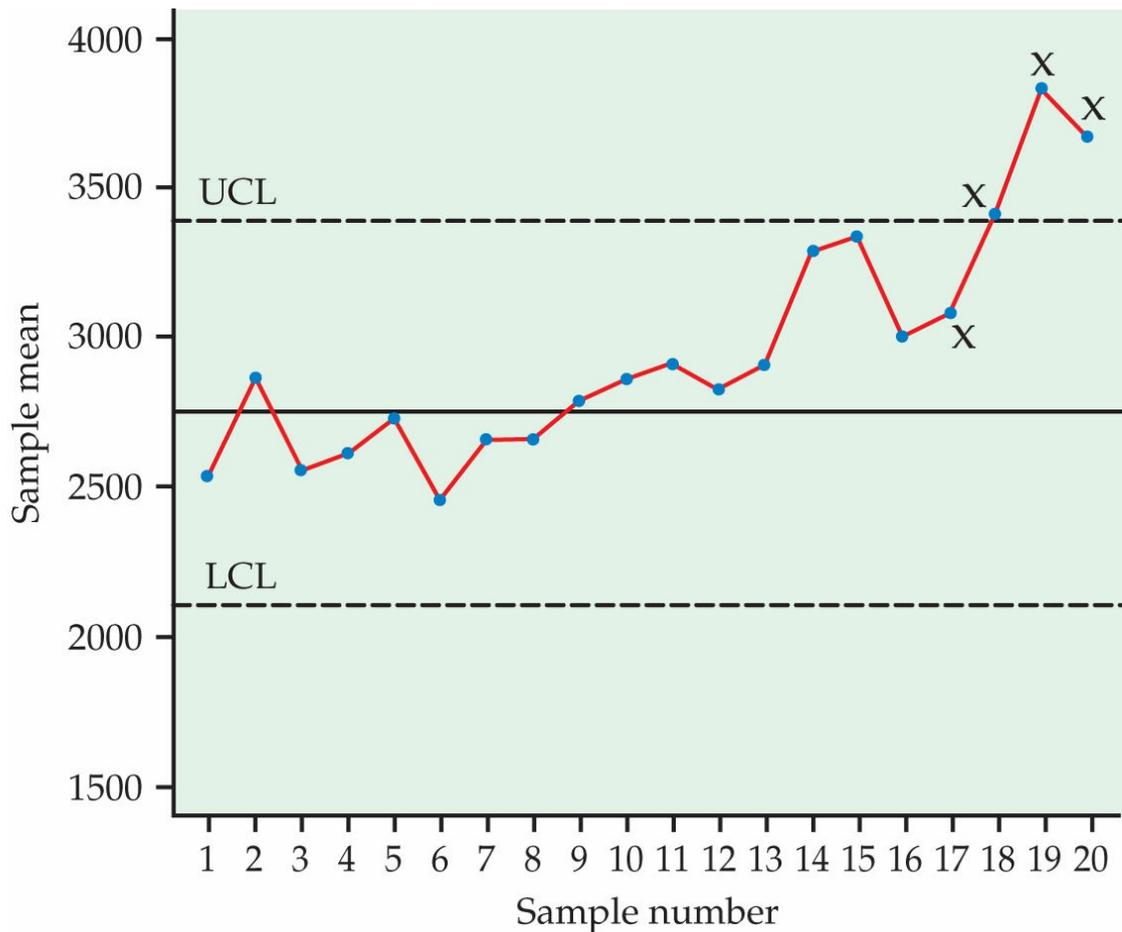
- (a) **One-point-out:** A single point lies outside the  $3\sigma$  control limits of either chart.

(b) **Run:** The  $\bar{x}$  chart shows 9 consecutive points above the center line or 9 consecutive points below the center line. The signal occurs when we see the 9th point of the run.

## EXAMPLE

### 17.10 Effectiveness of the runs rule.

Figure 17.12 reproduces the  $\bar{x}$  chart from Figure 17.6. The process center began a gradual upward drift at Sample 11. The chart shows the effect of the drift—the sample means plotted on the chart move gradually upward, with some random variation. The one-point-out rule does not call for action until Sample 18 finally produces an  $\bar{x}$  above the UCL. The runs rule reacts slightly more quickly: Sample 17 is the 9th consecutive point above the center line.



**FIGURE 17.12**

The  $\bar{x}$  chart for water resistance data when the process center drifts upward, for Example 17.10. The “run of 9” signal gives an out-of-control warning at Sample 17.

It is a mathematical fact that the runs rule responds to a gradual drift more quickly (on the average) than the one-point-out rule does. The motivation for a runs rule is that when a process is in control, half the points on an  $\bar{x}$  chart should lie above the center line and half below. That’s true on the average in the long term. In the short term, we will see runs of points above or below, just as we see runs of heads or tails in tossing a coin.

To determine how long a run must be to suggest that the process center has moved, we once again concern ourselves with the cost of false alarms. The 99.7 part of the 68–95–99.7 rule says that we will get a point outside the  $3\sigma$  control limits about 3 times for every 1000 points plotted when the process is in control. The chance of 9 straight points above the center line when the process is in control is  $(1/2)^9 = 1/512$ , or about 2 per 1000. The chance for a run of 9 below the center line is the same. Combined, that’s about 4 false alarms per 1000 plotted points overall when the process is in control. This is very close to the false-alarm rate for one-point-out.

There are many other patterns that can be added to the rules for responding to  $\bar{x}$  and  $s$  or  $\bar{x}$  and  $R$  charts. In our enthusiasm to detect various special kinds of loss of control, it is easy to forget that adding rules always increases

the frequency of false alarms. Frequent false alarms are so annoying that the people responsible for responding soon begin to ignore out-of-control signals. *It is better to use only a few out-of-control rules and to reserve rules other than one-point-out and runs for processes that are known to be prone to specific special causes for which there are tailor-made detection rules.*<sup>8</sup>



## USE YOUR KNOWLEDGE

### 17.27 What's wrong?

For each of the following, explain what is wrong and why.

- (a) For the one-point-out rule, you could reduce the frequency of false alarms by using  $2\sigma$  control limits.
- (b) In speeding up the response of a control chart to lack of control, we decrease the frequency of false alarms.
- (c) The runs rule is designed to quickly detect a large and sudden shift in the process.

### 17.28 The effect of special cause variation.

Is each of the following examples of a special cause most likely to first result in (i) one-point-out on the  $s$  or  $R$  chart, (ii) one-point-out on the  $\bar{x}$  chart, or (iii) a run on the  $\bar{x}$  chart? In each case, briefly explain your reasoning.

- (a) An etching solution deteriorates as more items are etched.
- (b) Buildup of dirt reduces the precision with which parts are placed for machining.
- (c) A new customer service representative for a Spanish-language help line is not a native speaker and has difficulty understanding customers.
- (d) A data entry employee grows less attentive as his shift continues.

## Setting up control charts

When you first encounter a process that has not been carefully studied, it is quite likely that the process is not in control. Your first goal is to discover and remove

special causes and so bring the process into control. Control charts are an important tool. Control charts for *process monitoring* follow the process forward in time to keep it in control. Control charts at the *chart setup* stage, on the other hand, look back in an attempt to discover the present state of the process. An example will illustrate the method.

## EXAMPLE

### 17.11 Monitoring the viscosity of a material.



The viscosity of a material is its resistance to flow when under stress. Viscosity is a critical characteristic of rubber and rubber-like compounds called elastomers, which have many uses in consumer products. Viscosity is measured by placing specimens of the material above and below a slowly rotating roller, squeezing the assembly, and recording the drag on the roller. Measurements are in “Mooney units,” named after the inventor of the instrument.

**TABLE 17.6**  $\bar{x}$  and  $s$  for 24 Samples of Elastomer Viscosity (in Mooneys)

Sample	$\bar{x}$	$s$	Sample	$\bar{x}$	$s$
1	49.750	2.684	13	47.875	1.118
2	49.375	0.895	14	48.250	0.895
3	50.250	0.895	15	47.625	0.671
4	49.875	1.118	16	47.375	0.671
5	47.250	0.671	17	50.250	1.566
6	45.000	2.684	18	47.000	0.895
7	48.375	0.671	19	47.000	0.447
8	48.500	0.447	20	49.625	1.118
9	48.500	0.447	21	49.875	0.447
10	46.250	1.566	22	47.625	1.118
11	49.000	0.895	23	49.750	0.671
12	48.125	0.671	24	48.625	0.895

A specialty chemical company is beginning production of an elastomer that

is supposed to have viscosity  $45 \pm 5$  Mooneys. Each lot of the elastomer is produced by “cooking” raw material with catalysts in a reactor vessel. Table 17.6 records  $\bar{x}$  and  $s$  from samples of size  $n = 4$  lots from the first 24 shifts as production begins.<sup>9</sup> An  $s$  chart therefore monitors variation among lots produced during the same shift. If the  $s$  chart is in control, an  $\bar{x}$  chart looks for shift-to-shift variation.

## Estimating $\mu$

We do not know the process mean  $\mu$  and standard deviation  $\sigma$ . What shall we do? Sometimes we can easily adjust the center of a process by setting some control, such as the depth of a cutting tool in a machining operation or the temperature of a reactor vessel in a pharmaceutical plant. In such cases it is common to simply take the process mean  $\mu$  to be the target value, the depth or temperature that the design of the process specifies as correct. The  $\bar{x}$  chart then helps us keep the process mean at this target value.

There is less likely to be a “correct value” for the process mean  $\mu$  if we are monitoring response times to customer calls or data entry errors. In Example 17.11, we have the target value 45 Mooneys, but there is no simple way to set viscosity at the desired level. In such cases, we want the  $\mu$  we use in our  $\bar{x}$  chart to describe the center of the process as it has actually been operating. To do this, take the mean of all the individual measurements in the past samples. Because the samples are all the same size, this is just the mean of the sample  $\bar{x}$ 's. The overall “mean of the sample means” is therefore usually called  $\bar{\bar{x}}$ . For the 24 samples in Table 17.6,

$$\begin{aligned}\bar{\bar{x}} &= 124(49.750 + 49.375 + \dots + 48.625) \\ &= 1161.12524 = 48.380\end{aligned}$$



## Estimating $\sigma$

*It is almost never safe to use a “target value” for the process standard deviation  $\sigma$  because it is almost never possible to directly adjust process variation. We must estimate  $\sigma$  from past data. We want to combine the sample standard deviations  $s$  from past samples rather than use the standard deviation of all the individual observations in those samples. That is, in Example 17.11, we want to combine the 24 sample standard deviations in Table 17.6 rather than calculate the standard deviation of the 96 observations in these samples. The reason is that it is the *within-sample* variation that is the benchmark against which we compare the longer-term process variation. Even if the process has been in control, we want*

only the variation over the short time period of a single sample to influence our value for  $\sigma$ .

There are several ways to estimate  $\sigma$  from the sample standard deviations. Software may use a somewhat sophisticated method and then calculate the control limits for you. Here, we use a simple method that is traditional in quality control because it goes back to the era before software. If we are basing chart setup on  $k$  past samples, we have  $k$  sample standard deviations  $s_1, s_2, \dots, s_k$ . Just average these to get

$$\bar{s} = \frac{1}{k}(s_1 + s_2 + \dots + s_k)$$

For the viscosity example, we average the  $s$ -values for the 24 samples in Table 17.6,

$$\begin{aligned}\bar{s} &= \frac{1}{24}(2.684 + 0.895 + \dots + 0.895) \\ &= \frac{24.15624}{24} = 1.0065\end{aligned}$$

Combining the sample  $s$ -values to estimate  $\sigma$  introduces a complication: the samples used in process control are often small (size  $n = 4$  in the viscosity example), so  $s$  has some bias as an estimator of  $\sigma$ . The estimator  $\bar{s}$  inherits this bias. A proper estimate of  $\sigma$  corrects this bias. Thus, our estimator is



**mean of  $s$ , p. 17-13**

$$\hat{\sigma} = \bar{s} c_4$$

We get control limits from past data by using the estimates  $\bar{x}$  and  $\hat{\sigma}$  in place of the  $\mu$  and  $\sigma$  used in charts at the process-monitoring stage. Here are the results.<sup>10</sup>

## $\bar{x}$ AND $s$ CONTROL CHARTS USING PAST DATA

Take regular samples of size  $n$  from a process. Estimate the process mean  $\mu$  and the process standard deviation  $\sigma$  from past samples by

$$\hat{\mu} = \bar{x} \text{ (or use at target value)}$$

$$\hat{\sigma} = \bar{s} c_4$$

The center line and control limits for an  $\bar{x}$  chart are

$$UCL = \hat{\mu} + 3\hat{\sigma}/n$$

$$CL = \hat{\mu}$$

$$LCL = \mu^{\wedge} - 3\sigma^{\wedge}$$

The center line and control limits for an *s* chart are

$$UCL = B_6\sigma^{\wedge}$$

$$CL = \bar{s} = c_4\sigma^{\wedge}$$

$$LCL = B_5\sigma^{\wedge}$$

If the process was not in control when the samples were taken, these should be regarded as trial control limits.

### Chart setup

We are now ready to outline the chart setup procedure for the elastomer viscosity.

**Step 1.** As usual, we look first at an *s* chart. For chart setup, control limits are based on the same past data that we will plot on the chart. Based on Table 17.6,

$$\bar{s} = 1.0065$$

$$\sigma^{\wedge} = \bar{s} / c_4 = 1.0065 / 0.9213 = 1.0925$$

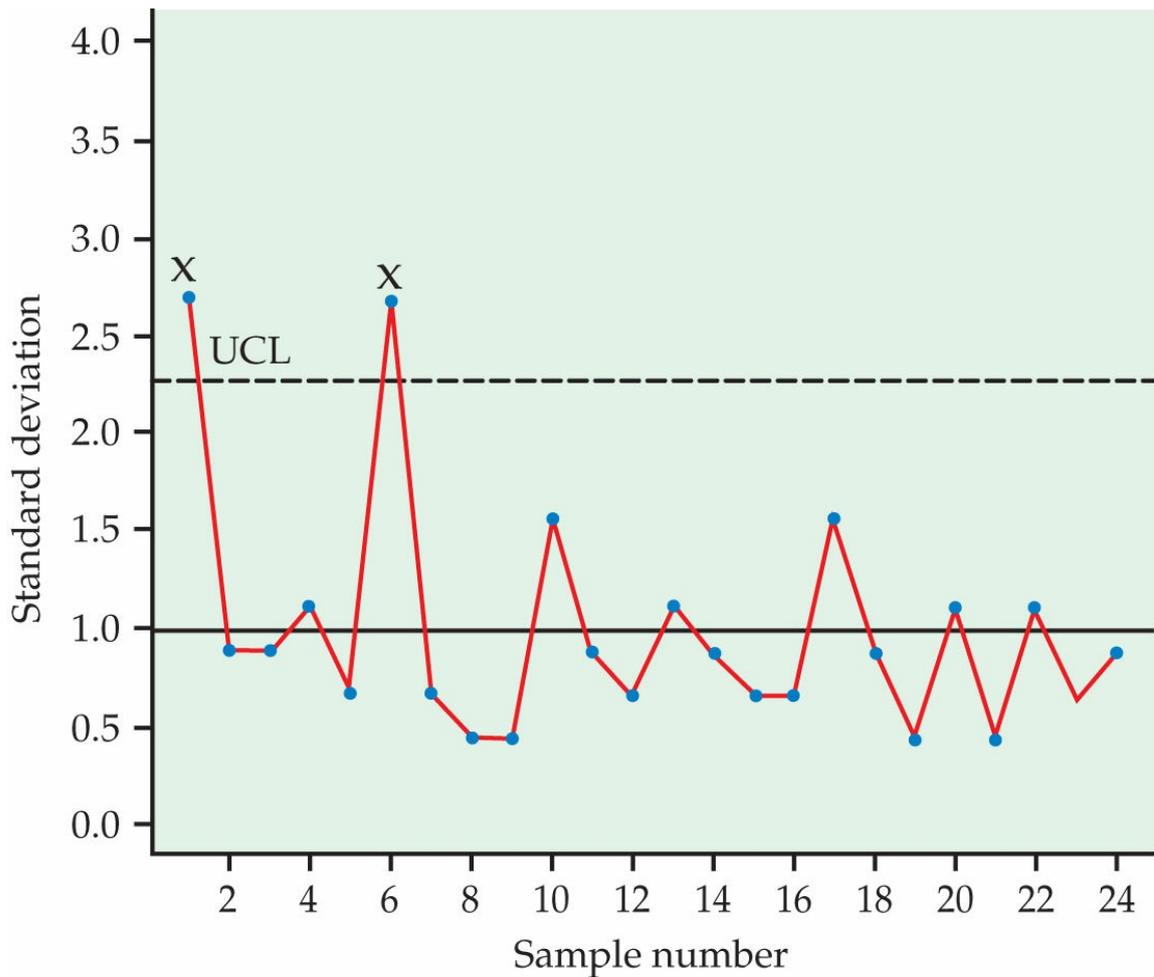
So the center line and control limits for the *s* chart are

$$UCL = B_6\sigma^{\wedge} = (2.088)(1.0925) = 2.281$$

$$CL = \bar{s} = 1.0065$$

$$LCL = B_5\sigma^{\wedge} = (0)(1.0925) = 0$$

Figure 17.13 is the *s* chart. The points for Shifts 1 and 6 lie above the UCL. Both are near the beginning of production. Investigation finds that the reactor operator made an error on one lot in each of these samples. The error changed the viscosity of those two lots and increased *s* for each of the samples. The error will not be repeated now that the operators have gained experience. That is, this special cause has already been removed.



**FIGURE 17.13**

The  $s$  chart based on past data for the viscosity data of Table 17.6. The control limits are based on the same  $s$ -values that are plotted on the chart. Points 1 and 6 are out of control.

**Step 2.** Remove the two values of  $s$  that were out of control. This is proper because the special cause responsible for these readings is no longer present. From the remaining 22 shifts

$$\bar{s} = 0.854 \quad \text{and} \quad \hat{\sigma} = 0.8540.9213 = 0.927$$

The new  $s$  chart center line and control limits are

$$UCL = B_6 \hat{\sigma} = (2.088)(0.927) = 1.936$$

$$CL = \bar{s} = 0.854$$

$$LCL = B_5 \hat{\sigma} = (0)(0.927) = 0$$

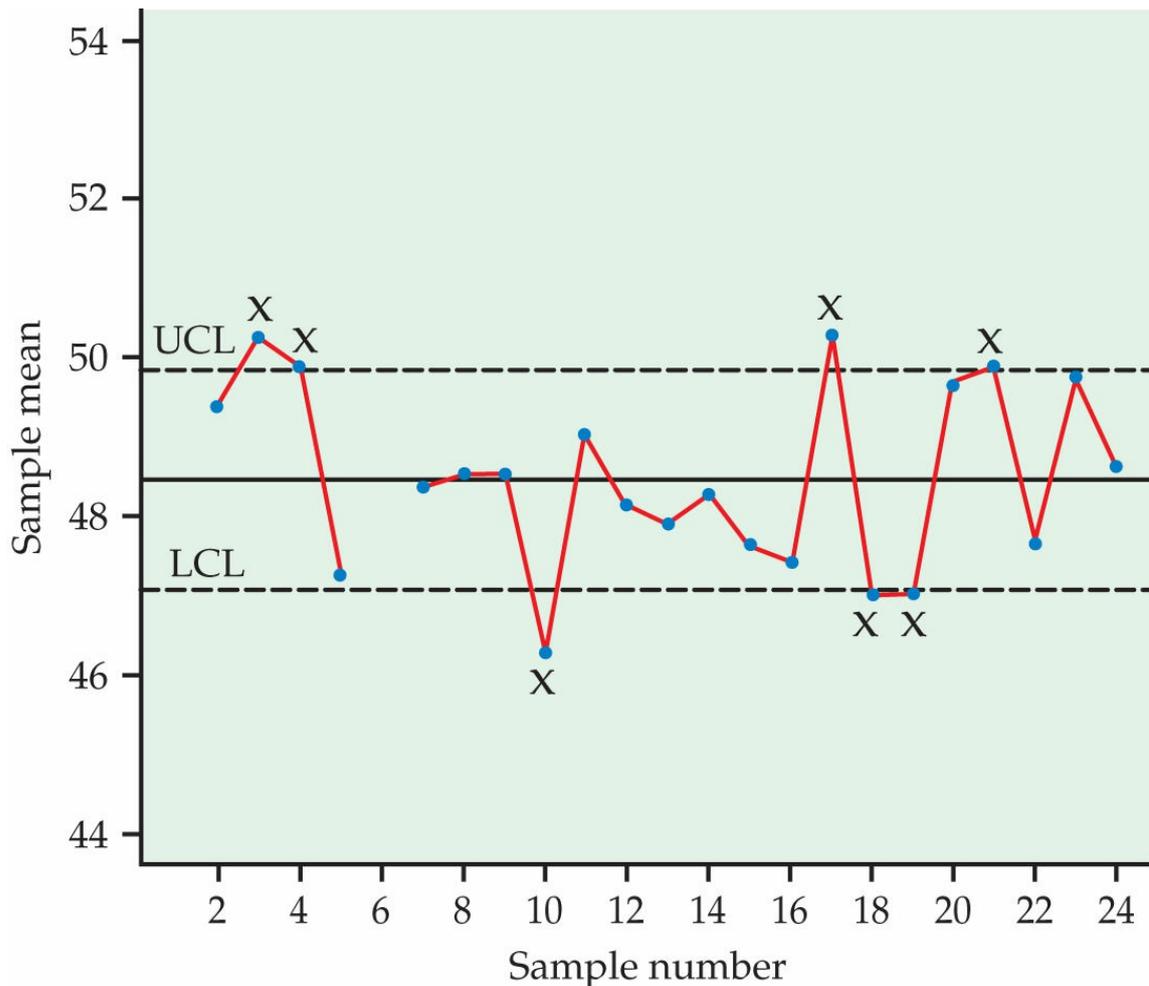
We don't show this chart, but you can see from Table 17.6 and Figure 17.13 that none of the remaining  $s$ -values lies above the new, lower UCL; the largest remaining  $s$  is 1.566. If additional points were out of control, we would repeat the process of finding and eliminating  $s$ -type causes until the  $s$  chart for the remaining shifts is in control. In practice, this is often a challenging task.

**Step 3.** Once  $s$ -type causes have been eliminated, make an  $\bar{x}$  chart *using only the samples that remain* after dropping those that had out-of-control  $s$ -values. For the 22 remaining samples, we calculate  $\bar{x} = 48.4716$  and we know that  $\hat{\sigma} = 0.927$ . The center line and control limits for the  $\bar{x}$  chart are

$$UCL = \bar{x} + 3\hat{\sigma} = 48.4716 + 3(0.927) = 49.862$$

$$CL = \bar{x} = 48.4716$$

$$LCL = \bar{x} - 3\hat{\sigma} = 48.4716 - 3(0.927) = 47.081$$



**FIGURE 17.14**

The  $\bar{x}$  chart based on past data for the viscosity data of Table 17.6. The samples for Shifts 1 and 6 have been removed because  $s$ -type special causes active in those samples are no longer active. The  $\bar{x}$  chart shows poor control.

Figure 17.14 is the  $\bar{x}$  chart. Shifts 1 and 6 were already dropped. Seven of the remaining 22 points are beyond the  $3\sigma$  limits, four high and three low. Although within-shift variation is now stable, there is excessive variation from shift to shift. To find the cause, we must understand the details of the process, but knowing that the special cause or causes operate between shifts is a big help. If the reactor is set up anew at the beginning of each shift, that's one place to look more closely.

**Step 4.** Once the  $\bar{x}$  and  $s$  charts are both in control (looking backward), use the estimates  $\hat{\mu}$  and  $\hat{\sigma}$  from the points in control to set tentative control limits to monitor the process going forward. If it remains in control, we can update the charts and move to the process-monitoring stage.

## USE YOUR KNOWLEDGE

### 17.29 Updating control chart limits.



MEATWGT

Suppose that when the process improvement project of Example 17.11 (page 17-26) is complete, the points remaining after removing special causes have  $\bar{x}=47.2$  and  $\bar{s}=1.03$ . What are the center line and control limits for the  $\bar{x}$  and  $s$  charts you would use to monitor the process going forward?

### 17.30 More on updating control chart limits.

In Exercise 17.15, control limits for the weight of ground beef were obtained using historical results. Using Table 17.3 (page 17-19), estimate the process  $\mu$  and process  $\sigma$ . Do either of these values suggest a change in the process center and spread?

## Comments on statistical control

Having seen how  $\bar{x}$  and  $s$  (or  $\bar{x}$  and  $R$ ) charts work, we can turn to some important comments and cautions about statistical control in practice.

### Focus on the process rather than on the product

This is perhaps the fundamental idea in statistical process control. We might attempt to attain high quality by careful inspection of the finished product and reviewing every outgoing invoice and expense account payment. Inspection of finished products can ensure good quality, but it is expensive.

Perhaps more important, final inspection often comes too late: when something goes wrong early in a process, much bad product may be produced before final inspection discovers the problem. This adds to the expense, because the bad product must then be scrapped or reworked.

The small samples that are the basis of control charts are intended to monitor the process at key points, not to ensure the quality of the particular items in the samples. If the process is kept in control, we know what to expect in the finished product. We want to do it right the first time, not inspect and fix finished product.

Choosing the “key points” at which we will measure and monitor the process is important. The choice requires that you understand the process well enough to know where problems are likely to arise. Flowcharts and cause-and-effect diagrams can help. It should be clear that control charts that monitor only the final output are

often *not* the best choice.

## Rational subgroups

The interpretation of control charts depends on the distinction between  $\bar{x}$ -type special causes and  $s$ -type special causes. This distinction in turn depends on how we choose the samples from which we calculate  $s$  (or  $R$ ). We want the variation *within* a sample to reflect only the item-to-item chance variation that (when in control) results from many small common causes. Walter Shewhart, the founder of statistical process control, used the term **rational subgroup** to emphasize that we should think about the process when deciding how to choose samples.

rational subgroup

### EXAMPLE

#### 17.12 Selecting the sample.

A pharmaceutical manufacturer forms tablets by compressing a granular material that contains the active ingredient and various fillers. To monitor the compression process, we will measure the hardness of a sample from each 10 minutes' production of tablets. Should we choose a random sample of tablets from the several thousand produced in a 10-minute period?

A random sample would contain tablets spread across the entire 10 minutes. It fairly represents the 10-minute period, but that isn't what we want for process control. If the setting of the press drifts or a new lot of filler arrives during the 10 minutes, the spread of the sample will be increased. That is, a random sample contains both the short-term variation among tablets produced in quick succession and the longer-term variation among tablets produced minutes apart. We prefer to measure a rational subgroup of 5 consecutive tablets every 10 minutes. We expect the process to be stable during this very short time period, so that variation within the subgroups is a benchmark against which we can see special cause variation.

Samples of consecutive items are rational subgroups when we are monitoring the output of a single activity that does the same thing over and over again. Several consecutive items is the most common type of sample for process control.

When the stream of product contains output from several machines or several people, however, the choice of samples is more complicated. Do you want to

include variation due to different machines or different people within your samples? If you decide that this variation is common cause variation, be sure that the sample items are spread across machines or people. If all the items in each sample have a common origin,  $s^{\bar{}}$  will be small and the control limits for the  $\bar{x}$  chart will be narrow. Points on the  $\bar{x}$  chart from samples representing different machines or different people will often be out of control, some high and some low.



*There is no formula for deciding how to form rational subgroups. You must think about causes of variation in your process and decide which you are willing to think of as common causes that you will not try to eliminate.* Rational subgroups are samples chosen to express variation due to these causes and no others. Because the choice requires detailed process knowledge, we will usually accept samples of consecutive items as being rational subgroups. Just remember that real processes are messier than textbooks suggest.

### Why statistical control is desirable

To repeat, if the process is kept in control, we know what to expect in the finished product. The process mean  $\mu$  and standard deviation  $\sigma$  remain stable over time, so (assuming Normal variation) the 99.7 part of the 68–95–99.7 rule tells us that almost all measurements on individual products will lie in the range  $\mu \pm 3\sigma$ . These are sometimes called the **natural tolerances** for the product. Be careful to distinguish  $\mu \pm 3\sigma$ , the range we expect for *individual measurements*, from the  $\bar{x}$  chart control limits  $\mu \pm 3\sigma/n$  which mark off the expected range of *sample means*.

natural tolerances

### EXAMPLE

#### 17.13 Estimating the tolerances for the water resistance study.

The process of waterproofing the jackets has been operating in control. The  $\bar{x}$  and  $s$  charts were based on  $\mu = 2750$  mm and  $\sigma = 430$  mm. The  $s$  chart in Figure 17.7 and a calculation (see Exercise 17.35, page 17-37) suggest that the process  $\sigma$  is now less than 430 mm. We may prefer to calculate the natural

tolerances from the recent data on 20 samples (80 jackets) in Table 17.1. The estimate of the mean is  $\bar{x}=2750.7$ , very close to the target value.

Now a subtle point arises. The estimate  $\hat{\sigma}=s/c_4$  used for past-data control charts is based entirely on variation *within the samples*. That's what we want for control charts, because within-sample variation is likely to be "pure common cause" variation.

Even when the process is in control, there is some additional variation from sample to sample, just by chance. So the variation in the process output will be greater than the variation within samples. *To estimate the natural tolerances, we should estimate  $\sigma$  from all 80 individual jackets rather than by averaging the 20 within-sample standard deviations.* The standard deviation for all 80 jackets is

$$s = 383.8$$

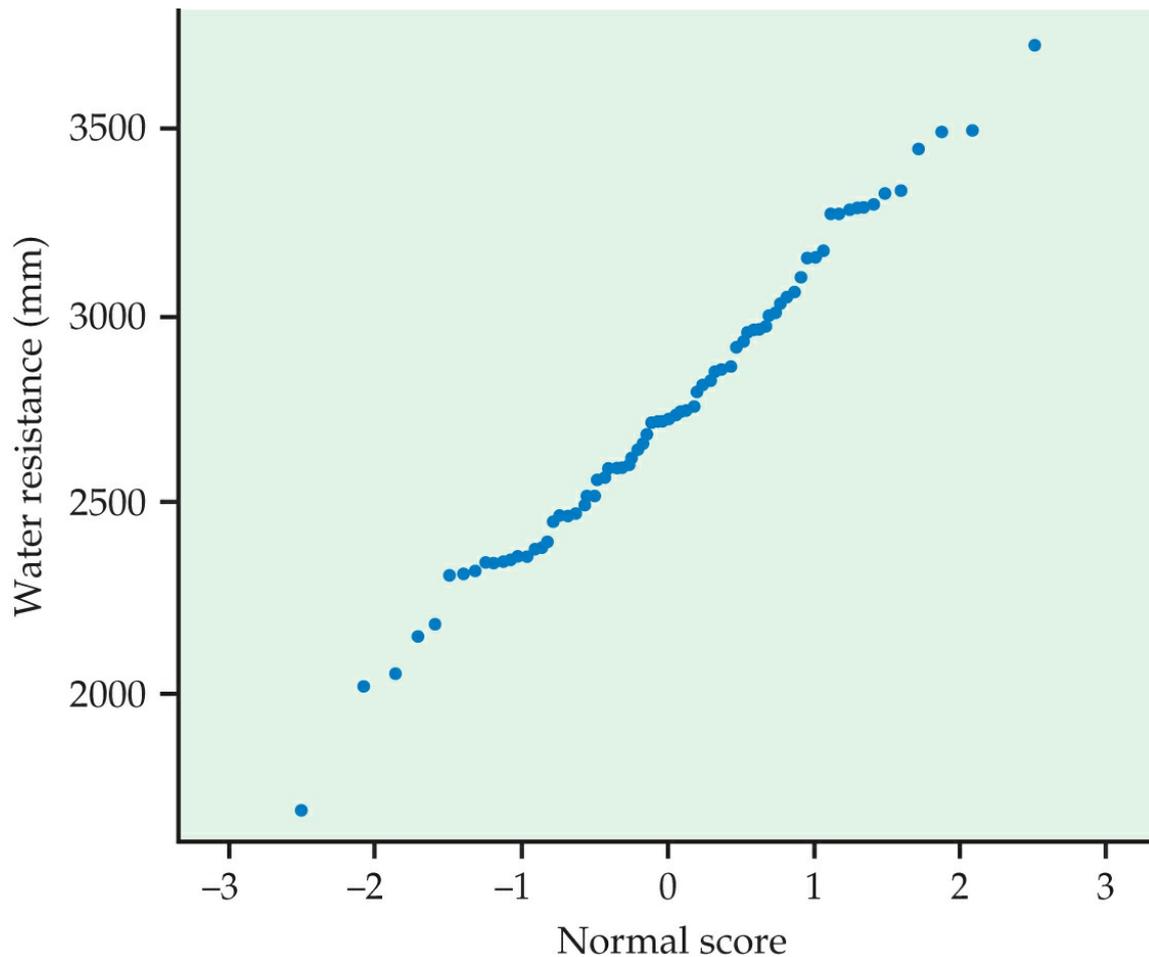
For a sample of size 80,  $c_4$  is very close to 1, so we can ignore it. We are therefore confident that almost all individual jackets will have a water resistance reading between

$$\bar{x} \pm 3s = 2750.7 \pm (3)(383.8) = 2750.7 \pm 1151.4$$

We expect water resistance measurements to vary between 1599 and 3902 mm. You see that the spread of individual measurements is wider than the spread of sample means used for the control limits of the  $\bar{x}$  chart.

The natural tolerances in Example 17.13 depend on the fact that the water resistance of individual jackets follows a Normal distribution. We know that the process was in control when the 80 measurements in Table 17.1 were made, so we can use them to assess Normality. Figure 17.15 is a Normal quantile plot of these measurements. There are no strong deviations from Normality. All 80 observations, including the one point that may appear suspiciously low in Figure 17.15, lie within the natural tolerances. Examining the data strengthens our confidence in the natural tolerances.

Because we can predict the performance of the waterproofing process, we can tell the buyers of our jackets what to expect. What is more, if a process is in control, we can see the effect of any changes we make. A process operating out of control is erratic. We can't do reliable statistical studies on such a process, and if we make a change in the process, we can't clearly see the results of the change—they are hidden by erratic special cause variation. If we want to improve a process, we must first bring it into control so that we have a stable starting point from which to improve.



**FIGURE 17.15**

Normal quantile plot for the 80 water resistance measurements of Table 17.1. Calculations about individual measurements, such as natural tolerances, depend on approximate Normality.

### Don't confuse control with capability!



A process in control is stable over time and we know how much variation the finished product will show. Control charts are, so to speak, the voice of the process telling us what state it is in. *There is no guarantee that a process in control produces products of satisfactory quality.* “Satisfactory quality” is measured by comparing the product to some standard outside the process, set by technical specifications, customer expectations, or the goals of the organization. These external standards are unrelated to the internal state of the process, which is all that statistical control pays attention to.

## CAPABILITY

**Capability** refers to the ability of a process to meet or exceed the requirements placed on it.

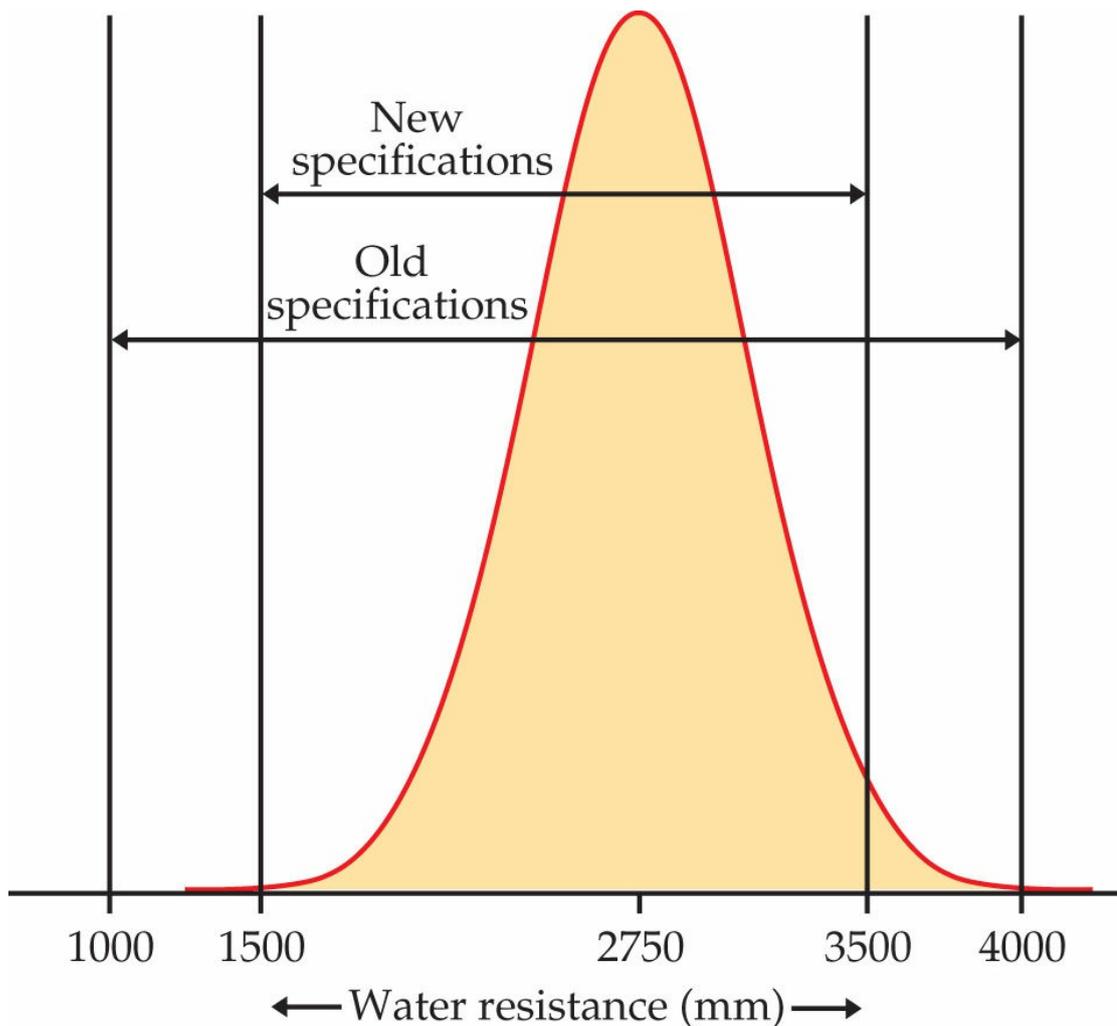
Capability has nothing to do with control—except for the very important point that if a process is not in control, it is hard to tell if it is capable or not.

## **EXAMPLE**

### **17.14 Assessing the capability of the waterproofing process.**

An outfitting company is a large buyer of this jacket. They informed us that they need water resistance levels between 1000 and 4000 mm. Although the waterproofing process is in control, we know (Example 17.13) that almost all jackets will have water resistance levels between 1599 and 3902 mm. The process is capable of meeting the customer's requirement.

Figure 17.16 compares the distribution of water resistance levels for individual jackets with the customer specifications. The distribution of water resistance is approximately Normal, and we estimate its mean to be very close to 2750 mm and the standard deviation to be about 384 mm. The distribution is safely within the specifications.



**FIGURE 17.16**

Comparison of the distribution of water resistance (Normal curve) with original and tightened specifications, for Example 17.14. The process in its current state is not capable of meeting the new specifications.

Times change, however. The outfitting company demands more similarity in jackets and decides to require that the water resistance level lie between 1500 and 3500 mm. These new specification limits also appear in Figure 17.16. The process is not capable of meeting the new requirements. The process remains in control. The change in its capability is entirely due to a change in external requirements.

Because the waterproofing process is in control, we know that it is not capable of meeting the new specifications. That's an advantage of control, but the fact remains that control does not guarantee capability. We will discuss numerical measures of capability in Section 17.3.

Managers must understand, that *if a process that is in control does not have adequate capability, fundamental changes in the process are needed.* The process is doing as well as it can and displays only the chance variation that is natural to its present state. Slogans to encourage the workers or disciplining the workers for

poor performance will not change the state of the process. Better training for workers is a change in the process that may improve capability. New equipment or more uniform material may also help, depending on the findings of a careful investigation.

## SECTION 17.2 Summary

An **R chart** based on the **range** of observations in a sample is often used in place of an  $s$  chart. Interpret  $\bar{x}$  and  $R$  charts exactly as you would interpret  $\bar{x}$  and  $s$  charts.

It is common to use **out-of-control rules** in addition to “one point outside the control limits.” In particular, a **runs rule** for the  $\bar{x}$  chart allows the chart to respond more quickly to a gradual drift in the process center.

**Control charts based on past data** are used at the **chart setup** stage for a process that may not be in control. Start with control limits calculated from the same past data that you are plotting. Beginning with the  $s$  chart, narrow the limits as you find special causes, and remove the points influenced by these causes. When the remaining points are in control, use the resulting limits to monitor the process.

Statistical process control maintains quality more economically than inspecting the final output of a process. Samples that are **rational subgroups** are important to effective control charts. A process in control is stable, so that we can predict its behavior. If individual measurements have a Normal distribution, we can give the **natural tolerances**.

A process is **capable** if it can meet the requirements placed on it. Control (stability over time) does not in itself imply capability. Remember that control describes the internal state of the process, whereas capability relates the state of the process to external specifications.

## SECTION 17.2 Exercises

*For Exercise 17.26, see page 17-24; for Exercises 17.27 and 17.28, see page 17-26; and for Exercises 17.29 and 17.30, see page 17-31.*

### 17.31 Setting up a control chart.

In Exercise 17.12 (page 17-18) the  $\bar{x}$  and  $s$  control charts for the placement of the rum label were based on historical results. Suppose that a new labeling machine has been purchased and new control limits need to be determined. Table 17.7 contains the means and standard deviations of the first 24 batch samples. We will use these to determine tentative control limits.  LABEL

- Estimate the center line and control limits for the  $s$  chart using all 24 samples.
- Does the variation within samples appear to be in control? If not, remove any out-of-control samples and recalculate the limits. We'll assume that any out-of-control samples are due to the operators adjusting to the new machine.

(c) Using the remaining samples, estimate the center line and control limits for the  $\bar{x}$  chart. Again remove any out-of-control samples and recalculate.

(d) How do these control limits compare with the ones obtained in Exercise 17.12?

### 17.32 Setting up another control chart.

Refer to the previous exercise. Table 17.8 contains another set of 24 samples. Repeat parts (a) to (c) using this data set.  LABEL1

Sample	$\bar{x}$	$s$	Sample	$\bar{x}$	$s$
1	1.9824	0.0472	13	1.9949	0.0964
2	2.0721	0.0479	14	2.0287	0.0607
3	2.0031	0.0628	15	1.9391	0.0481
4	2.0088	0.1460	16	1.9801	0.1133
5	2.0445	0.0850	17	1.9991	0.0482
6	2.0322	0.0676	18	1.9834	0.0572
7	2.0209	0.0651	19	2.0348	0.0734
8	1.9927	0.1291	20	1.9935	0.0584
9	2.0164	0.0889	21	1.9866	0.0628
10	2.0462	0.0662	22	1.9599	0.0829
11	2.0438	0.0554	23	2.0018	0.0541
12	2.0269	0.0493	24	1.9954	0.0566

Sample	$\bar{x}$	$s$	Sample	$\bar{x}$	$s$
1	2.0309	0.1661	13	1.9907	0.0620
2	2.0066	0.1366	14	1.9612	0.0748
3	2.0163	0.0369	15	2.0312	0.0421
4	2.0970	0.1088	16	2.0293	0.0932
5	1.9499	0.0905	17	1.9758	0.0252
6	1.9859	0.1683	18	2.0255	0.0728
7	1.9456	0.0920	19	1.9574	0.0186
8	2.0213	0.0478	20	2.0320	0.0151
9	1.9621	0.0489	21	1.9775	0.0294
10	1.9529	0.0456	22	1.9612	0.0911
11	1.9995	0.0519	23	2.0042	0.0365
12	1.9927	0.0762	24	1.9933	0.0293

### 17.33 Control chart for an unusual sampling situation.

Invoices are processed and paid by two clerks, one very experienced and the other newly hired. The

experienced clerk processes invoices quickly. The new hire often refers to the procedures handbook and is much slower. Both are quite consistent, so that their times vary little from invoice to invoice. Suppose that each daily sample of four invoice-processing times comes from only one of the clerks. Thus, some samples are from one and some from the other clerk. Sketch the  $\bar{x}$  chart pattern that will result.

### 17.34 Altering the sampling plan.

Refer to Exercise 17.33. Suppose instead that each sample contains an equal number of invoices from each clerk.

- Sketch the  $\bar{x}$  and  $s$  chart patterns that will result.
- The process in this case will appear in control. When might this be an acceptable conclusion?

### 17.35 Reevaluating the process parameters.

The  $\bar{x}$  and  $s$  control charts for the waterproofing example were based on  $\mu = 2750$  mm and  $\sigma = 430$  mm. Table 17.1 (page 17-10) gives the 20 most recent samples from this process.  H2ORES

- Estimate the process  $\mu$  and  $\sigma$  based on these 20 samples.
- Your calculations suggest that the process  $\sigma$  may now be less than 430 mm. Explain why the  $s$  chart in Figure 17.7 (page 17-15) suggests the same conclusion. (If this pattern continues, we would eventually update the value of  $\sigma$  used for control limits.)

### 17.36 Estimating the control chart limits from past data.

Table 17.9 gives data on the losses (in dollars) incurred by a hospital in treating DRG 209 (major joint replacement) patients.<sup>11</sup> The hospital has taken from its records a random sample of 8 such patients each month for 15 months.  DRG

- Make an  $s$  control chart using center lines and limits calculated from these past data. There are no points out of control.
- Because the  $s$  chart is in control, base the  $\bar{x}$  chart on all 15 samples. Make this chart. Is it also in control?

### 17.37 Efficient process control.

A company that makes cellular phones requires that their microchip supplier practice statistical process control and submit control charts for verification. This allows the company to eliminate inspection of the microchips as they arrive, a considerable cost savings. Explain carefully why incoming inspection can safely be eliminated.

### 17.38 Determining the tolerances for losses from DRG 209 patients.

Table 17.9 gives data on hospital losses for samples of DRG 209 patients. The distribution of losses has been stable over time. What are the natural tolerances within which you expect losses on nearly all such patients to fall?  DRG

### 17.39 Checking the Normality of losses.

Do the losses on the 120 individual patients in Table 17.9 appear to come from a single Normal distribution? Make a Normal quantile plot and discuss what it shows. Are the natural tolerances you found in the previous exercise trustworthy? Explain your answer.  DRG

### 17.40 The percent of products that meet specifications.

If the water resistance readings of individual jackets follow a Normal distribution, we can describe capability by giving the percent of jackets that meet specifications. The old specifications for water resistance are 1000 to 4000 mm. The new specifications are 1500 to 3500 mm. Because the process is in control, we can estimate (Example 17.13) that water resistance has mean 2750 mm and standard deviation 384 mm.  H2ORES

**TABLE 17.9** Hospital Losses for 15 Samples of DRG 209 Patients

Sample	Loss (dollars)								Sample mean	Standard deviation
1	6835	5843	6019	6731	6362	5696	7193	6206	6360.6	521.7
2	6452	6764	7083	7352	5239	6911	7479	5549	6603.6	817.1
3	7205	6374	6198	6170	6482	4763	7125	6241	6319.8	749.1
4	6021	6347	7210	6384	6807	5711	7952	6023	6556.9	736.5
5	7000	6495	6893	6127	7417	7044	6159	6091	6653.2	503.7
6	7783	6224	5051	7288	6584	7521	6146	5129	6465.8	1034.3
7	8794	6279	6877	5807	6076	6392	7429	5220	6609.2	1104.0
8	4727	8117	6586	6225	6150	7386	5674	6740	6450.6	1033.0
9	5408	7452	6686	6428	6425	7380	5789	6264	6479.0	704.7
10	5598	7489	6186	5837	6769	5471	5658	6393	6175.1	690.5
11	6559	5855	4928	5897	7532	5663	4746	7879	6132.4	1128.6
12	6824	7320	5331	6204	6027	5987	6033	6177	6237.9	596.6
13	6503	8213	5417	6360	6711	6907	6625	7888	6828.0	879.8
14	5622	6321	6325	6634	5075	6209	4832	6386	5925.5	667.8
15	6269	6756	7653	6065	5835	7337	6615	8181	6838.9	819.5

- (a) What percent of jackets meet the old specifications?
- (b) What percent meet the new specifications?

### 17.41 Improving the capability of the process.

Refer to the previous exercise. The center of the specifications for waterproofing is 2500 mm, but the center of our process is 2750 mm. We can improve capability by adjusting the process to have center 2500 mm. This is an easy adjustment that does not change the process variation. What percent of jackets now meet the new specifications?

### 17.42 Monitoring the calibration of a densitometer.

Loss of bone density is a serious health problem for many people, especially older women. Conventional X-rays often fail to detect loss of bone density until the loss reaches 25% or more. New equipment such as the Lunar bone densitometer is much more sensitive. A health clinic installs one of these machines. The manufacturer supplies a “phantom,” an aluminum piece of known density that can be used to keep the machine calibrated. Each morning, the clinic makes two measurements on the phantom before measuring the first patient. Control charts based on these measurements alert the operators if the machine has lost calibration. Table 17.10 contains data for the first 30 days of operation.<sup>12</sup> The units are grams per square centimeter (for technical reasons, area rather than volume is measured).  **DENSITY**

- (a) Calculate  $\bar{x}$  and  $s$  for the first 2 days to verify the table entries for those quantities.
- (b) What kind of variation does the  $s$  chart monitor in this setting? Make an  $s$  chart and comment on control. If any points are out of control, remove them and recompute the chart limits until all remaining points are in control. (That is, assume that special causes are found and removed.)
- (c) Make an  $\bar{x}$  chart using the samples that remain after you have completed part (b). What kind of variation will be visible on this chart? Comment on the stability of the machine over these 30 days based on both charts.

### 17.43 Determining the natural tolerances for the distance between holes.

Figure 17.10 (page 17-22) displays a record sheet for 18 samples of distances between mounting holes in an electrical meter. In Exercise 17.21 (page 17-21), you found that Sample 5 was out of control on the process-monitoring  $s$  chart. The special cause responsible was found and removed. Based on the 17 samples that were in control, what are the natural tolerances for the distance between the holes?  **MOUNT**

### 17.44 Determining the natural tolerances for the densitometer.

Remove any samples in Table 17.10 that your work in Exercise 17.42 showed to be out of control on either chart. Estimate the mean and standard deviation of individual measurements on the phantom. What are the natural tolerances for these measurements?  **DENSITY**

### 17.45 Determining the percent of meters that meet specifications.

The record sheet in Figure 17.10 gives the specifications as  $0.6054 \pm 0.0010$  inch. That’s  $54 \pm 10$  as the data are coded on the record. Assuming that the distance varies Normally from meter to meter, about what percent of meters meet the specifications?  **DENSITY**

<b>TABLE 17.10</b>		<b>Daily Calibration Samples for a Lunar Bone Densitometer</b>		
<b>Day</b>	<b>Measurements (g/cm<sup>2</sup>)</b>		<b><math>\bar{x}</math></b>	<b><math>s</math></b>
1	1.261	1.260	1.2605	0.000707
2	1.261	1.268	1.2645	0.004950
3	1.258	1.261	1.2595	0.002121
4	1.261	1.262	1.2615	0.000707
5	1.259	1.262	1.2605	0.002121
6	1.269	1.260	1.2645	0.006364

7	1.262	1.263	1.2625	0.000707
8	1.264	1.268	1.2660	0.002828
9	1.258	1.260	1.2590	0.001414
10	1.264	1.265	1.2645	0.000707
11	1.264	1.259	1.2615	0.003536
12	1.260	1.266	1.2630	0.004243
13	1.267	1.266	1.2665	0.000707
14	1.264	1.260	1.2620	0.002828
15	1.266	1.259	1.2625	0.004950
16	1.257	1.266	1.2615	0.006364
17	1.257	1.266	1.2615	0.006364
18	1.260	1.265	1.2625	0.003536
19	1.262	1.266	1.2640	0.002828
20	1.265	1.266	1.2655	0.000707
21	1.264	1.257	1.2605	0.004950
22	1.260	1.257	1.2585	0.002121
23	1.255	1.260	1.2575	0.003536
24	1.257	1.259	1.2580	0.001414
25	1.265	1.260	1.2625	0.003536
26	1.261	1.264	1.2625	0.002121
27	1.261	1.264	1.2625	0.002121
28	1.260	1.262	1.2610	0.001414
29	1.260	1.256	1.2580	0.002828
30	1.260	1.262	1.2610	0.001414

### 17.46 Assessing the Normality of the densitometer measurements.

Are the 60 individual measurements in Table 17.10 at least approximately Normal, so that the natural tolerances you calculated in Exercise 17.44 can be trusted? Make a Normal quantile plot (or another graph if your software is limited) and discuss what you see.  DENSITY

### 17.47 Assessing the Normality of the distance between holes.

Make a Normal quantile plot of the 85 distances in the data file MOUNT that remain after removing Sample 5. How does the plot reflect the limited precision of the measurements (all of which end in 4)? Is there any departure from Normality that would lead you to discard your conclusion from Exercise 17.43? (If your software will not make Normal quantile plots, use a histogram to assess Normality.)  MOUNT

### 17.48 Determining the natural tolerances for the weight of ground beef.

Table 17.3 (page 17-19) gives data on the weight of ground beef sections. Since the distribution of weights has been stable, use the data in Table 17.3 to construct the natural tolerances within which you expect almost all the weights to fall.  MEATWGT

### 17.49 Assessing the Normality of the weight measurements.

Refer to the previous exercise. Do the weights of the 60 individual sections in Table 17.3 appear to come from a single Normal distribution? Make a Normal quantile plot and discuss whether the natural tolerances you found in the previous exercise are trustworthy.  MEATWGT

### 17.50 Control charts for the bore diameter of a bearing.

A sample of 5 skateboard bearings is taken near the end of each hour of production. Table 17.11 gives  $\bar{x}$  and  $s$  for the first 21 samples, coded in units of 0.001 mm from the target value. The specifications allow a range of  $\pm 0.004$  mm about the target (a range of  $-4$  to  $+4$  as coded).  BEARINGS

- (a) Make an  $s$  chart based on past data and comment on control of short-term process variation.
- (b) Because the data are coded about the target, the process mean for the data provided is  $\mu = 0$ . Make an  $\bar{x}$  chart and comment on control of long-term process variation. What special  $\bar{x}$ -type cause probably explains the lack of control of  $\bar{x}$ ?

### 17.51 Detecting special cause variation.

Is each of the following examples of a special cause most likely to first result in (i) a sudden change in level on the  $s$  or  $R$  chart, (ii) a sudden change in level on the  $\bar{x}$  chart, or (iii) a gradual drift up or down on the  $\bar{x}$  chart? In each case, briefly explain your reasoning.

- (a) An airline pilots' union puts pressure on management during labor negotiations by asking its members to "work to rule" in doing the detailed checks required before a plane can leave the gate.
- (b) Measurements of part dimensions that were formerly made by hand are now made by a very accurate laser system. (The process producing the parts does not change—measurement methods can also affect control charts.)
- (c) Inadequate air conditioning on a hot day allows the temperature to rise during the afternoon in an office that prepares a company's invoices.

### 17.52 Deming speaks.

The following comments were made by the quality guru W. Edwards Deming (1900–1993).<sup>13</sup> Choose one of these sayings. Explain carefully what facts about improving quality the saying attempts to summarize.

- (a) "People work in the system. Management creates the system."
- (b) "Putting out fires is not improvement. Finding a point out of control, finding the special cause and removing it, is only putting the process back to where it was in the first place. It is not improvement of the process."
- (c) "Eliminate slogans, exhortations and targets for the workforce asking for zero defects and new levels of productivity."

### 17.53 Monitoring the winning times of the Boston Marathon.

The Boston Marathon has been run each year since 1897. Winning times were highly variable in the early years, but control improved as the best runners became more professional. A clear downward trend continued until the 1980s. Sam plans to make a control chart for the winning times from 1980 to the present. Calculation from the winning times from 1980 to 2013 gives

$$\bar{x} = 129.52 \text{ minutes} \text{ and } s = 2.19 \text{ minutes}$$

Sam draws a center line at  $\bar{x}$  and control limits at  $\bar{x} \pm 3s$  for a plot of individual winning times. Explain carefully why these control limits are too wide to effectively signal unusually fast or slow times.

**TABLE 17.11**  $\bar{x}$  and  $s$  for Samples of Bore Diameter

Sample	$\bar{x}$	$s$	Sample	$\bar{x}$	$s$
1	0.0	1.225	12	0.8	3.899
2	0.4	1.517	13	2.0	1.581
3	0.6	2.191	14	0.2	2.049
4	1.0	3.162	15	0.6	2.302
5	-0.8	2.280	16	1.2	2.588
6	-1.0	2.345	17	2.8	1.924
7	1.6	1.517	18	2.6	3.130
8	1.0	1.414	19	1.8	2.387
9	0.4	2.608	20	0.2	2.775
10	1.4	2.608	21	1.6	1.949
11	0.8	1.924			

### 17.54 Monitoring weight.

Joe has recorded his weight, measured at the gym after a workout, for several years. The mean is 181 pounds and the standard deviation is 1.7 pounds, with no signs of lack of control. An injury keeps Joe away from the gym for several months. The data below give his weight, measured once each week for the first 16 weeks after he returns from the injury:

Week	1	2	3	4	5	6	7	8
Weight	185.2	185.5	186.3	184.3	183.1	180.8	183.8	182.1
Week	9	10	11	12	13	14	15	16
Weight	181.1	180.1	178.7	181.2	183.1	180.2	180.8	182.2

Joe wants to plot these individual measurements on a control chart. When each “sample” is just one measurement, short-term variation is estimated by advanced techniques.<sup>14</sup> The short-term variation in Joe’s weight is estimated to be about  $\sigma = 1.6$  pounds. Joe has a target of  $\mu = 181$  pounds for his weight. Make a control chart for his measurements, using control limits  $\mu \pm 2\sigma$ . It is common to use these narrower limits on an “individuals chart.” Comment on individual points out of control and on runs. Is Joe’s weight stable or does it change systematically over this period?  **JOEWGT**

## 17.3 Process Capability Indexes

When you complete this section, you will be able to

- Estimate the percent of product that meets specifications using the Normal distribution.
- Explain why the percent of product meeting specifications is not a good measure of capability.
- Compute and interpret the  $C_p$  and  $C_{pk}$  capability indexes.
- Identify issues that affect the interpretation of capability indexes.

Capability describes the quality of the output of a process relative to the needs or requirements of the users of that output. To be more precise, capability relates the *actual performance* of a process in control, after special causes have been removed, to the *desired* performance.

Suppose, to take a simple but common setting, that there are *specifications* set for some characteristic of the process output. The viscosity of the elastomer in Example 17.11 (page 17-26) is supposed to be  $45 \pm 5$  Mooneys. The speed with which calls are answered at a corporate customer service call center is supposed to be no more than 30 seconds.

In this setting, we might measure capability by the *percent of output that meets the specifications*. When the variable we measure has a Normal distribution, we can estimate this percent using the mean and standard deviation estimated from past control chart samples. When the variable is not Normal, we can use the actual percent of the measurements in the samples that meet the specifications.

### EXAMPLE

#### 17.15 What is the probability of meeting specifications?

(a) Before concluding the process improvement study begun in Example 17.11, we found and fixed special causes and eliminated from our data the samples on which those causes operated. The remaining viscosity measurements have  $\bar{x} = 48.7$  and  $s = 0.85$ . Note once again that to draw conclusions about viscosity for individual lots we estimate the standard

deviation  $\sigma$  from all individual lots, not from the average  $\bar{s}$  of sample standard deviations.

The specifications call for the viscosity of the elastomer to lie in the range  $45 \pm 5$ . A Normal quantile plot shows the viscosities to be quite Normal. Figure 17.17(a) shows the Normal distribution of lot viscosities with the specification limits  $45 \pm 5$ . These are marked **LSL** for *lower specification limit* and **USL** for *upper specification limit*. The percent of lots that meet the specifications is about

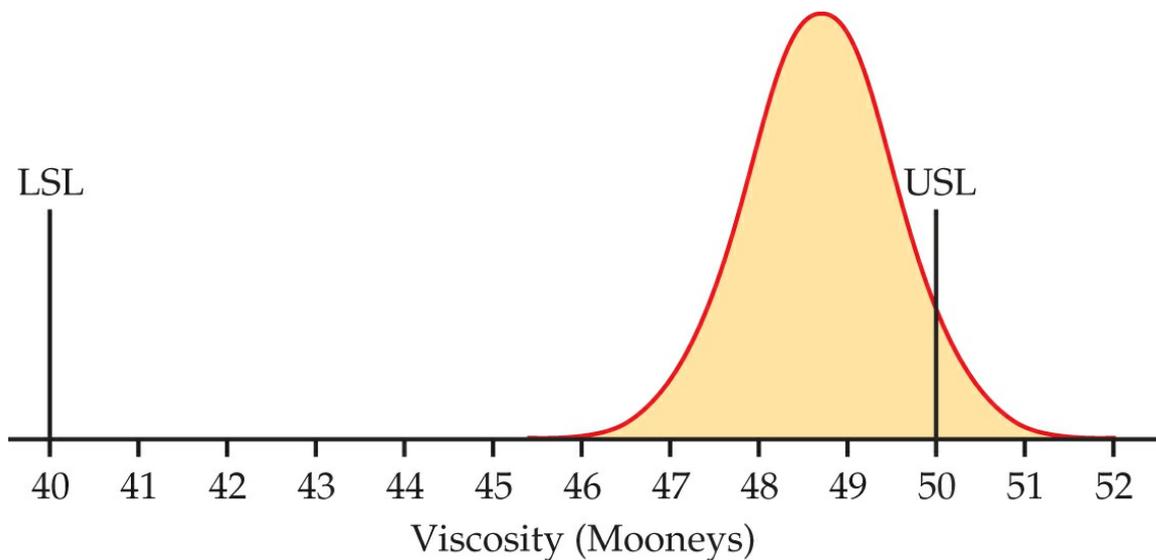
LSL

USL

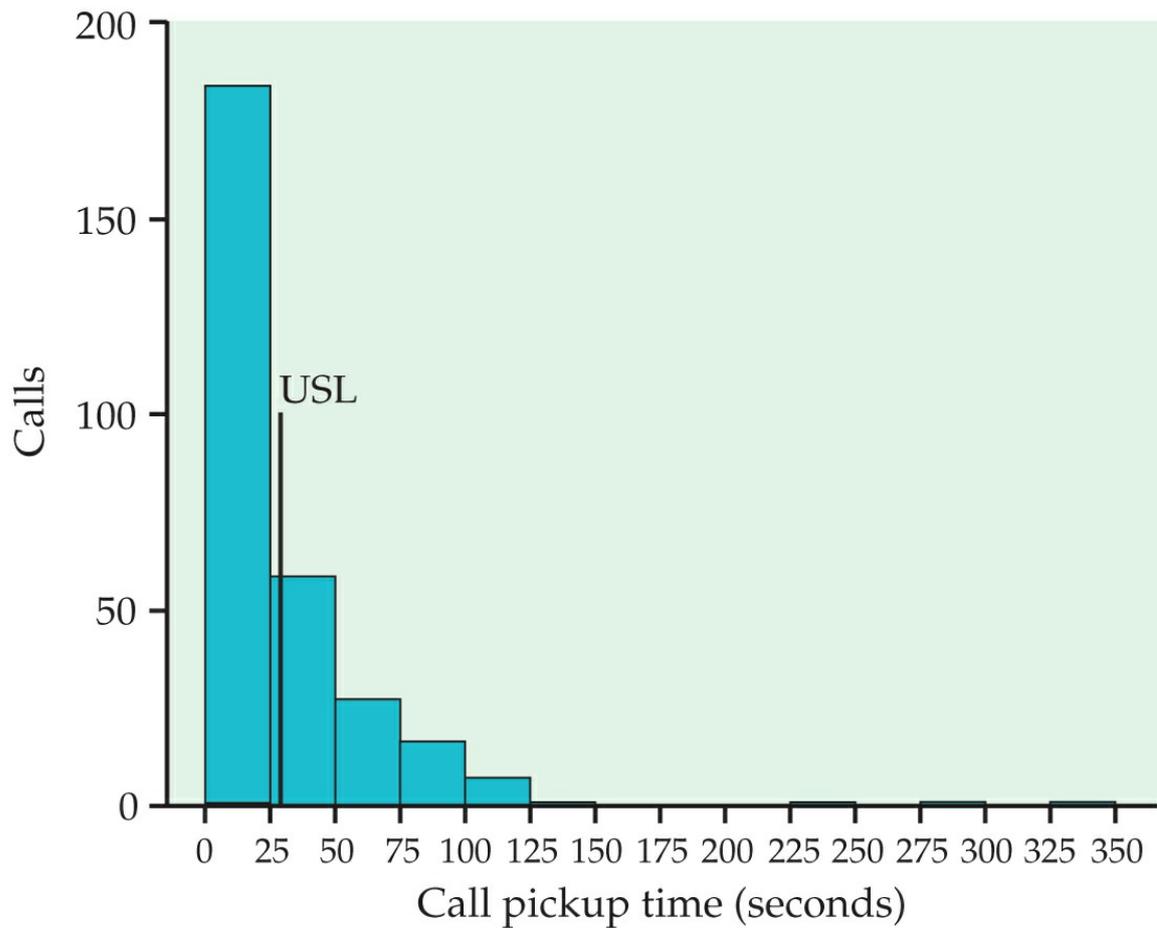
$$\begin{aligned} P(40 \leq \text{viscosity} \leq 50) &= P(40 - 48.7085 \leq Z \leq 50 - 48.7085) \\ &= P(-10.2 \leq Z \leq 1.53) = 0.937 \end{aligned}$$

Roughly 94% of the lots meet the specifications. If we can adjust the process center to the center of the specifications,  $\mu = 45$ , it is clear from Figure 17.17(a) that essentially 100% of lots will meet the specifications.

(b) Times to answer calls to a corporate customer service center are usually right-skewed. Figure 17.17(b) is a histogram of the times for 300 calls to the call center of a large bank.<sup>15</sup> The specification limit of 30 seconds is marked USL. The median is 20 seconds, but the mean is 32 seconds. Of the 300 calls, 203 were answered in no more than 30 seconds. That is,  $203/300 = 68\%$  of the times meet the specifications.



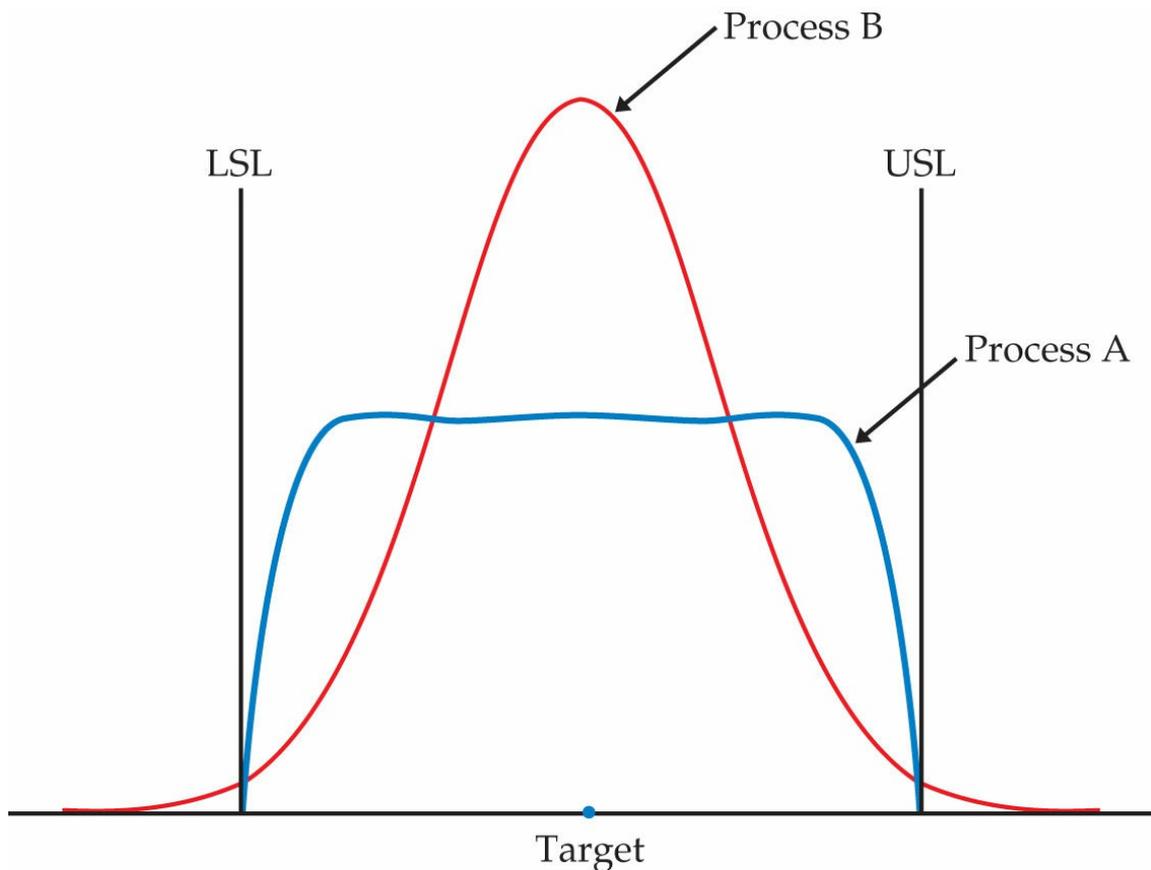
(a)



(b)

**FIGURE 17.17**

Comparing distributions of individual measurements with specifications for, Example 17.15. (a) Viscosity has a Normal distribution. The capability is poor but will be good if we can properly center the process. (b) Response times to customer calls have a right-skewed distribution and only an upper specification limit. Capability is again poor.



**FIGURE 17.18**

Two distributions for part diameters. All the parts from Process A meet the specifications, but a higher proportion of parts from Process B have diameters close to the target.



*Turns out, however, that the percent meeting specifications is a poor measure of capability.* Figure 17.18 shows why. This figure compares the distributions of the diameter of the same part manufactured by two processes. The target diameter and the specification limits are marked. All the parts produced by Process A meet the specifications, but about 1.5% of those from Process B fail to do so.

Nonetheless, Process B appears superior to Process A because it is less variable: much more of Process B's output is close to the target. Process A produces many parts close to LSL and USL. These parts meet the specifications, but they will likely fit and perform more poorly than parts with diameters close to the center of the specifications. A distribution like that for Process A might result from inspecting all the parts and discarding those whose diameters fall outside the specifications. That's not an efficient way to achieve quality.

We need a way to measure process capability that pays attention to the variability of the process (smaller is better). The standard deviation does that, but it doesn't measure capability because it takes no account of the specifications that the output must meet.

*Capability indexes* start with the idea of comparing process variation with the specifications. Process B will beat Process A by such a measure. Capability indexes also allow us to measure process improvement—we can continue to drive down variation, and so improve the process, long after 100% of the output meets specifications. Continual improvement of processes is our goal, not just reaching “satisfactory” performance. The real importance of capability indexes is that they give us numerical measures to describe ever-better process quality.

## The capability indexes $C_p$ and $C_{pk}$

Capability indexes are numerical measures of process capability that, unlike percent meeting specifications, have no upper limit such as 100%. We can use capability indexes to measure continuing improvement of a process. Of course, reporting just one number has limitations. What is more, the usual indexes are based on thinking about Normal distributions. They are not meaningful for distinctly non-Normal output distributions like the call center response times in Figure 17.17(b).

### CAPABILITY INDEXES

Consider a process with specification limits LSL and USL for some measured characteristic of its output. The process mean for this characteristic is  $\mu$  and the standard deviation is  $\sigma$ . The **capability index**  $C_p$  is

$$C_p = \frac{USL - LSL}{6\sigma}$$

The **capability index**  $C_{pk}$  is

$$C_{pk} = \frac{\min\{\mu - \text{lower spec limit}, \text{upper spec limit} - \mu\}}{3\sigma}$$

Set  $C_{pk} = 0$  if the process mean  $\mu$  lies outside the specification limits. Large values of  $C_p$  or  $C_{pk}$  indicate more capable processes.

Capability indexes start from the fact that *Normal distributions are in practice about 6 standard deviations wide*. That’s the 99.7 part of the 68–95–99.7 rule. Conceptually,  $C_p$  is the specification width as a multiple of the process width  $6\sigma$ . When  $C_p = 1$ , the process output will just fit within the specifications if the center is midway between LSL and USL.

Larger values of  $C_p$  are better—the process output can fit within the specs with room to spare. But a process with high  $C_p$  can produce poor-quality product if it is not correctly centered.

$C_{pk}$  remedies this deficiency by considering both the center  $\mu$  and the variability

$\sigma$  of the measurements. The denominator  $3\sigma$  in  $C_{pk}$  is half the process width. It is the space needed on either side of the mean if essentially all the output is to lie between LSL and USL. When  $C_{pk} = 1$ , the process has just this much space between the mean and the nearer of LSL and USL. Again, higher values are better.  $C_{pk}$  is the most common capability index, but starting with  $C_p$  helps us see how the indexes work.

## EXAMPLE

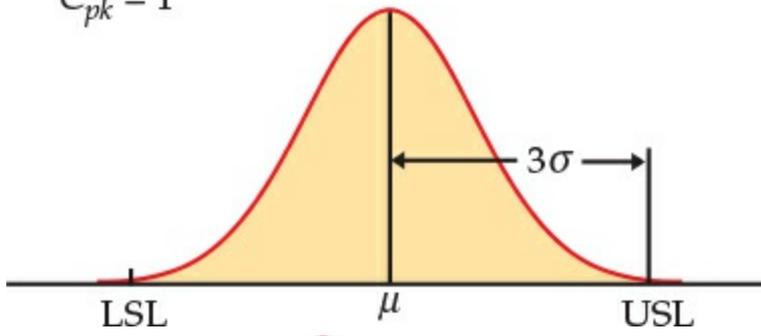
### 17.16 A comparison of the $C_p$ and $C_{pk}$ indexes.

Consider the series of pictures in Figure 17.19. We might think of a process that machines a metal part. Measure a dimension of the part that has LSL and USL as its specification limits. As usual, there is variation from part to part. The dimensions vary Normally with mean  $\mu$  and standard deviation  $\sigma$ .

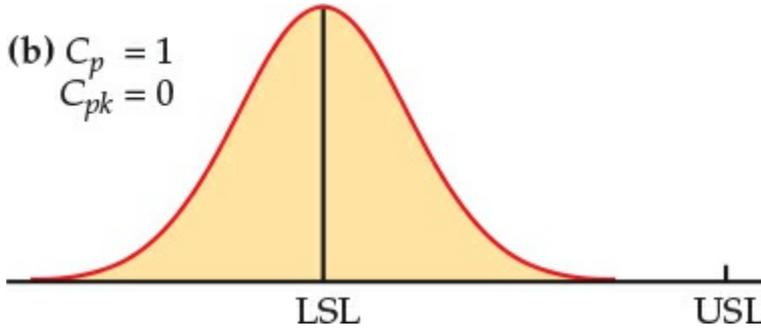
Figure 17.19(a) shows process width equal to the specification width. That is,  $C_p = 1$ . Almost all the parts will meet specifications *if*, as in this figure, the process mean  $\mu$  is at the center of the specs. Because the mean is centered, it is  $3\sigma$  from both LSL and USL, so  $C_{pk} = 1$  also. In Figure 17.19(b), the mean has moved down to LSL. Only half the parts will meet the specifications.  $C_p$  is unchanged because the process width has not changed. But  $C_{pk}$  sees that the center  $\mu$  is right on the edge of the specifications,  $C_{pk} = 0$ . The value remains 0 if  $\mu$  moves outside the specifications.

In Figures 17.19(c) and (d), the process  $\sigma$  has been reduced to half the value it had in (a) and (b). The process width  $6\sigma$  is now half the specification width, so  $C_p = 2$ . In Figure 17.19(c) the center is just 3 of the new  $\sigma$ 's above LSL, so that  $C_{pk} = 1$ . Figure 17.19(d) shows the same smaller  $\sigma$  accompanied by mean  $\mu$  correctly centered between LSL and USL.  $C_{pk}$  rewards the process for moving the center from  $3\sigma$  to  $6\sigma$  away from the nearer limit by increasing from 1 to 2. You see that  $C_p$  and  $C_{pk}$  are equal if the process is properly centered. If not,  $C_{pk}$  is smaller than  $C_p$ .

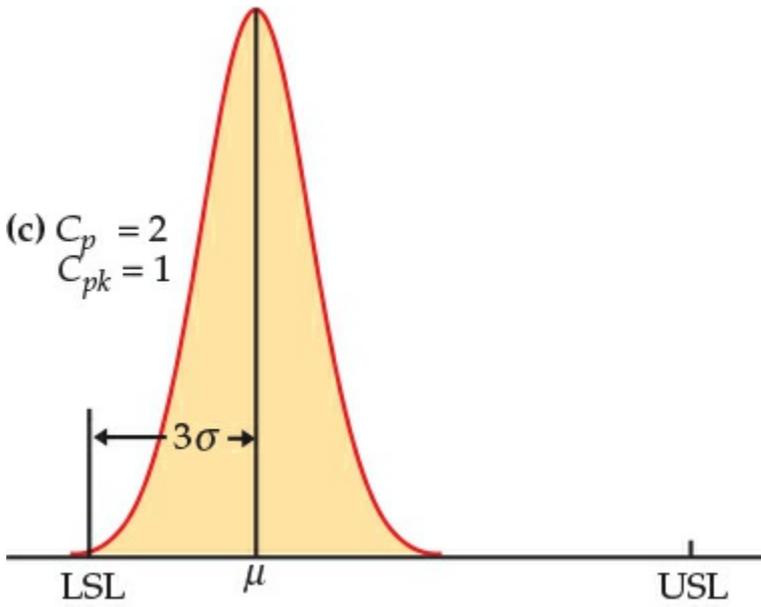
(a)  $C_p = 1$   
 $C_{pk} = 1$

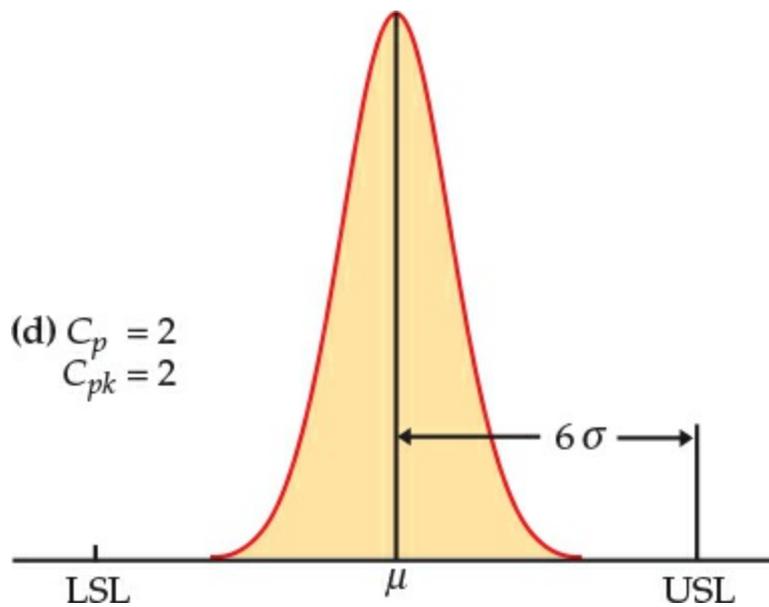


(b)  $C_p = 1$   
 $C_{pk} = 0$



(c)  $C_p = 2$   
 $C_{pk} = 1$





**FIGURE 17.19**

How capability indexes work. (a) Process centered, process width equal to specification width. (b) Process off-center, process width equal to specification width. (c) Process off-center, process width equal to half the specification width. (d) Process centered, process width equal to half the specification width.

## EXAMPLE

### 17.17 Computing $C_p$ and $C_{pk}$ for the viscosity process.

Figure 17.17(a) compares the distribution of the viscosities of lots of elastomers with the specifications  $LSL = 40$  and  $USL = 50$ . The distribution here, as is always true in practice, is *estimated* from past observations on the process. The estimates are

$$\hat{\mu} = \bar{\bar{x}} = 48.7$$

$$\hat{\sigma} = s = 0.85$$

Because capability describes the distribution of individual measurements, we once more estimate  $\sigma$  from individual measurements rather than using the estimate  $\bar{s}/c_4$  that we employ for control charts.

These estimates may be quite accurate if we have data on many past lots. Estimates based on only a few observations may, however, be inaccurate because statistics from small samples can have large sampling variability. This important point is often not appreciated when capability indexes are used in practice. To emphasize that we can only estimate the indexes, we write  $\hat{C}_p$

and  $C^{pk}$  for values calculated from sample data. They are

$$\begin{aligned}C^p &= \frac{USL - LSL}{6\sigma} \\ &= \frac{50 - 10}{6(0.85)} = 1.96 \\ C^{pk} &= \frac{|\mu - \text{nearer limit}|}{3\sigma} \\ &= \frac{50 - 48.7}{3(0.85)} = 0.51\end{aligned}$$

$C^p = 1.96$  is quite satisfactory because it indicates that the process width is only about half the specification width. The small value of  $C^{pk}$  reflects the fact that the process center is not close to the center of the specs. If we can move the center  $\mu$  to 45, then  $C^{pk}$  will also be 1.96.

## USE YOUR KNOWLEDGE

### 17.55 Specification limits versus control limits.

The manager you report to is confused by LSL and USL versus LCL and UCL. The notations look similar. Carefully explain the conceptual difference between specification limits for individual measurements and control limits for  $\bar{x}$ .

### 17.56 Interpreting the capability indexes.

Sketch Normal curves that represent measurements on products from a process with

- (a)  $C_p = 1.0$  and  $C_{pk} = 0.5$ .
- (b)  $C_p = 1.0$  and  $C_{pk} = 1.0$ .
- (c)  $C_p = 2.0$  and  $C_{pk} = 1.0$ .

## Cautions about capability indexes

Capability indexes are widely used, especially in manufacturing. Some large manufacturers even set standards, such as  $C_{pk} \geq 1.33$ , that their suppliers must meet. That is, suppliers must show that their processes are in control (through control charts) and also that they are capable of high quality (as measured by  $C_{pk}$ ). There are good reasons for requiring  $C_{pk}$ : it is a better description of process

quality than “100% of output meets specs,” and it can document continual improvement. Nonetheless, it is easy to trust  $C_{pk}$  too much. We will point to three possible pitfalls.

### How to cheat on $C_{pk}$

Estimating  $C_{pk}$  requires estimates of the process mean  $\mu$  and standard deviation  $\sigma$ . The estimates are usually based on samples measured in order to keep control charts. There is only one reasonable estimate of  $\mu$ . This is the mean  $\bar{x}$  of all measurements in recent samples, which is the same as the mean  $\bar{\bar{x}}$  of the sample means.

There are two different ways of estimating  $\sigma$ , however. The standard deviation  $s$  of all measurements in recent samples will usually be larger than the control chart estimate  $\bar{s}/c_4$  based on averaging the sample standard deviations. For  $C_{pk}$ , the proper estimate is  $s$  because we want to describe all the variation in the process output. Larger  $C_{pk}$ 's are better, and a supplier wanting to satisfy a customer can make  $C_{pk}$  a bit larger simply by using the smaller estimate  $\bar{s}/c_4$  for  $\sigma$ . That's cheating.

### Non-Normal distributions

Many business processes, and some manufacturing processes as well, give measurements that are clearly right-skewed rather than approximately Normal. Measuring the times required to deal with customer calls or prepare invoices typically gives a right-skewed distribution—there are many routine cases and a few unusual or difficult situations that take much more time. Other processes have “heavy tails,” with more measurements far from the mean than in a Normal distribution.



Process capability concerns the behavior of individual outputs, so the central limit theorem effect that improves the Normality of  $\bar{x}$  does not help us. Capability indexes are therefore more strongly affected by non-Normality than are control charts. *It is hard to interpret  $C_{pk}$  when the measurements are strongly non-Normal.* Until you gain experience, it is best to apply capability indexes only when Normal quantile plots show that the distribution is at least roughly Normal.

### Sampling variation

We know that all statistics are subject to sampling variation. If we draw another sample from the same process at the same time, we get slightly different  $\bar{x}$  and  $s$  due to the luck of the draw in choosing samples. In process control language, the samples differ due to the common cause variation that is always present.

$C_p$  and  $C_{pk}$  are in practice calculated from process data because we don't know the true process mean and standard deviation. That is, these capability indexes are statistics subject to sampling variation. A supplier under pressure from a large customer to measure  $C_{pk}$  often may base calculations on small samples from the process. The resulting estimate  $\hat{C}_{pk}$  can differ from the true process  $C_{pk}$  in either direction.

## EXAMPLE

### 17.18 Can we adequately measure $C_{pk}$ ?

Suppose that the process of waterproofing is in control at its original level. Water resistance measurements are Normally distributed with mean  $\mu = 2750$  mm and standard deviation  $\sigma = 430$  mm. The tightened specification limits are  $LSL = 1500$  and  $USL = 3500$ , so the true capability is

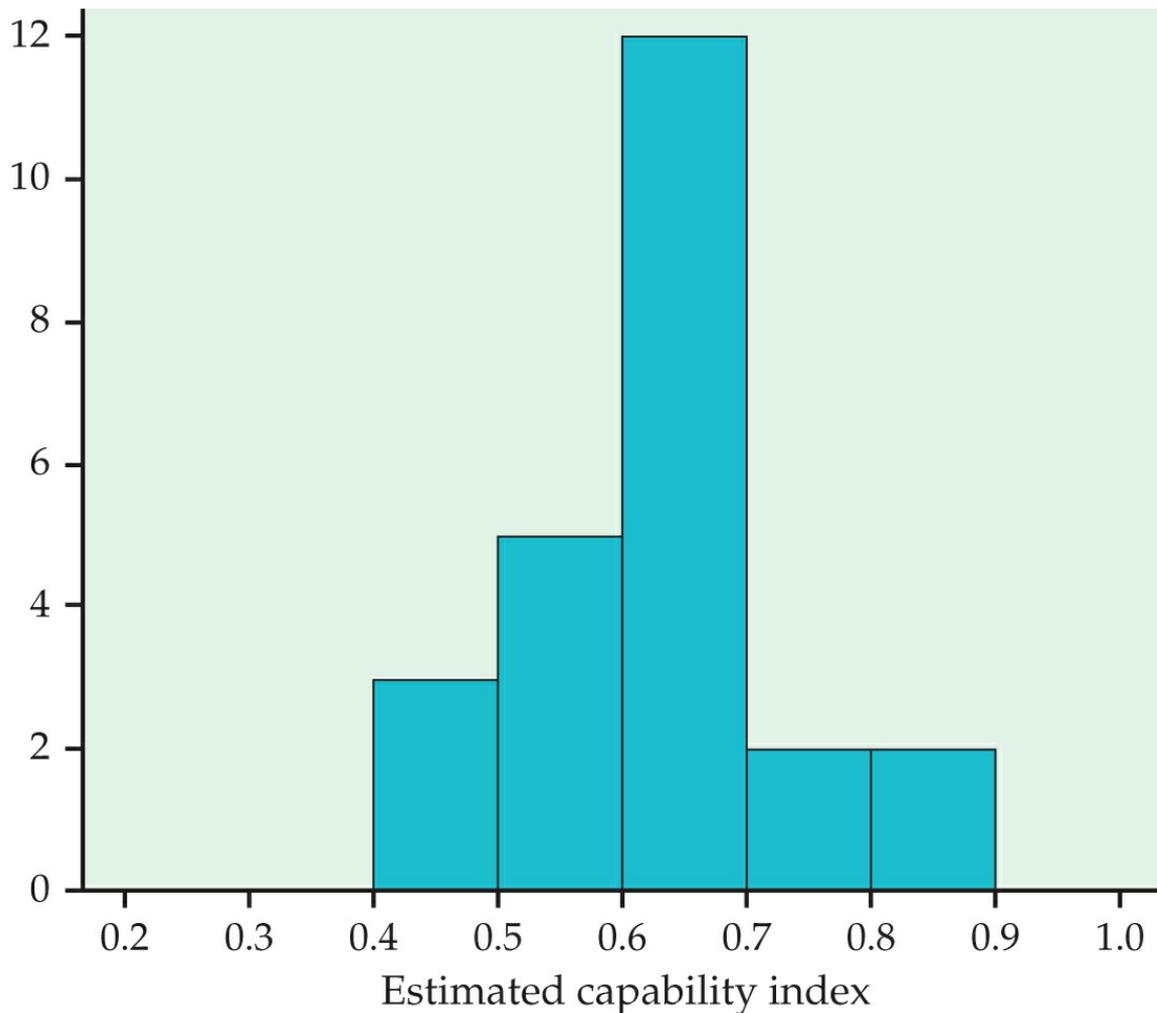
$$C_{pk} = \frac{3500 - 2750}{3(430)} = 0.58$$

Suppose also that the manufacturer measures 4 jackets each four-hour shift and then calculates  $\hat{C}_{pk}$  at the end of 8 shifts. That is,  $\hat{C}_{pk}$  uses measurements from 32 jackets.

Figure 17.20 is a histogram of 24 computer-simulated  $\hat{C}_{pk}$ 's from this setting. They vary from 0.44 to 0.84, almost a two-to-one spread. It is clear that 32 measurements are not enough to reliably estimate  $C_{pk}$ .



*As a very rough rule of thumb, don't trust  $\hat{C}_{pk}$  unless it is based on at least 100 measurements.*



**FIGURE 17.20**

Capability indexes estimated from sample will vary from sample to sample. The histogram shows the variation in  $\hat{C}_{pk}$  in 24 samples, each of size 32, for Example 17.18. The process capability is in fact  $C_{pk} = 0.58$ .

### SECTION 17.3 Summary

**Capability indexes** measure process variability ( $C_p$ ) or process center and variability ( $C_{pk}$ ) against the standard provided by external specifications for the output of the process. Larger values indicate higher capability.

Interpretation of  $C_p$  and  $C_{pk}$  requires that measurements on the process output have a roughly Normal distribution. These indexes are not meaningful unless the process is in control so that its center and variability are stable.

Estimates of  $C_p$  and  $C_{pk}$  can be quite inaccurate when based on small numbers of observations, due to sampling variability. You should mistrust estimates not based on at least 100 measurements.

### SECTION 17.3 Exercises

For Exercises 17.55 and 17.56, see page 17-46.

### 17.57 Capability indexes for the waterproofing process.

Table 17.1 (page 17-10) gives 20 process control samples of the water resistance of a particular outdoor jacket. In Example 17.13, we estimated from these samples that  $\hat{\mu} = \bar{x} = 2750.7$  mm and  $\hat{\sigma} = s = 383.8$  mm.

- (a) The original specifications for water resistance were LSL = 1000 mm and USL = 4000 mm. Estimate  $C_p$  and  $C_{pk}$  for this process.
- (b) A major customer tightened the specifications to LSL = 1500 mm and USL = 3500 mm. Now what are  $C_p$  and  $C_{pk}$ ?

### 17.58 Capability indexes for the waterproofing process, continued.

We could improve the performance of the waterproofing process discussed in the previous exercise by making an adjustment that moves the center of the process to  $\mu = 2500$  mm, the center of the specifications. We should do this even if the original specifications remain in force, because this will require less sealer and therefore cost less. Suppose that we succeed in moving  $\mu$  to 2500 with no change in the process variability  $\sigma$ , estimated by  $s = 383.8$ .

- (a) What are  $C_p$  and  $C_{pk}$  with the original specifications? Compare the values with those from part (a) of the previous exercise.
- (b) What are  $C_p$  and  $C_{pk}$  with the tightened specifications? Again compare with the previous results.

### 17.59 Capability indexes for the meat-packaging process.

Table 17.3 (page 17-19) gives 20 process control samples of the weight of ground beef sections. The lower and upper specifications for the 1-pound sections are 0.96 and 1.10.  MEATWGT

- (a) Using these data, estimate  $C_p$  and  $C_{pk}$  for this process.
- (b) What may be a reason for the specifications being centered at a weight that is slightly greater than the desired 1 pound?

### 17.60 Can we improve the capability of the meat-packaging process?

Refer to Exercise 17.59. The average weight of each section can be increased (or decreased) by increasing (or decreasing) the time between slices of the machine. Based on the results of the previous exercise, would a change in the slicing-time interval improve capability? If so, what value of the average weight should the company seek to attain, and what are  $C_p$  and  $C_{pk}$  with this new process mean?

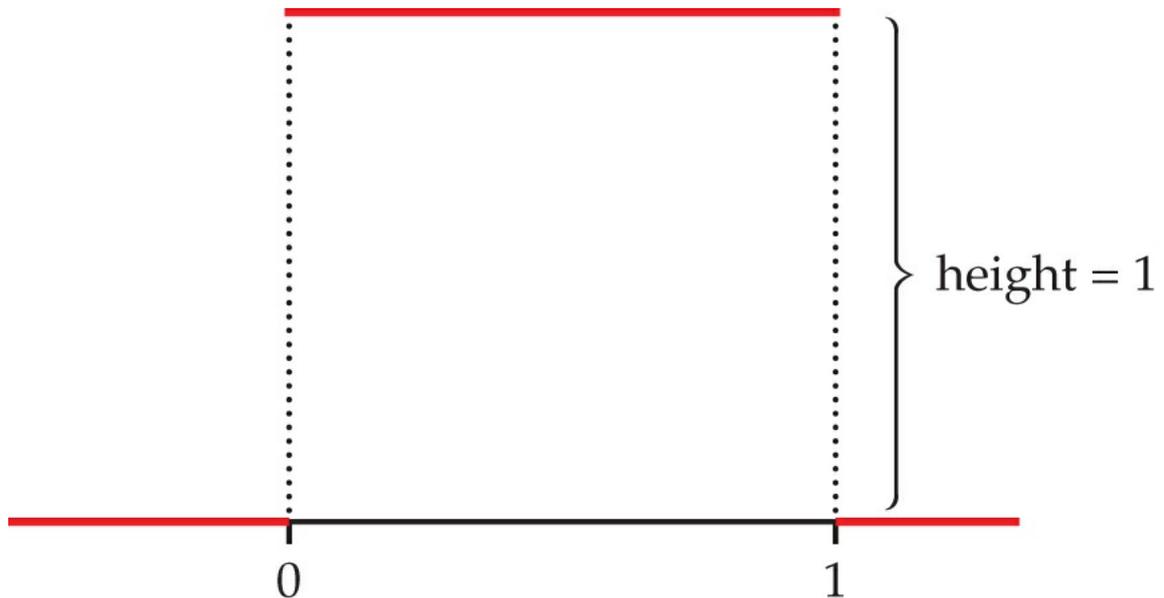
### 17.61 Capability of a characteristic with a uniform distribution.

Suppose that a quality characteristic has the uniform distribution on 0 to 1. Figure 17.21 shows the density curve. You can see that the process mean (the balance point of the density curve) is  $\mu = 1/2$ . The standard deviation turns out to be  $\sigma = 0.289$ . Suppose also that LSL = 1/4 and USL = 3/4.

- (a) Mark LSL and USL on a sketch of the density curve. What is  $C_{pk}$ ? What percent of the output meets the specifications?
- (b) For comparison, consider a process with Normally distributed output having mean  $\mu = 1/2$  and standard deviation  $\sigma = 0.289$ . This process has the same  $C_{pk}$  that you found in part (a). What percent of its output

meets the specifications?

(c) What general fact do your calculations illustrate?



**FIGURE 17.21**

Density curve for the uniform distribution on 0 to 1, for Exercise 17.61.

### 17.62 An alternative estimate for $C_{pk}$ of the waterproofing process.

In Exercise 17.58(b) you found  $C_{pk}$  for specifications  $LSL = 1500$  and  $USL = 3500$  using the standard deviation  $s = 383.8$  for all 80 individual jackets in Table 17.1. Repeat the calculation using the control chart estimate  $\hat{\sigma} = \bar{s}/c_4$ . You should find this  $C_{pk}$  to be slightly larger.

### 17.63 Estimating capability indexes for the distance between holes.

Figure 17.10 (page 17-22) displays a record sheet on which operators have recorded 18 samples of measurements on the distance between two mounting holes on an electrical meter. Sample 5 was out of control on an  $s$  chart. We remove it from the data after the special cause has been fixed. In Exercise 17.47 (page 17-39), you saw that the measurements are reasonably Normal.  **MOUNT**

(a) Based on the remaining 17 samples, estimate the mean and standard deviation of the distance between holes for the population of all meters produced by this process. Make a sketch comparing the Normal distribution with this mean and standard deviation with the specification limits  $54 \pm 10$ .

(b) What are  $C_p$  and  $C_{pk}$  based on the data? How would you characterize the capability of the process? (Mention both center and variability.)

### 17.64 Calculating capability indexes for the DRG 209 hospital losses.

Table 17.9 (page 17-38) gives data on a hospital's losses for 120 DRG 209 patients, collected as 15 monthly samples of 8 patients each. The process has been in control and losses have a roughly Normal distribution. The hospital decides that suitable specification limits for its loss in treating one such patient are  $LSL = \$4500$  and  $USL = \$7500$ .  **DRG**

- (a) Estimate the percent of losses that meet the specifications.
- (b) Estimate  $C_p$ .
- (c) Estimate  $C_{pk}$ .

### 17.65 Assessing the capability of the skateboard bearings process.

Recall the skateboard bearings process described in Exercise 17.50 (page 17-40). The bore diameter has specifications (7.9920, 8.000) mm. The process is monitored by  $\bar{x}$  and  $s$  charts based on samples of 5 consecutive bearings each hour. Control has recently been excellent. The 200 individual measurements from the past week's 40 samples have

$$\bar{x} = 7.996 \text{ mm} \quad s = 0.0023 \text{ mm}$$

A Normal quantile plot shows no important deviations from Normality.

- (a) What percent of bearings will meet specifications if the process remains in its current state?
- (b) Estimate the capability index  $C_{pk}$ .

### 17.66 Will these actions help the capability?

Based on the results of the previous exercise, you conclude that the capability of the bearing-making process is inadequate. Here are some suggestions for improving the capability of this process. Comment on the usefulness of each action suggested.

- (a) Narrowing the control limits so that the process is adjusted more often.
- (b) Additional training of operators to ensure correct operating procedures.
- (c) A capital investment program to install new fabricating machinery.
- (d) An award program for operators who produce the fewest nonconforming bearings.
- (e) Purchasing more uniform (and more expensive) metal stock from which to form the bearings.

### 17.67 $C_{pk}$ and “six-sigma.”

A process with  $C_p \geq 2$  is sometimes said to have “six-sigma quality.” Sketch the specification limits and a Normal distribution of individual measurements for such a process when it is properly centered. Explain from your sketch why this is called six-sigma quality.

### 17.68 More on “six-sigma quality.”

The originators of the “six-sigma quality” idea reasoned as follows. Short-term process variation is described by  $\sigma$ . In the long term, the process mean  $\mu$  will also vary. Studies show that in most manufacturing processes,  $\pm 1.5\sigma$  is adequate to allow for changes in  $\mu$ . The six-sigma standard is intended to allow the mean  $\mu$  to be as much as  $1.5\sigma$  away from the center of the specifications and still meet high standards for percent of output lying outside the specifications.

- (a) Sketch the specification limits and a Normal distribution for process output when  $C_p = 2$  and the mean is  $1.5\sigma$  away from the center of the specifications.

(b) What is  $C_{pk}$  in this case? Is six-sigma quality as strong a requirement as  $C_{pk} \geq 2$ ?

(c) Because most people don't understand standard deviations, six-sigma quality is usually described as guaranteeing a certain level of parts per million of output that fails to meet specifications. Based on your sketch in part (a), what is the probability of an outcome outside the specification limits when the mean is  $1.5\sigma$  away from the center? How many parts per million is this? (You will need software or a calculator for Normal probability calculations, because the value you want is beyond the limits of the standard Normal table.)

Table 17.12 gives the process control samples that lie behind the histogram of call center response times in Figure 17.17(b) on page 17-42. A sample of 6 calls is recorded each shift for quality improvement purposes. The time from the first ring until a representative answers the call is recorded. Table 17.12 gives data for 50 shifts, 300 calls total. Exercises 17.69 to 17.71 make use of this setting.

### 17.69 Choosing the sample.

The 6 calls each shift are chosen at random from all calls received during the shift. Discuss the reasons behind this choice and those behind a choice to time 6 consecutive calls.

### 17.70 Constructing and interpreting the $s$ chart.

Table 17.12 also gives  $\bar{x}$  and  $s$  for each of the 50 samples.

(a) Make an  $s$  chart and check for points out of control.

(b) If the  $s$ -type cause responsible is found and removed, what would be the new control limits for the  $s$  chart? Verify that no points  $s$  are now out of control.

(c) Use the remaining 46 samples to find the center line and control limits for an  $\bar{x}$  chart. Comment on the control (or lack of control) of  $\bar{x}$ . (Because the distribution of response times is strongly skewed,  $\bar{s}$  is large and the control limits for  $\bar{x}$  are wide. Control charts based on Normal distributions often work poorly when measurements are strongly skewed.)

### 17.71 More on interpreting the $s$ chart.

Each of the 4 out-of-control values of  $s$  in part (a) of the previous exercise is explained by a single outlier, a very long response time to one call in the sample. You can see these outliers in Figure 17.17(b). What are the values of these outliers, and what are the  $s$ -values for the 4 samples when the outliers are omitted? (The interpretation of the data is, unfortunately, now clear. Few customers will wait 5 minutes for a call to be answered, as the customer whose call took 333 seconds to answer did. We suspect that other customers hung up before their calls were answered. If so, response time data for the calls that were answered don't adequately picture the quality of service. We should now look at data on calls lost before being answered to see a fuller picture.)

**TABLE 17.12** Fifty Control Chart Samples of Call Center Response Times

Sample	Time (seconds)						Sample mean	Standard deviation
1	59	13	2	24	11	18	21.2	19.93
2	38	12	46	17	77	12	33.7	25.56
3	46	44	4	74	41	22	38.5	23.73
4	25	7	10	46	78	14	30.0	27.46

5	6	9	122	8	16	15	29.3	45.57
6	17	17	9	15	24	70	25.3	22.40
7	9	9	10	32	9	68	22.8	23.93
8	8	10	41	13	17	50	23.2	17.79
9	12	82	97	33	76	56	59.3	32.11
10	42	19	14	21	12	44	25.3	14.08
11	63	5	21	11	47	8	25.8	23.77
12	12	4	111	37	12	24	33.3	39.76
13	43	37	27	65	32	3	34.5	20.32
14	9	26	5	10	30	27	17.8	10.98
15	21	14	19	44	49	10	26.2	16.29
16	24	11	10	22	43	70	30.0	22.93
17	27	10	32	96	11	29	34.2	31.71
18	7	28	22	17	9	24	17.8	8.42
19	15	14	34	5	38	29	22.5	13.03
20	16	65	6	5	58	17	27.8	26.63
21	7	44	14	16	4	46	21.8	18.49
22	32	52	75	11	11	17	33.0	25.88
23	31	8	36	25	14	85	33.2	27.45
24	4	46	23	58	5	54	31.7	24.29
25	28	6	46	4	28	11	20.5	16.34
26	111	6	3	83	27	6	39.3	46.34
27	83	27	2	56	26	21	35.8	28.88
28	276	14	30	8	7	12	57.8	107.20
29	4	29	21	23	4	14	15.8	10.34
30	23	22	19	66	51	60	40.2	21.22
31	14	111	20	7	7	87	41.0	45.82
32	22	11	53	20	14	41	26.8	16.56
33	30	7	10	11	9	9	12.7	8.59
34	101	55	18	20	77	14	47.5	36.16
35	13	11	22	15	2	14	12.8	6.49
36	20	83	25	10	34	23	32.5	25.93
37	21	5	14	22	10	68	23.3	22.82
38	8	70	56	8	26	7	29.2	27.51
39	15	7	9	144	11	109	49.2	60.97
40	20	4	16	20	124	16	33.3	44.80
41	16	47	97	27	61	35	47.2	28.99
42	18	22	244	19	10	6	53.2	93.68
43	43	20	77	22	7	33	33.7	24.49
44	67	20	4	28	5	7	21.8	24.09
45	118	18	1	35	78	35	47.5	43.00
46	71	85	24	333	50	11	95.7	119.53
47	12	11	13	19	16	91	27.0	31.49
48	4	63	14	22	43	25	28.5	21.29
49	18	55	13	11	6	13	19.3	17.90
50	4	3	17	11	6	17	9.7	6.31

## 17.4 Control Charts for Sample Proportions

When you complete this section, you will be able to

- Know when to use a  $p$  chart rather than an  $\bar{x}$  chart.
- Compute the center line and control limits for a  $p$  chart and utilize the chart for process monitoring.

We have considered control charts for just one kind of data: measurements of a quantitative variable in some meaningful scale of units. We describe the distribution of measurements by its center and spread and use  $\bar{x}$  and  $s$  or  $\bar{x}$  and  $R$  charts for process control. There are control charts for other statistics that are appropriate for other kinds of data. The most common of these is the  $p$  chart for use when the data are proportions.

### $p$ CHART

A  $p$  chart is a control chart based on plotting sample proportions  $\hat{p}$  from regular samples from a process against the order in which the samples were taken.

### EXAMPLE

#### 17.19 Examples of the $p$ chart.

Here are two examples of the usefulness of  $p$  charts:

**Manufacturing.** Measure two dimensions of a part and also grade its surface finish by eye. The part conforms if both dimensions lie within their specifications and the finish is judged acceptable. Otherwise, it is nonconforming. Plot the proportion of nonconforming parts in samples of parts from each shift.

**School absenteeism.** An urban school system records the percent of its eighth-grade students who are absent three or more days each month. Because students with high absenteeism in eighth grade often fail to complete high

school, the school system has launched programs to reduce absenteeism. These programs include calls to parents of absent students, public-service messages to change community expectations, and measures to ensure that the schools are safe and attractive. A  $p$  chart will show if the programs are having an effect.



The manufacturing example illustrates an advantage of  $p$  charts: they can combine several specifications in a single chart. Nonetheless,  *$p$  charts have been rendered outdated in many manufacturing applications by improvements in typical levels of quality.* When the proportion of nonconforming parts is very small, even large samples of parts will rarely contain any bad parts. The sample proportions will almost all be 0, so that plotting them is uninformative.

It is better to choose important measured characteristics—voltage at a critical circuit point, for example—and keep  $\bar{x}$  and  $s$  charts. Even if the voltage is satisfactory, quality can be improved by moving it yet closer to the exact voltage specified in the design of the part.

The school absenteeism example is a management application of  $p$  charts. More than 19% of all American eighth-graders miss three or more days of school per month, and this proportion is higher in large cities and for certain ethnic groups.<sup>16</sup> A  $p$  chart will be useful. Proportions of “things going wrong” are often higher in business processes than in manufacturing, so that  $p$  charts are an important tool in business.

## Control limits for $p$ charts

We studied the sampling distribution of a sample proportion  $\hat{p}$  in Chapter 5. The center line and control limits for a  $3\sigma$  control chart follow directly from the facts stated there, in the box on page 330. We ought to call such charts “ $\hat{p}$  charts” because they plot sample proportions. Unfortunately, they have always been called  $p$  charts in quality control circles. We will keep the traditional name but also keep our usual notation:  $p$  is a *process* proportion and  $\hat{p}$  is a *sample* proportion.

### $p$ CHART USING PAST DATA

Take regular samples from a process that has been in control. The samples need not all have the same size. Estimate the process proportion  $p$  of “successes” by

$$\bar{p} = \frac{\text{total number of successes in past samples}}{\text{number of opportunities in these samples}}$$

The center line and control limits for a ***p* chart** for future samples of size  $n$  are

$$UCL = \bar{p} + 3\bar{p}(1-\bar{p})n$$

$$CL = \bar{p}$$

$$LCL = \bar{p} - 3\bar{p}(1-\bar{p})n$$

Common **out-of-control signals** are one sample proportion  $\hat{p}$  outside the control limits or a run of 9 sample proportions on the same side of the center line.

If we have  $k$  past samples of the *same* size  $n$ , then  $\bar{p}$  is just the average of the  $k$  sample proportions. In some settings, you may meet samples of unequal size—differing numbers of students enrolled in a month or differing numbers of parts inspected in a shift. The average  $\bar{p}$  estimates the process proportion  $p$  even when the sample sizes vary. Note that the control limits use the actual size  $n$  of a sample.

## EXAMPLE

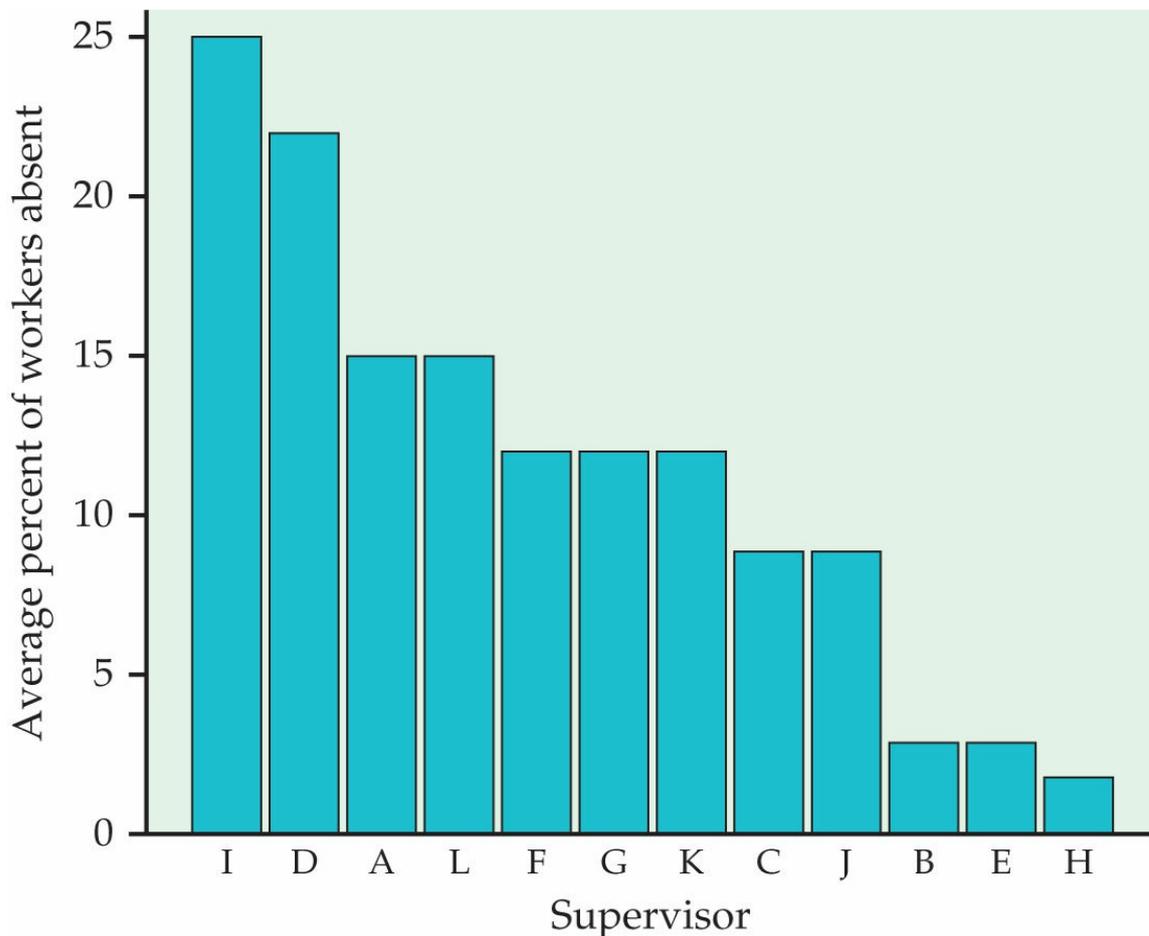
### 17.20 Monitoring employees' absences.

Unscheduled absences by clerical and production workers are an important cost in many companies. Reducing the rate of absenteeism is therefore an important goal for a company's human relations department. A rate of absenteeism above 5% is a serious concern. Many companies set 3% absent as a desirable target. You have been asked to improve absenteeism in a production facility where 12% of the workers are now absent on a typical day.

You first do some background study—in greater depth than this very brief summary. Companies try to avoid hiring workers who are likely to miss work often, such as substance abusers. They may have policies that reward good attendance or penalize frequent absences by individual workers. Changing those policies in this facility will have to wait until the union contract is renegotiated. What might you do with the current workers under current policies?

Studies of absenteeism by clerical and production workers who do repetitive, routine work under close supervision point to unpleasant work environment and harsh or unfair treatment by supervisors as factors that increase absenteeism. It's now up to you to apply this general knowledge to your specific problem.

First, collect data. Daily absenteeism data are already available. You carry out a sample survey that asks workers about their absences and the reasons for them (responses are anonymous, of course). Workers who are more often absent complain about their supervisors and about the lighting at their workstations. Female workers complain that the rest rooms are dirty and unpleasant. You do more data analysis:



**FIGURE 17.22**

Pareto chart of the average absenteeism rate for workers reporting to each of 12 supervisors.

- A Pareto chart of average absenteeism rate for the past month broken down by supervisor (Figure 17.22) shows important differences among supervisors. Only supervisors B, E, and H meet the level of 5% or less absenteeism. Workers supervised by I and D have particularly high rates.
- Another Pareto chart (not shown) by type of workstation shows that a few types of workstation have high absenteeism rates.

Now you take action. You retrain all the supervisors in human relations skills, using B, E, and H as discussion leaders. In addition, a trainer works individually with supervisors I and D. You ask supervisors to talk with any absent worker when he or she returns to work. Working with the engineering department, you study the workstations with high absenteeism rates and make changes such as better lighting.

You refurbish the rest rooms (for both genders even though only women complained) and schedule more frequent cleaning.

## EXAMPLE

### 17.21 Are your actions effective?

You hope to see a reduction in absenteeism. To view progress (or lack of progress), you will keep a  $p$  chart of the proportion of absentees. The plant has 987 production workers. For simplicity, you just record the number who are absent from work each day. Only unscheduled absences count, not planned time off such as vacations. Each day you will plot

$$\hat{p} = \frac{\text{number of workers absent}}{987}$$

You first look back at data for the past three months. There were 64 workdays in these months. The total workdays available for the workers was

$$(64)(987) = 63,168 \text{ person-days}$$

Absences among all workers totaled 7580 person-days. The average daily proportion absent was therefore

$$\begin{aligned} \bar{p} &= \frac{\text{total days absent}}{\text{total days available for work}} \\ &= \frac{7580}{63,168} = 0.120 \end{aligned}$$

The daily rate has been in control at this level.

These past data allow you to set up a  $p$  chart to monitor future proportions absent:

$$\begin{aligned} \text{UCL} &= \bar{p} + 3\bar{p}\sqrt{1-\bar{p}}/\sqrt{n} = 0.120 + 3(0.120)\sqrt{(0.880)/987} \\ &= 0.120 + 0.031 = 0.151 \end{aligned}$$

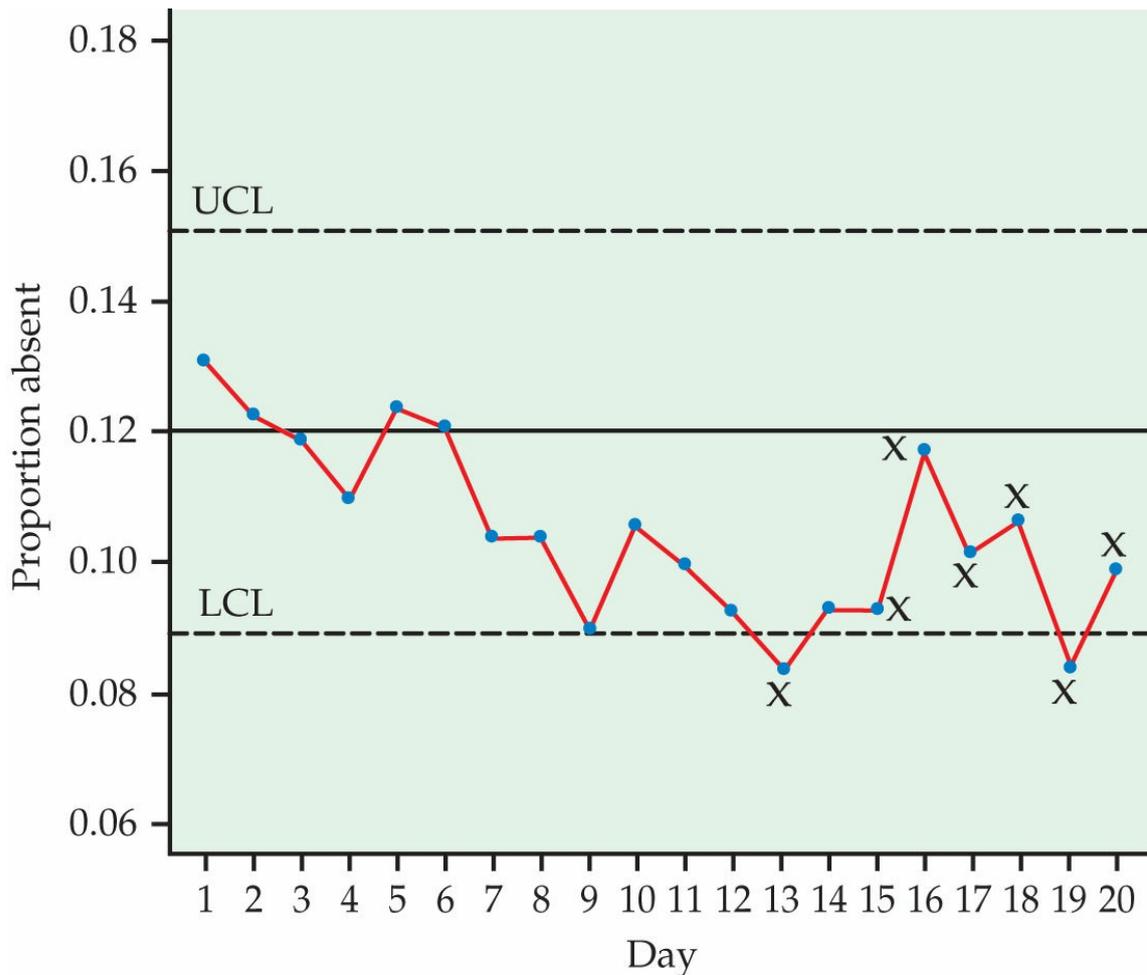
$$\text{CL} = \bar{p} = 0.120$$

$$\begin{aligned} \text{LCL} &= \bar{p} - 3\bar{p}\sqrt{1-\bar{p}}/\sqrt{n} = 0.120 - 3(0.120)\sqrt{(0.880)/987} \\ &= 0.120 - 0.031 = 0.089 \end{aligned}$$

Table 17.13 gives the data for the next four weeks. Figure 17.23 is the  $p$  chart.

**TABLE 17.13****Proportions of Workers Absent During Four Weeks**

<b>Day</b>	<b>M</b>	<b>T</b>	<b>W</b>	<b>Th</b>	<b>F</b>	<b>M</b>	<b>T</b>	<b>W</b>	<b>Th</b>	<b>F</b>
Workers absent	129	121	117	109	122	119	103	103	89	105
Proportion $p^{\wedge}$	0.131	0.123	0.119	0.110	0.124	0.121	0.104	0.104	0.090	0.106
<b>Day</b>	<b>M</b>	<b>T</b>	<b>W</b>	<b>Th</b>	<b>F</b>	<b>M</b>	<b>T</b>	<b>W</b>	<b>Th</b>	<b>F</b>
Workers absent	99	92	83	92	92	115	101	106	83	98
Proportion $p^{\wedge}$	0.100	0.093	0.084	0.093	0.093	0.117	0.102	0.107	0.084	0.099



**FIGURE 17.23**

The  $p$  chart for daily proportion of workers absent over a four-week period, for Example 17.21. The lack of control shows an improvement (decrease) in absenteeism. Update the chart to continue monitoring the process.

Figure 17.23 shows a clear downward trend in the daily proportion of workers who are absent. Days 13 and 19 lie below LCL, and a run of 9 days below the center line is achieved at Day 15 and continues. The points marked “x” are therefore all out of control. It appears that a special cause (the various actions you took) has reduced the absenteeism rate from around 12% to around 10%. The last two weeks’ data suggest that the rate has stabilized at this level. You will update the chart based on the new data. If the rate does not decline further (or even rises again as the effect of your actions wears off), you will consider further changes.

Example 17.21 is a bit oversimplified. The number of workers available did not remain fixed at 987 each day. Hirings, resignations, and planned vacations change the number a bit from day to day. The control limits for a day’s  $\hat{p}$  depend on  $n$ , the number of workers that day. If  $n$  varies, the control limits will move in and out from day to day. Software will do the extra arithmetic needed for a different  $n$  each day, but as long as the count of workers remains close to 987, the greater detail will not change your conclusion.

A single  $p$  chart for all workers is not the only, or even the best, choice in this

setting. Because of the important role of supervisors in absenteeism, it would be wise to also keep separate  $p$  charts for the workers under each supervisor. These charts may show that you must reassign some supervisors.

### SECTION 17.4 Summary

There are control charts for several different types of process measurements. One important type is the  $p$  chart for sample proportions  $\hat{p}$ .

The interpretation of  $p$  charts is very similar to that of  $\bar{x}$  charts. The out-of-control rules used are also the same.

### SECTION 17.4 Exercises

#### 17.72 Constructing a $p$ chart for absenteeism.

After inspecting Figure 17.23, you decide to monitor the next four weeks' absenteeism rates using a center line and control limits calculated from the second two weeks' data recorded in Table 17.13. Find  $\bar{p}$  for these 10 days and give the new values of CL, LCL, and UCL. (Until you have more data, these are trial control limits. As long as you are taking steps to improve absenteeism, you have not reached the process-monitoring stage.)

#### 17.73 Constructing a $p$ chart for unpaid invoices.

The controller's office of a corporation is concerned that invoices that remain unpaid after 30 days are damaging relations with vendors. To assess the magnitude of the problem, a manager searches payment records for invoices that arrived in the past 10 months. The average number of invoices is 2635 per month, with relatively little month-to-month variation. Of all these invoices, 957 remained unpaid after 30 days.

- What is the total number of opportunities for unpaid invoices? What is  $\bar{p}$ ?
- Give the center line and control limits for a  $p$  chart on which to plot the future monthly proportions of unpaid invoices.

#### 17.74 Constructing a $p$ chart for mishandled baggage.

The Department of Transportation reports that 3.09 of every 1000 passengers on domestic flights of the 10 largest U.S. airlines file a report of mishandled baggage.<sup>17</sup> Starting with this information, you plan to sample records for 2500 passengers per day at a large airport to monitor the effects of efforts to reduce mishandled baggage. What are the initial center line and control limits for a chart of the daily proportion of mishandled baggage reports? (You will find that  $LCL < 0$ . Because proportions  $\hat{p}$  are always 0 or positive, take  $LCL = 0$ .)

#### 17.75 Constructing a $p$ chart for damaged eggs.

An egg farm wants to monitor the effects of some new handling procedures on the percent of eggs arriving at the packaging center with cracked or broken shells. In the past, 2.31% of the eggs were damaged. A machine will allow the farm to inspect 500 eggs per hour. What are the initial center line and control limits for a chart of the hourly percent of damaged eggs?

### 17.76 More on constructing a $p$ chart for damaged eggs.

Refer to Exercise 17.75. Suppose that there are two machine operators, each working four-hour shifts. The first operator is very skilled and can inspect 500 eggs per hour. The second operator is less experienced and can inspect only 400 eggs per hour. Construct a  $p$  chart for an eight-hour day showing the appropriate center line and control limits.

### 17.77 Constructing a $p$ chart for missing or deformed rivets.

After completion of an aircraft wing assembly, inspectors count the number of missing or deformed rivets. There are hundreds of rivets in each wing, but the total number varies depending on the aircraft type. Recent data for wings with a total of 38,370 rivets show 194 missing or deformed. The next wing contains 1520 rivets. What are the appropriate center line and control limits for plotting the  $\hat{p}$  from this wing on a  $p$  chart?

### 17.78 Constructing the $p$ chart limits for incorrect or illegible prescriptions.

A regional chain of retail pharmacies finds that about 1% of prescriptions it receives from doctors are incorrect or illegible. The chain puts in place a secure online system that doctors' offices can use to enter prescriptions directly. It hopes that fewer prescriptions entered online will be incorrect or illegible. A  $p$  chart will monitor progress. Use information about past prescriptions to set initial center line and control limits for the proportion of incorrect or illegible prescriptions on a day when the chain fills 90,000 online prescriptions. What are the center line and control limits for a day when only 45,000 online prescriptions are filled?

### 17.79 Calculating the $p$ chart limits for school absenteeism.

Here are data from an urban school district on the number of eighth-grade students with three or more unexcused absences from school during each month of a school year. Because the total number of eighth-graders changes a bit from month to month, these totals are also given for each month.

Month	Sept.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June
Students	911	947	939	942	918	920	931	925	902	883
Absent	291	349	364	335	301	322	344	324	303	344

- Find  $\bar{p}$ . Because the number of students varies from month to month, also find  $\bar{n}$ , the average per month.
- Make a  $p$  chart using control limits based on  $\bar{n}$  students each month. Comment on control.
- The exact control limits are different each month because the number of students  $n$  is different each month. This situation is common in using  $p$  charts. What are the exact limits for October and June, the months with the largest and smallest  $n$ ? Add these limits to your  $p$  chart, using short lines spanning a single month. Do exact limits affect your conclusions?

### 17.80 $p$ chart for a high-quality process.

A manufacturer of consumer electronic equipment makes full use not only of statistical process control but also of automated testing equipment that efficiently tests all completed products. Data from the testing equipment show that finished products have only 2.9 defects per million opportunities.

- What is  $\bar{p}$  for the manufacturing process? If the process turns out 5000 pieces per day, how many

defects do you expect to see per day? In a typical month of 24 working days, how many defects do you expect to see?

(b) What are the center line and control limits for a  $p$  chart for plotting daily defect proportions?

(c) Explain why a  $p$  chart is of no use at such high levels of quality.

### **17.81 More on monitoring a high-quality process.**

Because the manufacturing quality in the previous exercise is so high, the process of writing up orders is the major source of quality problems: the defect rate there is 8000 per million opportunities. The manufacturer processes about 500 orders per month.

(a) What is  $\bar{p}$  for the order-writing process? How many defective orders do you expect to see in a month?

(b) What are the center line and control limits for a  $p$  chart for plotting monthly proportions of defective orders? What is the smallest number of bad orders in a month that will result in a point above the upper control limit?

## CHAPTER 17 Exercises

### 17.82 Describing a process that is in control.

A manager who knows no statistics asks you, “What does it mean to say that a process is in control? Is being in control a guarantee that the quality of the product is good?” Answer these questions in plain language that the manager can understand.

### 17.83 Constructing a Pareto chart.

You manage the customer service operation for a maker of electronic equipment sold to business customers. Traditionally, the most common complaint is that equipment does not operate properly when installed, but attention to manufacturing and installation quality will reduce these complaints. You hire an outside firm to conduct a sample survey of your customers. Here are the percents of customers with each of several kinds of complaints:

Category	Percent
Accuracy of invoices	25
Clarity of operating manual	8
Complete invoice	24
Complete shipment	16
Correct equipment shipped	15
Ease of obtaining invoice adjustments/credits	33
Equipment operates when installed	6
Meeting promised delivery date	11
Sales rep returns calls	4
Technical competence of sales rep	12

- (a) Why do the percents not add to 100%?
- (b) Make a Pareto chart. What area would you choose as a target for improvement?

### 17.84 Choice of control chart.

What type of control chart or charts would you use as part of efforts to assess quality? Explain your choices.

- (a) Time to get security clearance
- (b) Percent of job offers accepted
- (c) Thickness of steel washers
- (d) Number of dropped calls per day

### 17.85 Interpreting signals.

Explain the difference in the interpretation of a point falling beyond the upper control limit of the  $\bar{x}$  chart versus a point falling beyond the upper control limit of an  $s$  chart.

### 17.86 Selecting the appropriate control chart and limits.

At the present time, about 5 out of every 1000 lots of material arriving at a plant site from outside vendors are rejected because they do not meet specifications. The plant receives about 350 lots per week. As part of an effort to reduce errors in the system of placing and filling orders, you will monitor the proportion of rejected lots each week. What type of control chart will you use? What are the initial center line and control limits?

*You have just installed a new system that uses an interferometer to measure the thickness of polystyrene film. To control the thickness, you plan to measure 3 film specimens every 10 minutes and keep  $\bar{x}$  and  $s$  charts. To establish control, you measure 22 samples of 3 films each at 10-minute intervals. Table 17.14 gives  $\bar{x}$  and  $s$  for these samples. The units are millimeters  $\times 10^{-4}$ . Exercises 17.87 to 17.91 are based on this process improvement setting.*

### 17.87 Constructing the $s$ chart.

Calculate control limits for  $s$ , make an  $s$  chart, and comment on control of short-term process variation.  **THICK**

### 17.88 Recalculating the $\bar{x}$ and $s$ charts.

Interviews with the operators reveal that in Samples 1 and 10 mistakes in operating the interferometer resulted in one high-outlier thickness reading that was clearly incorrect. Recalculate  $\bar{x}$  and  $s$  after removing Samples 1 and 10. Recalculate UCL for the  $s$  chart and add the new UCL to your  $s$  chart from the previous exercise. Control for the remaining samples is excellent. Now find the appropriate center line and control limits for an  $\bar{x}$  chart, make the  $\bar{x}$  chart, and comment on control.  **THICK**

### 17.89 Capability of the film thickness process.

The specifications call for film thickness  $830 \pm 25 \text{ mm} \times 10^{-4}$ .

- What is the estimate  $\hat{\sigma}$  of the process standard deviation based on the sample standard deviations (after removing Samples 1 and 10)? Estimate the capability ratio  $C_p$  and comment on what it says about this process.
- Because the process mean can easily be adjusted,  $C_p$  is more informative than  $C_{pk}$ . Explain why this is true.
- The estimate of  $C_p$  from part (a) is probably too optimistic as a description of the film produced. Explain why.

### 17.90 Calculating the percent that meet specifications.

Examination of individual measurements shows that they are close to Normal. If the process mean is set to the target value, about what percent of films will meet the specifications?  **THICK**

### 17.91 More on the film thickness process.

Previously, control of the process was based on categorizing the thickness of each film inspected as satisfactory or not. Steady improvement in process quality has occurred, so that just 15 of the last 5000 films inspected were unsatisfactory.  **THICK**

(a) What type of control chart would be used in this setting, and what would be the control limits for a sample of 100 films?

(b) The chart in part (a) is of little practical value at current quality levels. Explain why.

### 17.92 Probability of an out-of-control signal.

There are other out-of-control rules that are sometimes used with  $\bar{x}$  charts. One is “15 points in a row within the  $1\sigma$  level.” That is, 15 consecutive points fall between  $\mu - \sigma/n$  and  $\mu + \sigma/n$ . This signal suggests either that the value of  $\sigma$  used for the chart is too large or that careless measurement is producing results that are suspiciously close to the target. Find the probability that the next 15 points will give this signal when the process remains in control with the given  $\mu$  and  $\sigma$ .

### 17.93 Probability of another out-of-control signal.

Another out-of-control signal is when four out of five successive points are on the same side of the center line and farther than  $\sigma/n$  from it. Find the probability of this event when the process is in control.

Sample	$\bar{x}$	$s$	Sample	$\bar{x}$	$s$
1	848	20.1	12	823	12.6
2	832	1.1	13	835	4.4
3	826	11.0	14	843	3.6
4	833	7.5	15	841	5.9
5	837	12.5	16	840	3.6
6	834	1.8	17	833	4.9
7	834	1.3	18	840	8.0
8	838	7.4	19	826	6.1
9	835	2.1	20	839	10.2
10	852	18.9	21	836	14.8
11	836	3.8	22	829	6.7

## CHAPTER 17 Notes and Data Sources

1. Texts on quality management give more detail about these and other simple graphical methods for quality problems. The classic reference is Kaoru Ishikawa, *Guide to Quality Control*, Asian Productivity Organization, 1986.
2. The flowchart and a more elaborate version of the cause-and-effect diagram for Example 17.3 were prepared by S. K. Bhat of the General Motors Technical Center as part of a course assignment at Purdue University.
3. Walter Shewhart's classic book, *Economic Control of Quality of Manufactured Product* (Van Nostrand, 1931), organized the application of statistics to improving quality.
4. We have adopted the terms "chart setup" and "process monitoring" from Andrew C. Palm's discussion of William H. Woodall, "Controversies and contradictions in statistical process control," *Journal of Quality Technology*, 32 (2000), pp. 341–350. Palm's discussion appears in the same issue, pp. 356–360. We have combined Palm's stages B ("process improvement") and C ("process monitoring") in writing for beginners because the distinction between them is one of degree.
5. It is common to call these "standards given"  $\bar{x}$  and  $s$  charts. We avoid this term because it easily leads to the common and serious error of confusing control limits (based on the process itself) with standards or specifications imposed from outside.
6. Data provided by Charles Hicks, Purdue University.
7. See, for example, Chapter 3 of Stephen B. Vardeman and J. Marcus Jobe, *Statistical Quality Assurance Methods for Engineers*, Wiley, 1999.
8. The classic discussion of out-of-control signals and the types of special causes that may lie behind special control chart patterns is the *AT&T Statistical Quality Control Handbook*, Western Electric, 1956.
9. The data in Table 17.6 are adapted from data on viscosity of rubber samples appearing in Table P3.3 of Irving W. Burr, *Statistical Quality Control Methods*, Marcel Dekker, 1976.
10. The control limits for the  $s$  chart based on past data are commonly given as  $B_4\bar{s}$  and  $B_3\bar{s}$ . That is,  $B_4 = B_6/c_4$  and  $B_3 = B_5/c_4$ . This is convenient for users, but we choose to minimize the number of control chart constants students must keep straight and to emphasize that process-monitoring and past-data charts are exactly the same except for the source of  $\mu$  and  $\sigma$ .
11. Simulated data based on information appearing in Arvind Salvekar, "Application of six sigma to DRG 209," found at the Smarter Solutions website, [www.smartersolutions.com](http://www.smartersolutions.com).
12. Data provided by Linda McCabe, Purdue University.
13. The first two Deming quotations are from *Public Sector Quality Report*, December 1993, p. 5. They were found online at [deming.eng.clemson.edu/pub/den/files/demqtes.txt](http://deming.eng.clemson.edu/pub/den/files/demqtes.txt). The third quotation is part of the 10th of Deming's "14 points of quality management," from his book *Out of the Crisis*, MIT Press, 1986.
14. Control charts for *individual measurements* cannot use within-sample standard deviations to estimate short-term process variability. The spread between successive observations is the next best thing. Texts such as that cited in Note 7 give the details.
15. The data in Figure 17.17(b) are simulated from a probability model for call pickup times. That pickup times for large financial institutions have median 20 seconds and mean 32 seconds is reported by Jon Anton, "A case study in benchmarking call centers," Purdue University Center for Customer-Driven Quality, no date.
16. These 2011 statistics can be found at [nces.ed.gov/programs/digest/d12/tables/dt12\\_187.asp](http://nces.ed.gov/programs/digest/d12/tables/dt12_187.asp).

17. Data obtained from “Air travel consumer report,” *Office of Aviation Enforcement and Proceedings*, March 2013.

# TABLES

**Table A** Standard Normal Probabilities

**Table B** Random Digits

**Table C** Binomial Probabilities

**Table D**  $t$  Distribution Critical Values

**Table E**  $F$  Critical Values

**Table F**  $\chi^2$  Distribution Critical Values

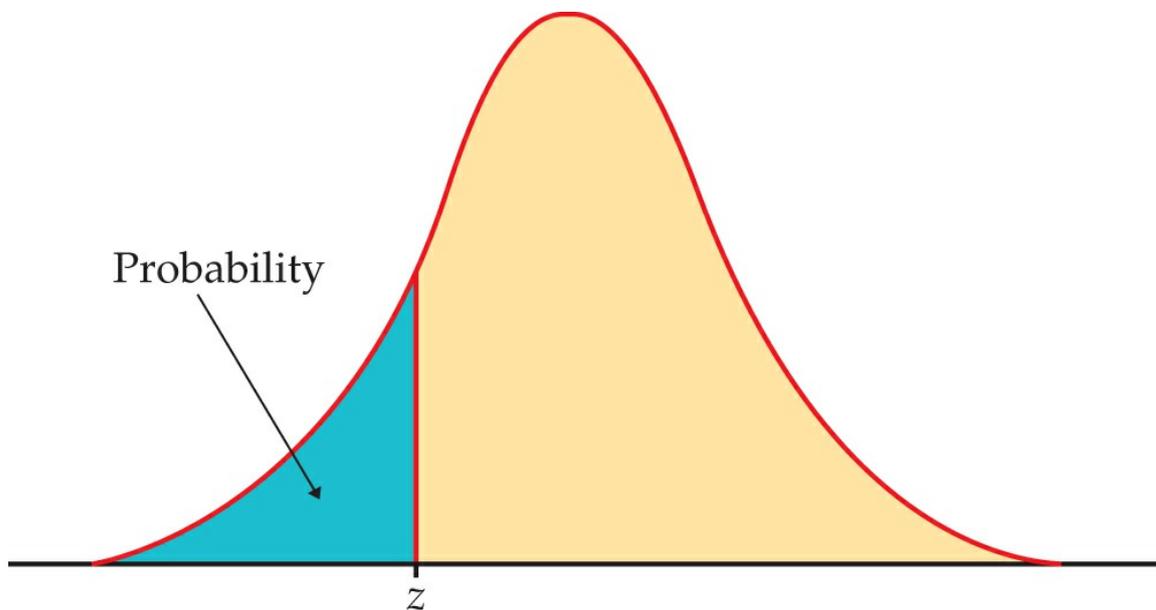


Table entry for  $z$  is the area under the standard Normal curve to the left of  $z$ .

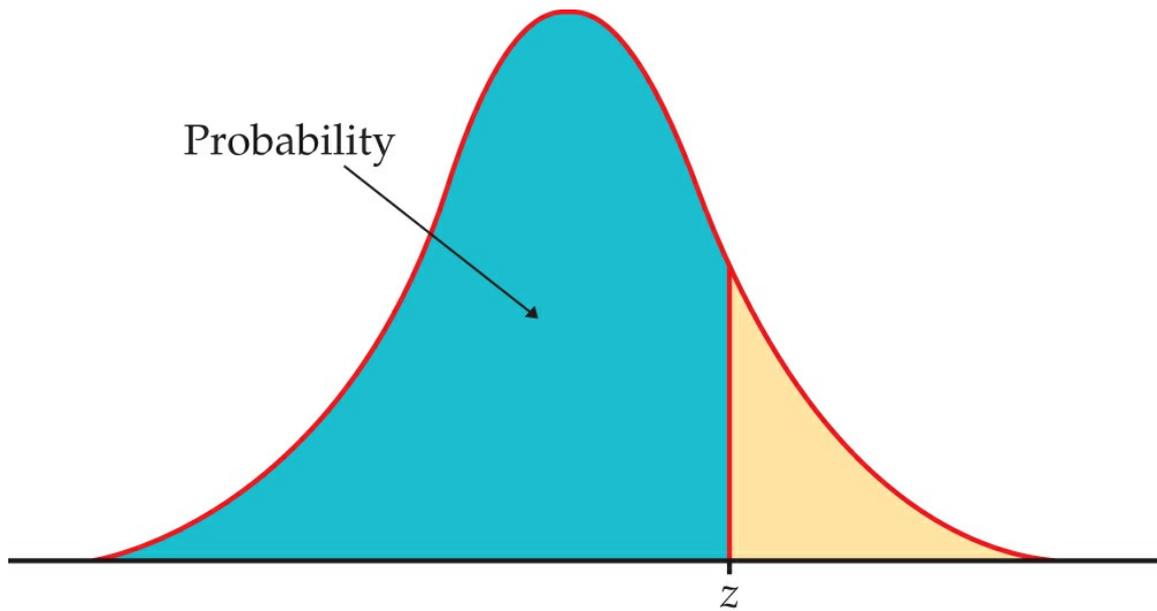


Table entry for  $z$  is the area under the standard Normal curve to the left of  $z$ .

<b>TABLE A</b>		<b>Standard Normal probabilities</b>								
$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379

-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

TABLE B

## Random digits

Line								
101	19223	95034	05756	28713	96409	12531	42544	82853
102	73676	47150	99400	01927	27754	42648	82425	36290
103	45467	71709	77558	00095	32863	29485	82226	90056
104	52711	38889	93074	60227	40011	85848	48767	52573
105	95592	94007	69971	91481	60779	53791	17297	59335
106	68417	35013	15529	72765	85089	57067	50211	47487
107	82739	57890	20807	47511	81676	55300	94383	14893
108	60940	72024	17868	24943	61790	90656	87964	18883
109	36009	19365	15412	39638	85453	46816	83485	41979
110	38448	48789	18338	24697	39364	42006	76688	08708
111	81486	69487	60513	09297	00412	71238	27649	39950
112	59636	88804	04634	71197	19352	73089	84898	45785
113	62568	70206	40325	03699	71080	22553	11486	11776
114	45149	32992	75730	66280	03819	56202	02938	70915
115	61041	77684	94322	24709	73698	14526	31893	32592
116	14459	26056	31424	80371	65103	62253	50490	61181
117	38167	98532	62183	70632	23417	26185	41448	75532
118	73190	32533	04470	29669	84407	90785	65956	86382
119	95857	07118	87664	92099	58806	66979	98624	84826
120	35476	55972	39421	65850	04266	35435	43742	11937
121	71487	09984	29077	14863	61683	47052	62224	51025
122	13873	81598	95052	90908	73592	75186	87136	95761
123	54580	81507	27102	56027	55892	33063	41842	81868
124	71035	09001	43367	49497	72719	96758	27611	91596
125	96746	12149	37823	71868	18442	35119	62103	39244
126	96927	19931	36089	74192	77567	88741	48409	41903
127	43909	99477	25330	64359	40085	16925	85117	36071
128	15689	14227	06565	14374	13352	49367	81982	87209
129	36759	58984	68288	22913	18638	54303	00795	08727
130	69051	64817	87174	09517	84534	06489	87201	97245
131	05007	16632	81194	14873	04197	85576	45195	96565
132	68732	55259	84292	08796	43165	93739	31685	97150
133	45740	41807	65561	33302	07051	93623	18132	09547
134	27816	78416	18329	21337	35213	37741	04312	68508
135	66925	55658	39100	78458	11206	19876	87151	31260
136	08421	44753	77377	28744	75592	08563	79140	92454
137	53645	66812	61421	47836	12609	15373	98481	14592
138	66831	68908	40772	21558	47781	33586	79177	06928
139	55588	99404	70708	41098	43563	56934	48394	51719
140	12975	13258	13048	45144	72321	81940	00360	02428
141	96767	35964	23822	96012	94591	65194	50842	53372
142	72829	50232	97892	63408	77919	44575	24870	04178
143	88565	42628	17797	49376	61762	16953	88604	12724
144	62964	88145	83083	69453	46109	59505	69680	00900

145	19687	12633	57857	95806	09931	02150	43163	58636
146	37609	59057	66967	83401	60705	02384	90597	93600
147	54973	86278	88737	74351	47500	84552	19909	67181
148	00694	05977	19664	65441	20903	62371	22725	53340
149	71546	05233	53946	68743	72460	27601	45403	88692
150	07511	88915	41267	16853	84569	79367	32337	03316
151	03802	29341	29264	80198	12371	13121	54969	43912
152	77320	35030	77519	41109	98296	18984	60869	12349
153	07886	56866	39648	69290	03600	05376	58958	22720
154	87065	74133	21117	70595	22791	67306	28420	52067
155	42090	09628	54035	93879	98441	04606	27381	82637
156	55494	67690	88131	81800	11188	28552	25752	21953
157	16698	30406	96587	65985	07165	50148	16201	86792
158	16297	07626	68683	45335	34377	72941	41764	77038
159	22897	17467	17638	70043	36243	13008	83993	22869
160	98163	45944	34210	64158	76971	27689	82926	75957
161	43400	25831	06283	22138	16043	15706	73345	26238
162	97341	46254	88153	62336	21112	35574	99271	45297
163	64578	67197	28310	90341	37531	63890	52630	76315
164	11022	79124	49525	63078	17229	32165	01343	21394
165	81232	43939	23840	05995	84589	06788	76358	26622
166	36843	84798	51167	44728	20554	55538	27647	32708
167	84329	80081	69516	78934	14293	92478	16479	26974
168	27788	85789	41592	74472	96773	27090	24954	41474
169	99224	00850	43737	75202	44753	63236	14260	73686
170	38075	73239	52555	46342	13365	02182	30443	53229
171	87368	49451	55771	48343	51236	18522	73670	23212
172	40512	00681	44282	47178	08139	78693	34715	75606
173	81636	57578	54286	27216	58758	80358	84115	84568
174	26411	94292	06340	97762	37033	85968	94165	46514
175	80011	09937	57195	33906	94831	10056	42211	65491
176	92813	87503	63494	71379	76550	45984	05481	50830
177	70348	72871	63419	57363	29685	43090	18763	31714
178	24005	52114	26224	39078	80798	15220	43186	00976
179	85063	55810	10470	08029	30025	29734	61181	72090
180	11532	73186	92541	06915	72954	10167	12142	26492
181	59618	03914	05208	84088	20426	39004	84582	87317
182	92965	50837	39921	84661	82514	81899	24565	60874
183	85116	27684	14597	85747	01596	25889	41998	15635
184	15106	10411	90221	49377	44369	28185	80959	76355
185	03638	31589	07871	25792	85823	55400	56026	12193
186	97971	48932	45792	63993	95635	28753	46069	84635
187	49345	18305	76213	82390	77412	97401	50650	71755
188	87370	88099	89695	87633	76987	85503	26257	51736
189	88296	95670	74932	65317	93848	43988	47597	83044
190	79485	92200	99401	54473	34336	82786	05457	60343

191	40830	24979	23333	37619	56227	95941	59494	86539
192	32006	76302	81221	00693	95197	75044	46596	11628
193	37569	85187	44692	50706	53161	69027	88389	60313
194	56680	79003	23361	67094	15019	63261	24543	52884
195	05172	08100	22316	54495	60005	29532	18433	18057
196	74782	27005	03894	98038	20627	40307	47317	92759
197	85288	93264	61409	03404	09649	55937	60843	66167
198	68309	12060	14762	58002	03716	81968	57934	32624
199	26461	88346	52430	60906	74216	96263	69296	90107
200	42672	67680	42376	95023	82744	03971	96560	55148

**TABLE C** Binomial probabilities

		Entry is $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$											
		<i>p</i>											
<i>n</i>	<i>k</i>	.01	.02	.03	.04	.05	.06	.07	.08	.09	.10	.15	.20
2	0	.9801	.9604	.9409	.9216	.9025	.8836	.8649	.8464	.8281	.8100	.7225	.6400
	1	.0198	.0392	.0582	.0768	.0950	.1128	.1302	.1472	.1638	.1800	.2550	.3200
	2	.0001	.0004	.0009	.0016	.0025	.0036	.0049	.0064	.0081	.0100	.0225	.0400
3	0	.9703	.9412	.9127	.8847	.8574	.8306	.8044	.7787	.7536	.7290	.6141	.5120
	1	.0294	.0576	.0847	.1106	.1354	.1590	.1816	.2031	.2236	.2430	.3251	.3840
	2	.0003	.0012	.0026	.0046	.0071	.0102	.0137	.0177	.0221	.0270	.0574	.0960
	3			.0001	.0001	.0002	.0003	.0005	.0007	.0010	.0010	.0034	.0080
4	0	.9606	.9224	.8853	.8493	.8145	.7807	.7481	.7164	.6857	.6561	.5220	.4096
	1	.0388	.0753	.1095	.1416	.1715	.1993	.2252	.2492	.2713	.2916	.3685	.4096
	2	.0006	.0023	.0051	.0088	.0135	.0191	.0254	.0325	.0402	.0486	.0975	.1536
	3			.0001	.0002	.0005	.0008	.0013	.0019	.0027	.0036	.0115	.0256
	4								.0001	.0001	.0005	.0016	
5	0	.9510	.9039	.8587	.8154	.7738	.7339	.6957	.6591	.6240	.5905	.4437	.3277
	1	.0480	.0922	.1328	.1699	.2036	.2342	.2618	.2866	.3086	.3280	.3915	.4096
	2	.0010	.0038	.0082	.0142	.0214	.0299	.0394	.0498	.0610	.0729	.1382	.2048
	3		.0001	.0003	.0006	.0011	.0019	.0030	.0043	.0060	.0081	.0244	.0512
	4					.0001	.0001	.0002	.0003	.0004	.0004	.0022	.0064
	5											.0001	.0003
6	0	.9415	.8858	.8330	.7828	.7351	.6899	.6470	.6064	.5679	.5314	.3771	.2621
	1	.0571	.1085	.1546	.1957	.2321	.2642	.2922	.3164	.3370	.3543	.3993	.3932
	2	.0014	.0055	.0120	.0204	.0305	.0422	.0550	.0688	.0833	.0984	.1762	.2458
	3		.0002	.0005	.0011	.0021	.0036	.0055	.0080	.0110	.0146	.0415	.0819
	4					.0001	.0002	.0003	.0005	.0008	.0012	.0055	.0154
	5										.0001	.0004	.0015
	6												.0001
7	0	.9321	.8681	.8080	.7514	.6983	.6485	.6017	.5578	.5168	.4783	.3206	.2097
	1	.0659	.1240	.1749	.2192	.2573	.2897	.3170	.3396	.3578	.3720	.3960	.3670
	2	.0020	.0076	.0162	.0274	.0406	.0555	.0716	.0886	.1061	.1240	.2097	.2753
	3		.0003	.0008	.0019	.0036	.0059	.0090	.0128	.0175	.0230	.0617	.1147

	4				.0001	.0002	.0004	.0007	.0011	.0017	.0026	.0109	.0287
	5								.0001	.0001	.0002	.0012	.0043
	6											.0001	.0004
	7												
8	0	.9227	.8508	.7837	.7214	.6634	.6096	.5596	.5132	.4703	.4305	.2725	.1678
	1	.0746	.1389	.1939	.2405	.2793	.3113	.3370	.3570	.3721	.3826	.3847	.3355
	2	.0026	.0099	.0210	.0351	.0515	.0695	.0888	.1087	.1288	.1488	.2376	.2936
	3	.0001	.0004	.0013	.0029	.0054	.0089	.0134	.0189	.0255	.0331	.0839	.1468
	4			.0001	.0002	.0004	.0007	.0013	.0021	.0031	.0046	.0185	.0459
	5							.0001	.0001	.0002	.0004	.0026	.0092
	6											.0002	.0011
	7												.0001
	8												
9	0	.9135	.8337	.7602	.6925	.6302	.5730	.5204	.4722	.4279	.3874	.2316	.1342
	1	.0830	.1531	.2116	.2597	.2985	.3292	.3525	.3695	.3809	.3874	.3679	.3020
	2	.0034	.0125	.0262	.0433	.0629	.0840	.1061	.1285	.1507	.1722	.2597	.3020
	3	.0001	.0006	.0019	.0042	.0077	.0125	.0186	.0261	.0348	.0446	.1069	.1762
	4			.0001	.0003	.0006	.0012	.0021	.0034	.0052	.0074	.0283	.0661
	5						.0001	.0002	.0003	.0005	.0008	.0050	.0165
	6										.0001	.0006	.0028
	7												.0003
	8												
	9												
10	0	.9044	.8171	.7374	.6648	.5987	.5386	.4840	.4344	.3894	.3487	.1969	.1074
	1	.0914	.1667	.2281	.2770	.3151	.3438	.3643	.3777	.3851	.3874	.3474	.2684
	2	.0042	.0153	.0317	.0519	.0746	.0988	.1234	.1478	.1714	.1937	.2759	.3020
	3	.0001	.0008	.0026	.0058	.0105	.0168	.0248	.0343	.0452	.0574	.1298	.2013
	4			.0001	.0004	.0010	.0019	.0033	.0052	.0078	.0112	.0401	.0881
	5					.0001	.0001	.0003	.0005	.0009	.0015	.0085	.0264
	6									.0001	.0001	.0012	.0055
	7											.0001	.0008
	8												.0001
	9												
	10												
12	0	.8864	.7847	.6938	.6127	.5404	.4759	.4186	.3677	.3225	.2824	.1422	.0687
	1	.1074	.1922	.2575	.3064	.3413	.3645	.3781	.3837	.3827	.3766	.3012	.2062
	2	.0060	.0216	.0438	.0702	.0988	.1280	.1565	.1835	.2082	.2301	.2924	.2835
	3	.0002	.0015	.0045	.0098	.0173	.0272	.0393	.0532	.0686	.0852	.1720	.2362
	4		.0001	.0003	.0009	.0021	.0039	.0067	.0104	.0153	.0213	.0683	.1329
	5				.0001	.0002	.0004	.0008	.0014	.0024	.0038	.0193	.0532
	6							.0001	.0001	.0003	.0005	.0040	.0155
	7											.0006	.0033
	8											.0001	.0005
	9												.0001
	10												
	11												

	12												
15	0	.8601	.7386	.6333	.5421	.4633	.3953	.3367	.2863	.2430	.2059	.0874	.0352
	1	.1303	.2261	.2938	.3388	.3658	.3785	.3801	.3734	.3605	.3432	.2312	.1319
	2	.0092	.0323	.0636	.0988	.1348	.1691	.2003	.2273	.2496	.2669	.2856	.2309
	3	.0004	.0029	.0085	.0178	.0307	.0468	.0653	.0857	.1070	.1285	.2184	.2501
	4		.0002	.0008	.0022	.0049	.0090	.0148	.0223	.0317	.0428	.1156	.1876
	5			.0001	.0002	.0006	.0013	.0024	.0043	.0069	.0105	.0449	.1032
	6						.0001	.0003	.0006	.0011	.0019	.0132	.0430
	7								.0001	.0001	.0003	.0030	.0138
	8											.0005	.0035
	9											.0001	.0007
	10												.0001
	11												
	12												
	13												
	14												
	15												
20	0	.8179	.6676	.5438	.4420	.3585	.2901	.2342	.1887	.1516	.1216	.0388	.0115
	1	.1652	.2725	.3364	.3683	.3774	.3703	.3526	.3282	.3000	.2702	.1368	.0576
	2	.0159	.0528	.0988	.1458	.1887	.2246	.2521	.2711	.2818	.2852	.2293	.1369
	3	.0010	.0065	.0183	.0364	.0596	.0860	.1139	.1414	.1672	.1901	.2428	.2054
	4		.0006	.0024	.0065	.0133	.0233	.0364	.0523	.0703	.0898	.1821	.2182
	5			.0002	.0009	.0022	.0048	.0088	.0145	.0222	.0319	.1028	.1746
	6				.0001	.0003	.0008	.0017	.0032	.0055	.0089	.0454	.1091
	7						.0001	.0002	.0005	.0011	.0020	.0160	.0545
	8								.0001	.0002	.0004	.0046	.0222
	9										.0001	.0011	.0074
	10											.0002	.0020
	11												.0005
	12												.0001
	13												
	14												
	15												
	16												
	17												
	18												
	19												
	20												

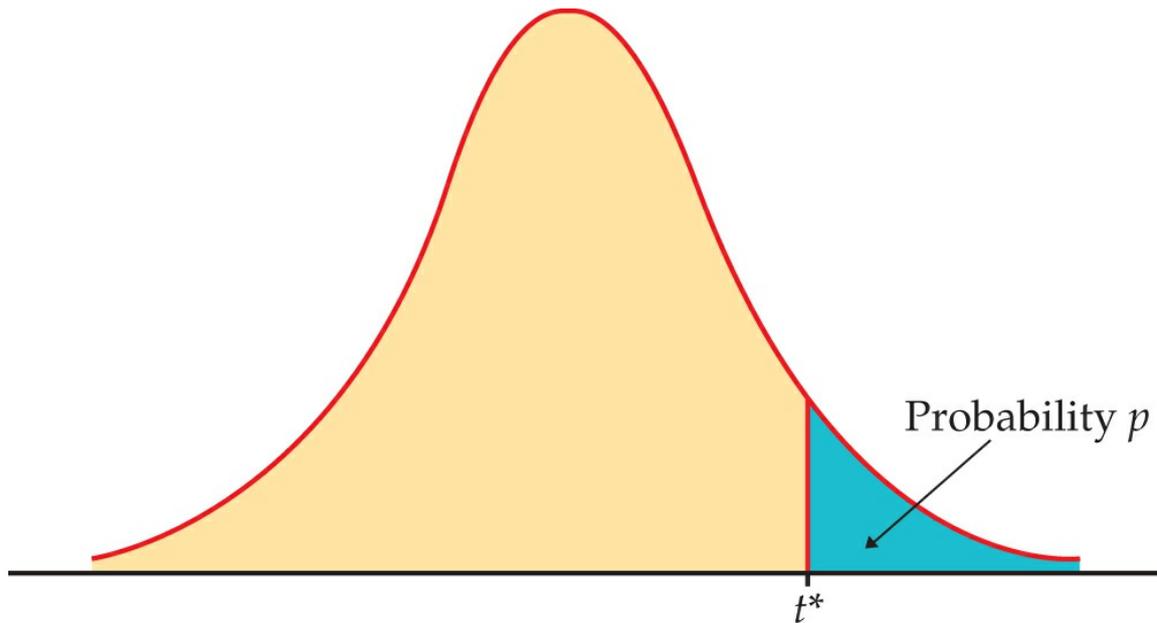


Table entry for  $p$  and  $C$  is the critical value  $t^*$  with probability  $p$  lying to its right and probability  $C$  lying between  $-t^*$  and  $t^*$ .

**TABLE D**

***t* distribution critical values**

df	Upper-tail probability $p$											
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768

24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
$z^*$	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level $C$											

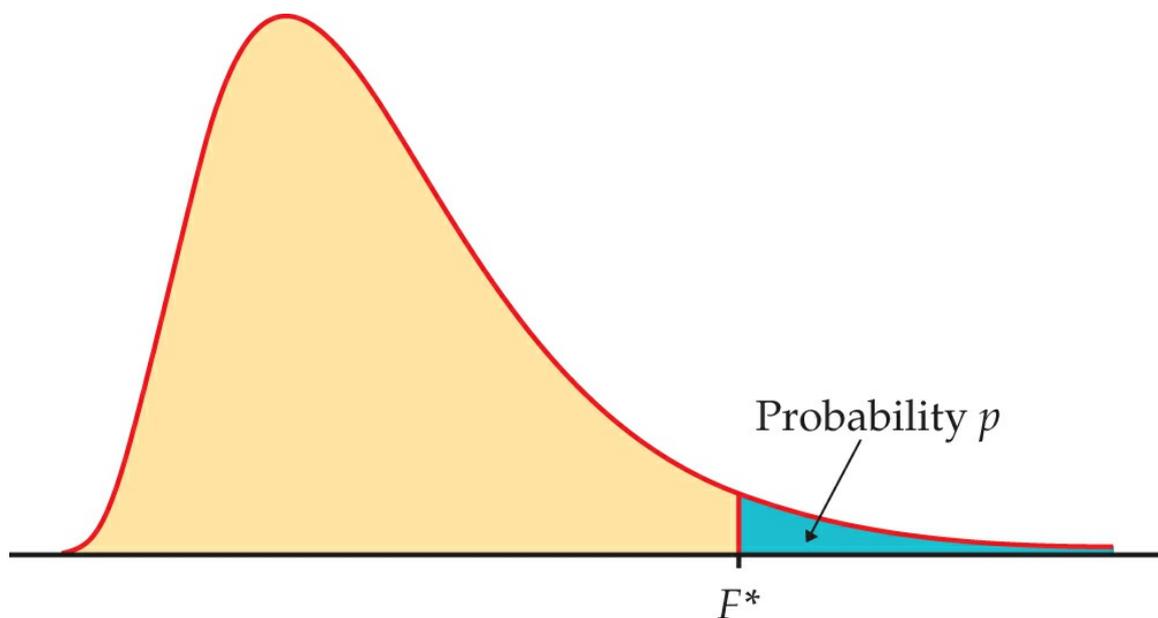


Table entry for  $p$  is the critical value  $F^*$  with probability  $p$  lying to its right.

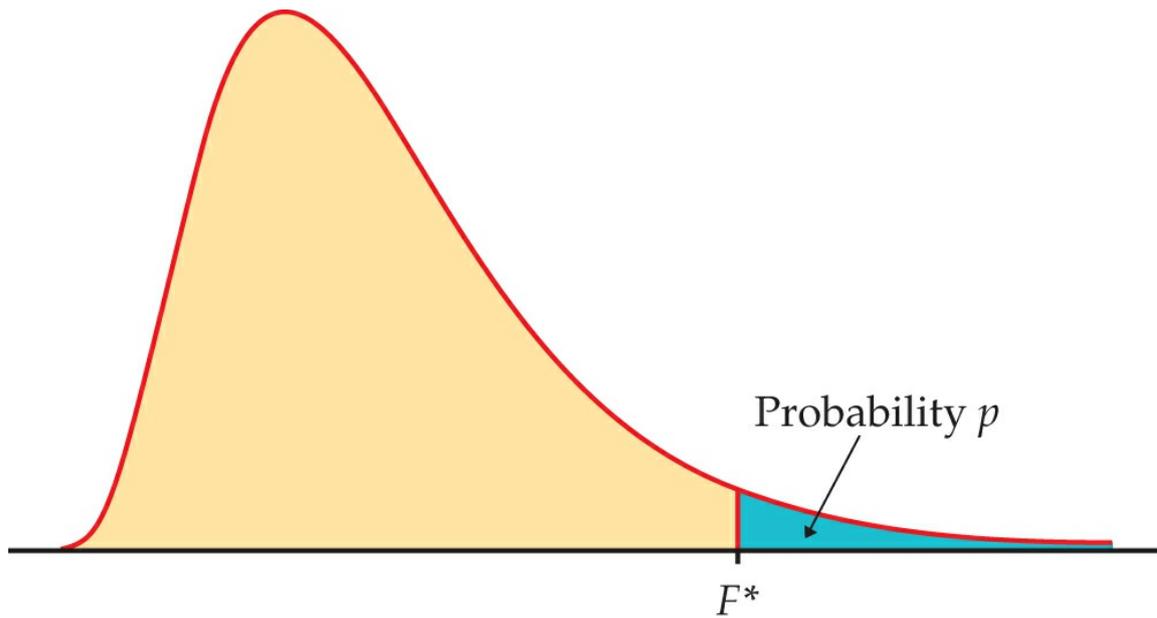


Table entry for  $p$  is the critical value  $F^*$  with probability  $p$  lying to its right.

TABLE E		<i>F</i> critical values									
		Degrees of freedom in the numerator									
	<i>p</i>	1	2	3	4	5	6	7	8	9	10
1	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.14
	.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
	.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28	968.61
	.010	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5	6055.3
	.001	405284	500000	540379	562500	576405	585937	592873	598144	602284	60562
2	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39
	.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40
	.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
	.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39	999.40
3	.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23
	.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.43
	.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.27
	.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86	129.22
4	.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.93
	.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.99
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.86
	.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.57
	.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47	48.02
5	.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.31
	.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.75
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.64
	.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.09

5	.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24	26.90
	.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94
	.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
	.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46
	.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
6	.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69	18.43
	.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70
	.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
	.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76
	.010	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
7	.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33	14.05
	.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54
	.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30
	.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
8	.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77	11.54
	.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42
	.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
	.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96
	.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
9	.001	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11	9.89
	.100	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
	.050	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
	.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72
	.010	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
10	.001	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96	8.75
	.100	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25
	.050	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
	.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53
	.010	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
11	.001	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12	7.92
	.100	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
	.050	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
	.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37
	.010	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
12	.001	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48	7.29
	.100	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14
	.050	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
	.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25
	.010	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
13	.001	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98	6.80
	.100	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10
	.050	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
	.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15
	.010	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
14	.001	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58	6.40

	.100	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
	.050	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
	.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06
	.010	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
15	.001	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08
	.100	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03
	.050	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
	.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99
	.010	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
16	.001	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98	5.81
	.100	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00
	.050	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
	.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92
	.010	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59
17	.001	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75	5.58
	.100	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98
	.050	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
	.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87
	.010	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
18	.001	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56	5.39
	.100	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96
	.050	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
	.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82
	.010	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
19	.001	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39	5.22
	.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
	.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
	.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77
	.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
20	.001	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08
	.100	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92
	.050	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
	.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73
	.010	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
21	.001	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11	4.95
	.100	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90
	.050	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
	.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70
	.010	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
22	.001	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99	4.83
	.100	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89
	.050	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
	.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67
	.010	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
23	.001	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89	4.73
	.100	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88

	.050	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
	.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64
	.010	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
24	.001	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80	4.64
	.100	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
	.050	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
	.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61
	.010	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
25	.001	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71	4.56
	.100	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86
	.050	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
	.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59
	.010	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
26	.001	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64	4.48
	.100	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85
	.050	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
	.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57
	.010	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
27	.001	13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57	4.41
	.100	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84
	.050	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
	.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55
	.010	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
28	.001	13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50	4.35
	.100	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83
	.050	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
	.025	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53
	.010	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
29	.001	13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45	4.29
	.100	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
	.050	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
	.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51
	.010	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
30	.001	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39	4.24
	.100	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76
	.050	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
	.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39
	.010	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
40	.001	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02	3.87
	.100	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73
	.050	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03
	.025	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.33
	.010	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.71
50	.001	12.22	7.96	6.34	5.46	4.90	4.51	4.22	4.00	3.82	3.67
	.100	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
	.050	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99

Degrees of freedom in the denominator	60	.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	
		.010	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.65	
		.001	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69	3.57	
	100	.100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69	1.66	1.63
		.050	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.90
		.025	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.19	2.16
	200	.010	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.45
		.001	11.50	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.44	3.30	3.20
		.100	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66	1.63	1.60
	1000	.050	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.89	1.86
		.025	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18	2.13	2.10
		.010	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.43	2.39
	.001	10.89	6.96	5.46	4.65	4.14	3.78	3.51	3.30	3.13	2.99	2.90	

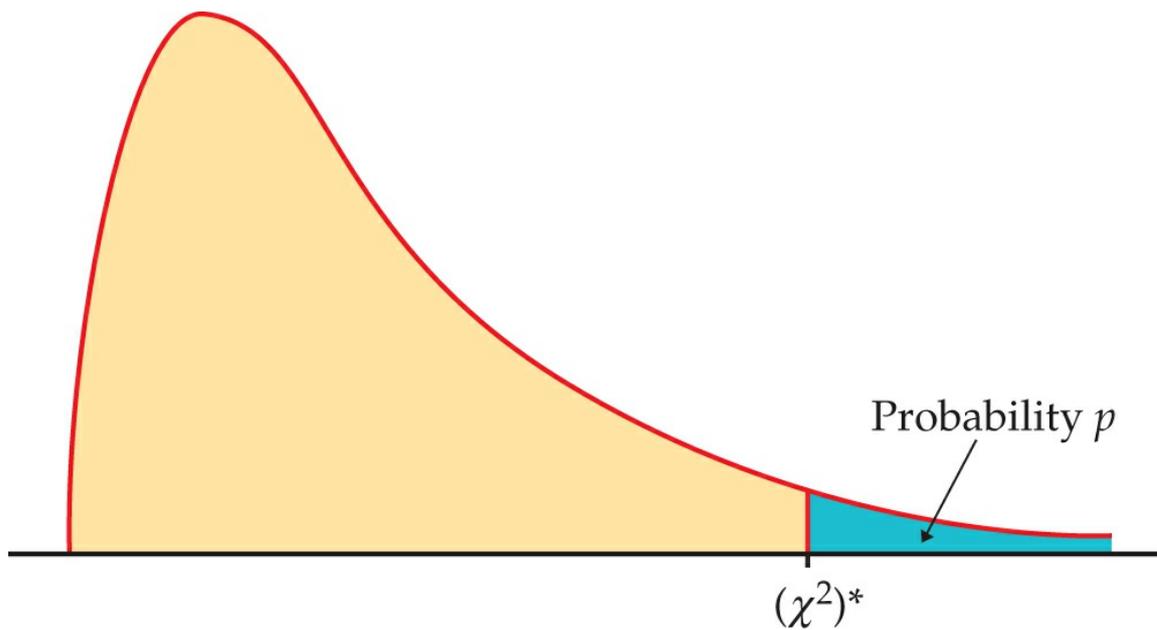


Table entry for  $p$  is the critical value  $(\chi^2)^*$  with probability  $p$  lying to its right.

TABLE F		$\sigma^2$ distribution critical values										
		Tail probability $p$										
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11

6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25	41.31
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79	42.88
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31	47.50
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80	49.01
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27	50.51
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73	52.00
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18	53.48
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62	54.95
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05	56.41
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48	57.86
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89	59.30
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30	60.73
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40	76.09
50	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66	89.56
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61	102.7
80	88.13	90.41	93.11	96.58	101.9	106.6	108.1	112.3	116.3	120.1	124.8	128.3
100	109.1	111.7	114.7	118.5	124.3	129.6	131.1	135.8	140.2	144.3	149.4	153.2

# ANSWERS TO ODD-NUMBERED EXERCISES

## CHAPTER 1

- 1.1** Working in seconds means avoiding decimals and fractions.
- 1.3** Exam1 = 95, Exam2 = 98, Final = 96.
- 1.5** Cases: apartments. Five variables: rent (quantitative), cable (categorical), pets (categorical), bedrooms (quantitative), distance to campus (quantitative).
- 1.7** Answers will vary. **(a)** For example, number of graduates could be used for similar-sized colleges. **(b)** One possibility might be to compare graduation rates between private and public colleges.
- 1.9** **(a)** Individual employees. **(b)** Employee ID number, last name, first name, and middle initial are labels. Department and Education level are categorical variables. Years with the company, Salary, and Age are quantitative.
- 1.11** Age: quantitative, possible values 16 to ? (what would the oldest student's age be?). Sing: categorical, yes/no. Play: categorical, no, a little, pretty well. Food: quantitative, possible values \$0 to ? (what would be the most a person might spend in a week?). Height: quantitative, possible values 2 to 9 feet (check the *Guinness World Records*).
- 1.13** Answers will vary. A few possibilities are graduation rate, student/professor ratio, and job placement rate.
- 1.15** Answers will vary. One possibility is alcohol-impaired fatalities per 100,000 residents. This allows comparing states with different populations; however, states with large seasonal populations (like Florida) might be overstated.
- 1.17** Scores range from 55 to 98. The center is about 80. Very few students scored less than 70.
- 1.19** **(a)** The first line for the 3 (30) stem is now blank. **(b)** Use two stems, even though one is blank. Seeing the gap is useful.
- 1.21** The larger classes hide a lot of detail; there are now only three bars in the histogram.
- 1.23** A stemplot or histogram can be used; the distribution is unimodal and left-skewed, centered near 80, and range from 55 to 98. There are no apparent outliers.
- 1.25** **(b)** Second class had the fewest passengers. Third class had the most; over twice as many as first class.
- (c)** A bar graph of relative frequency would have the same features.
- 1.27** A bar graph would be appropriate because each class is now a “whole” of interest.
- 1.29** **(b)** The overall pattern is unimodal (one major peak). The shape is roughly symmetric with center about 26 and spread from 19 to 33. There appears to be one possible low outlier.
- 1.31** **(a)** 2010 still has the highest usage in December and January. **(b)** The patterns are very similar, but

we don't see the increase between February and March that occurred in 2011; consumption in May was slightly higher in 2010. These differences are most likely due to weather.

**1.33** For example, opinions about least-favorite color are somewhat more varied than for favorite color. Interestingly, purple is liked and disliked by about the same percentage of people.

**1.35 (c)** Preferences will vary, but the ordered bars make it easier to pick out similar categories. The most frequently recycled types (Paper and Trimmings) stand out in both graphs. **(d)** We cannot make a pie chart because each garbage type is a "whole."

**1.37** Mobile browsers are dominated by Safari (on iPads and iPhones). Android has about one-fourth of the market. All others are minor players.

**1.39 (a and b)** Black is clearly more popular in Europe than in North America. The most popular four colors account for at least 70% of cars in both regions. **(c)** One possibility is to cluster the bars for the two regions together by color.

**1.41 (a)**

Region	% FB users	Region	% FB users
Africa	3.9	Middle East	9.4
Asia	5.0	North America	49.9
Caribbean	15.4	Oceania/Australia	38.9
Central America	26.5	South America	28.1
Europe	28.5		

**(b)** For example, when looking only at the absolute number of Facebook users, Europe is the leading region; however, when expressed as a percent of the population, North America has the most Facebook users. **(d)** The shape of the distribution might be right-skewed (there are numerical gaps between 28 and 38 and between 38 and 49). The center of the distribution is about 26% (Central America). This stemplot does not really indicate any major outliers. **(e)** Answers will vary, but one possibility is that the scaling in the stemplot actually hides the gaps in the distribution. **(f)** One possibility is that both the population and number of Facebook users are rounded.

**1.43 (a)** Four variables: GPA, IQ, and self-concept are quantitative; gender is categorical. **(c)** Unimodal and skewed to the left, centered near 7.8, spread from 0.5 to 10.8. **(d)** There is more variability among the boys; in fact, there seem to be two groups of boys: those with GPAs below 5 and those with GPAs above 5.

**1.45** Unimodal and skewed to the left, centered near 59.5; most scores are between 35 and 73, with a few below that and one high score of 80 (probably not quite an outlier).

**1.47** The new mean is 50.44 days.

**1.49** The sorted data are

5 5 5 5 6 7  
 7 7 8 12 12 13  
 13 15 18 18 27 28  
 36 48 52 60 66 94 694

Adding the outlier adds another observation but does not change the median at all.

**1.51**  $M = 83$ .

**1.53**  $\bar{x} = 196.575$  minutes (the value 197 in the text was rounded). The quartiles and median are in positions 20.5, 40.5, and 60.5.  $Q_1 = 54.5$ ,  $M = 103.5$ ,  $Q_3 = 200$ .

**1.55** Use the five-number summary from Exercise 1.54 (55, 75, 83, 93, 98). Be sure to give the plot a consistent, number-line axis.

**1.57**  $s^2 = 159.34$  and  $s = 12.62$ .

**1.59** Without Suriname,  $IQR = 25$ ; with Suriname,  $IQR = 35$ . The  $IQR$  increases because there is one additional large observation, but it does not increase as much as the sample mean does.

**1.61 (a)**  $\bar{x} = 122.92$ . **(b)**  $M = 102.5$ . **(c)** The data set is right-skewed with an outlier (London), so the median is a better measure of center.

**1.63 (a)**  $IQR = 62$ . **(b)** Outliers are below  $-26$  or above  $222$ . London is an outlier. **(c)** The first three quarters are about equal in length; the last (upper quarter) is extremely long. **(d)** The main part of the distribution is relatively symmetric; there is one extreme high outlier. **(f)** For example, the stemplot and the boxplot both indicate the same shape: relatively symmetric with an extremely high outlier.

**1.65 (a)**  $s = 8.80$ . **(b)**  $Q_1 = 43.8$  and  $Q_3 = 57.0$ . **(c)** For example, if you think that the median is the better center in Exercise 1.64, that statistic should be paired with the quartiles and not with the standard deviation.

**1.67 (a)** A histogram of the data shows a strong right-skew. Half the companies have values less than \$7.5 million. **(b)** Using software, we find the numerical summaries shown below.

Mean	StDev	Min	Q1	Med	Q3	Max
13,830	16,050	3731	4775	7516	15,537	77,839

**(c)** Answers will vary, but due to the severe right-skew, this distribution is best described by the five-number summary.

**1.69 (a)** With all the data,  $\bar{x} = 5.23$  and  $M = 4.9$ . Removing the outliers, we have  $\bar{x} = 4.93$  and  $M = 4.8$ . **(b)** With all the data,  $s = 1.429$ ;  $Q_1 = 4.4$ ,  $Q_3 = 5.6$ . Removing the outliers, we have  $s = 0.818$ ,  $Q_1 = 4.4$ , and  $Q_3 = 5.5$ .

**1.71 (a)** With a small data set, a stemplot is reasonable. There are clearly two clumps of data. Summary statistics are shown below.

Mean	StDev	Min	Q1	Med	Q3	Max
6.424	1.400	3.7	4.95	6.7	7.85	8

**(b)** Because of the clusters of data, one set of numerical summaries will not be adequate. **(c)** After separating the data, we have for the smaller weights:

Mean	StDev	Min	Q1	Med	Q3	Max
4.662	0.501	3.7	4.4	4.7	5.075	5.3

And for the larger weights:

Mean	StDev	Min	Q1	Med	Q3	Max
7.253	0.740	6	6.5	7.6	7.9	8

**1.73 (a)** 0, 0, 5.09, 9.47, 73.2. **(d)** Answers will vary. The distribution is unimodal and strongly skewed to the right with five high outliers.

**1.75** This distribution is unimodal and right-skewed and has no outliers. The five-number summary is 0.24, 0.355, 0.75, 1.03, 1.9.

**1.77** Some people, such as celebrities and business executives, make a very large amount of money and have very large assets (think Bill Gates of Microsoft, Warren Buffett, Oprah, etc.).

- 1.79** The mean is \$92,222.22. Eight of the employees make less than this.  $M = \$45,000$ .
- 1.81** The median doesn't change, but the mean increases to \$101, 111.11.
- 1.83** The average would be 2.5 or less (an earthquake that usually isn't felt). These do little or no damage.
- 1.85** For  $n = 2$  the median is also the average of the two values.
- 1.87 (a)** Place the new point at the current median.
- 1.89 (a)** *Bihai*:  $\bar{x} = 47.5975$ ,  $s = 1.2129$ . Red:  $\bar{x} = 39.7113$ ,  $s = 1.7988$ . Yellow:  $\bar{x} = 36.1800$ ,  $s = 0.9753$  (all in mm). **(b)** *Bihai* and red appear to be right-skewed (although it is difficult to tell with such small samples). Skewness would make these distributions unsuitable for  $\bar{x}$  and  $s$ .
- 1.91** Take six or more numbers, with the smallest number much smaller than  $Q_1$ .
- 1.93 (a)** Any set of four identical numbers works. **(b)** 0, 0, 20, 20 is the only possible answer.
- 1.95** Answers will vary with the technology. With newer technology, it is very hard to make this fail, until you reach the limits of the length of the number of digits allowed.
- 1.97**  $\bar{x} = 5.104$  pounds and  $s = 2.662$  pounds.
- 1.99** Full data set:  $\bar{x} = 196.575$  and  $M = 103.5$  minutes. The 10% and 20% trimmed means are  $\bar{x} = 127.734$  and  $\bar{x} = 111.917$  minutes, respectively.
- 1.101** 212 to 364.
- 1.103**  $z = 2.03$ .
- 1.105**  $z = 1.37$ . Using Table A, the proportion below 340 is 0.9147, and the proportion at or above is 0.0853. Using technology, the proportion below 340 is 0.9144.
- 1.107**  $x = \mu + z\sigma$  From Table A, we find the area to the left of  $z = 0.67$  is 0.7486 and the area to the left of  $z = 0.68$  is 0.7517. (Technology gives  $z = 0.6745$ .) If we approximate as  $z = 0.675$ , we have  $x = 313.65$ , or about 314.
- 1.109 (a)** In symmetric distributions, the mean and median are equal to each other. Examples are an equilateral triangle and a rectangle. **(b)** In left-skewed distributions, the mean is less than the median.
- 1.111 (c)** The distributions look the same, only shifted.
- 1.113 (c)** The table below indicates the desired ranges.

	Low	High
68%	256	320
95%	224	352
99.7%	192	384

**1.115**

Value	Percentile (Table A)	Percentile (Software)
150	50	50
140	38.6	38.8
100	7.6	7.7
180	80.5	80.4
230	98.9	98.9

**1.117** Using the  $N(153, 34)$  distribution, we find the values corresponding to the given percentiles as given below (using Table A). The actual scores are very close to the percentiles of the Normal distribution.

Percentile	Score	Score with $N(153, 34)$
10%	110	109
25%	130	130
50%	154	153
75%	177	176
90%	197	197

**1.119 (a)** Ranges are given in the table.

	Women	Men
68%	8489 to 20, 919	7158 to 22, 886
95%	2274 to 27, 134	-706 to 30, 750
99.7%	-3941 to 33, 349	-8570 to 38, 614

In both cases, some of the lower limits are negative, which does not make sense; this happens because the women's distribution is skewed, and the men's distribution has an outlier. Contrary to the conventional wisdom, the men's mean is slightly higher, although the outlier is at least partly responsible for that. **(b)** The means suggest that Mexican men and women tend to speak more than people of the same gender from the United States.

**1.121 (a)** F: -1.645. D: -1.04. C: 0.13. B: 1.04. **(b)** F: below 55.55. D: between 55.55 and 61.6. C: between 61.6 and 73.3. B: between 73.3 and 82.4. A: above 82.4. **(c)** Opinions will vary.

**1.123 (a)**  $1/5 = 0.2$ . **(b)**  $1/5 = 0.2$ . **(c)**  $2/5 = 0.4$ .

**1.125 (a)** Mean is C, median is B (the right-skew pulls the mean to the right). **(b)** Mean A, median A. **(c)** Mean A, median B (the left-skew pulls the mean to the left).

**1.127 (a)** The applet shows an area of 0.6826 between -1.000 and 1.000, while the 68-95-99.7 rule rounds this to 0.68. **(b)** Between -2.000 and 2.000, the applet reports 0.9544 (rather than the rounded 0.95 from the 68-95-99.7 rule). Between -3.000 and 3.000, the applet reports 0.9974 (rather than the rounded 0.997).

**1.129 (a)** 0.0446. **(b)** 0.9554. **(c)** 0.0287. **(d)** 0.9267.

**1.131 (a)** 0.77. **(b)** 0.77.

**1.133** 2.28%, or 0.0228.

**1.135** Anthony has a  $z$ -score of -1.48. Joshua's  $z$ -score is -0.83. Joshua's score is higher.

**1.137** About 2111.

**1.139** 20th percentile.

**1.141** About 1094 and lower.

**1.143** 1285, 1498, and 1711 (rounded to the nearest integer).

**1.145 (a)** From Table A, 33% of men have low values of HDL. (Software gives 32.95%.) **(b)** From Table A, 15.15% of men have protective levels of HDL. (Software gives 15.16%.) **(c)** 51.85% of men are in the intermediate range for HDL. (Software gives 51.88%.)

**1.147 (a)**  $\pm 1.2816$ . **(b)** 8.93 and 9.31 ounces.

1.149 (a) 1.3490. (b)  $c = 1.3490$ .

1.151

Percentile	10%	20%	30%	40%	50%
HDL level	35.2	42.0	46.9	51.1	55
Percentile	60%	70%	80%	90%	
HDL level	58.9	63.1	68.0	74.8	

1.153 (a) The yellow variety is the nearest to a straight line. (b) The other two distributions are both slightly right-skewed, and the *bihai* variety appears to have a couple of high outliers. (c) The deviations do not appear to be Normal. They seem to be right-skewed.

1.155 Histograms will suggest (but not exactly match) Figure 1.32. The uniform distribution does not extend as low or as high as a Normal distribution.

1.157 (a) The distribution appears to be roughly Normal, apart from two possible low and two possible high outliers. (b) The outliers on either end would inflate the standard deviation. The five-number summary is 8.5, 13.15, 15.4, 17.8, 23.8. (c) For example, smoking rates are typically 12% to 20%. Which states are high, and which are low?

1.159 For example, white is least popular in China, and silver is less common in Europe. Silver, white, gray, and black dominate the market worldwide.

1.163 (a) The distribution of 2010 Internet users is right-skewed. The five-number summary is 0.21, 10.31, 31.40, 55.65, 95.63. (b) The distribution of the change in users is right-skewed. The five-number summary is  $-1.285$ , 0.996, 2.570, 4.811, 22.000. (c) The percent change is also right-skewed. Two countries effectively tripled their Internet penetration (but it's still minuscule). The five-number summary is  $-12.50$ , 5.58, 10.75, 20.70, 327.32.

1.165 A bar graph is appropriate (there are other providers besides the 10 largest; we don't know who they are). There are two major providers and several smaller ones.

1.167 (a) For car makes (a categorical variable), use either a bar graph or a pie chart. For car age (a quantitative variable), use a histogram, stemplot, or boxplot. (b) Study time is quantitative, so use a histogram, stemplot, or boxplot. To show change over time, use a time plot (average hours studied against time). (c) Use a bar graph or pie chart to show radio station preferences. (d) Use a Normal quantile plot to see whether the measurements follow a Normal distribution.

1.169  $\sigma = 7.50$ .

1.171 (a) One option is to say  $\mu = 81.55$  (the average of the given 50th percentile and mean). The 5th percentile is  $42 = 81.55 - 1.645\sigma$ . The 95th percentile is  $142 = 81.55 + 1.645\sigma$ . If we average the two estimates, we would have  $\sigma = 30.4$ . (c) From the two distributions, over half of women consume more vitamin C than they need, but some consume far less.

1.173 (a) Not only are most responses multiples of 10; many are multiples of 30 and 60. The students who claimed 360 minutes (6 hours) and 300 minutes (5 hours) may have been exaggerating. (b) Women seem to generally study more (or claim to), as there are none that claim less than 60 minutes per night. The center (median) for women is 170; for men the median is 120 minutes. (c) Opinions will vary.

1.175 No to both questions; no summary can exactly describe a distribution that can include any number of values.

1.177 Simulation results will vary.

## CHAPTER 2

**2.1 (a)** The 30 students. **(b)** Attendance and score on the final exam. **(c)** Score on the final is quantitative. Attendance is most likely quantitative: number of classes attended (or missed).

**2.3** Cases: cups of Mocha Frappuccino. Variables: size and price (both quantitative).

**2.5 (a)** Tweets. **(b)** Click count and length of tweet are quantitative. Day of week and gender are categorical. Time of day could be quantitative (as hr:min) or categorical (if morning, afternoon, etc.). **(c)** Click count is the response. The others could all be potentially explanatory.

**2.7** Answers will vary. Some possible variables are condition, number of pages, and binding type (hardback or paperback), in addition to purchase price and buyback price. Cases are the individual textbooks; one might be interested in predicting buyback price based on other variables.

**2.9 (a)** Temperatures are usually similar from one day to the next (recording temperatures at noon each day, for example). One variable that would help is whether a front (cold or warm) came through. **(b)** No relationship. These are different individuals. **(c)** Answers will vary. It's possible that quality and price are related but not certain.

**2.11** Price per load looks right-skewed. Quality rating has two different clusters of values.

Variable	Mean	StDev	Min	Q1	Med	Q3	Max
Rating	43.88	10.77	26	33.5	47	51.5	61
PricePerLoad	14.21	5.99	5	10	13.5	17	30

**2.13 (a)** Divide each price by 100 to convert to dollars. **(c)** The only difference is the scaling of the  $x$  axis.

**2.15** For example, a new variable might be the ratio of the 2010 and 2009 debts.

**2.17 (a)** All the liquid detergents are at the upper right, and the powder detergents are at the lower left. **(b)** Answers will vary.

**2.19 (b)** The overall pattern is linear and increasing. There is one possible outlier at the upper right, far from the other points. **(c)** The relationship is roughly linear, increasing, and moderately strong. **(d)** The baseball player represented by the point at the far right is not as strong in his dominant arm as other players. **(e)** Other than the one outlier, the relationship is approximately linear.

**2.21 (a)** Population should be the explanatory variable and college students the response. **(b)** The graph shows a strong, linear, increasing relationship with one high outlier in both undergraduates and population (California).

**2.23 (b and c)** The relationship is very strong, linear, and decreasing. **(d)** There do not appear to be any outliers. **(e)** The relationship is linear.

**2.25 (a)** The description is for variables that are positively related. **(b)** The response variable is plotted on the  $y$  axis, and the explanatory on the  $x$  axis. **(c)** A histogram shows the distribution of a single variable, not the relationship between two variables.

**2.27 (b)** The relationship is linear, increasing, and much stronger than the relationship between carbohydrates and percent alcohol.

**2.29 (b)** The plot is much more linear.

**2.31 (a)** Examine for a relationship. **(b)** Use high school GPA as explanatory and college GPA as response. **(c)** Use square feet as explanatory and rental price as response. **(d)** Use amount of sugar as explanatory and sweetness as response. **(e)** Use temperature yesterday at noon as explanatory and temperature today at noon as response.

**2.33 (a)** In general, we expect more intelligent children to be better readers, and less intelligent children to be weaker readers. The plot does show this positive association. **(b)** These four have moderate IQs but poor reading scores. **(c)** Roughly linear but weak (much scatter).

**2.35 (b)** The association is positive and linear. Overall, the relationship is strong, but it is stronger for women than for men. Male subjects generally have both greater lean body mass and higher metabolic rates than women.

**2.37 (a)** Both show fairly steady improvement. Women have made more rapid progress but have not improved since 1993, while men's records may be dropping more rapidly in recent years. **(b)** The data support the first claim but do not seem to support the second.

**2.39 (a)** This is a linear transformation. Dollars =  $0 + 0.01 \times$  cents. **(b)**  $r = 0.671$ . **(c)** They are the same. **(d)** Changing units does not change  $r$ .

**2.41 (a)** No *linear* relationship. (There could be a nonlinear relationship, though.) **(b)** Strongly linear and negative. **(c)** Weakly linear and positive. **(d)** Strongly linear and positive.

**2.43 (a)**  $r = 0.905$ . **(b)** Correlation is a good summary for these data. The pattern is linear and appears to be strong. There is, however, one outlier at the upper right.

**2.45 (a)**  $r = 0.984$ . **(b)** The correlation may be a good summary for these data because the scatterplot is strongly linear. California, however, is an outlier that strengthens the relationship (makes  $r$  closer to 1). **(c)** Eliminate California, Texas, Florida, and New York.  $r = 0.971$ . Expanding the range of values can strengthen a relationship (if the new points follow the rest of the data).

**2.47 (a)**  $r = -0.999$ . **(b)** Correlation is a good numerical summary here because the scatterplot is very strongly linear. **(c)** You must be careful; there can be a strong correlation between two variables even when the relationship is curved.

**2.49** The correlation would be 1 in both cases. These are purely linear relationships.

**2.51**  $r = 0.521$ .

**2.53 (a)**  $r = -0.730$ . **(b)** The relationship is curved; birthrate declines with increasing Internet use until about 40 Internet users per 100 people. After that, there is a steady overall birthrate. Correlation is not a good numerical summary for this relationship.

**2.55 (a)**  $r = \pm 1$  for a line. **(c)** Leave some space above your vertical stack. **(d)** The curve must be higher at the right than at the left.

**2.57** The correlation is  $r = 0.481$ . The correlation is greatly lowered by the one outlier. Outliers tend to have fairly strong effects on correlation; it is even stronger here because there are so few observations.

**2.59** There is little linear association between research and teaching—for example, knowing a professor is a good researcher gives little information about whether she is a good or a bad teacher.

**2.61** Both relationships are somewhat linear; GPA/IQ ( $r = 0.634$ ) is stronger than GPA/self-concept ( $r = 0.542$ ). The two students with the lowest GPAs stand out in both plots; a few others stand out in at least one plot. Generally speaking, removing these points raises  $r$ , except for the lower-left point in the self-concept plot.

**2.63** 1.785 kilograms.

**2.65** Expressed as percents, these fractions are 64%, 16%, 4%, 0%, 9%, 25%, and 81%.

**2.67** The relationship is roughly linear. Bone strength in the dominant arm increases about 1.373 units for every unit increase in strength in the nondominant arm.

**2.69**  $22.854 \text{ cm}^4/1000$ .

**2.71 (a–c)**

		$\text{Count} = 602.8 - (74.7 \times \text{time})$
--	--	--

Time	Count	Predicted	Difference	Squared difference
1	578	528.1	49.9	2490.01
3	317	378.7	-61.7	3806.89
5	203	229.3	-26.3	691.69
7	118	79.9	38.1	1451.61

(d)

Count = 500 - (100 × time)				
Time	Count	Predicted	Difference	Squared difference
1	578	400	178	31,684
3	317	200	117	13,689
5	203	0	203	41,209
7	118	-200	318	101,124

2.73  $\hat{y} = -15,294.868 + 0.0533x$ , or, in context,

Student $\hat{s} = -15,294.868 + 0.0533$ Population

2.75 (a) 304, 505.13 students. (b) 299, 247.47 students. (c) Including the states with the largest populations (and largest numbers of undergraduates) increases the estimate by about 5000 students.

2.77 (a)

Student $\hat{s} = 8491.907 + 0.048$ Population. (b)  $r^2 = 0.942$ . (c) About 94.2% of the variability in number of undergraduates is explained by the regression on population. (d) The numerical output does not tell us whether the relation is linear.

2.79 (a)

Carb $\hat{s} = 2.606 + 1.789$ PercentAlcohol. (b)  $r^2 = 0.271$ .

2.81 (a) All correlations are approximately 0.816 or 0.817, and the regression lines are  $\hat{y} = 3.000 + 0.500x$ . We predict  $\hat{y} = 8$  when  $x = 10$ . (c) This regression formula is only appropriate for Set A.

2.83 (a) The added point is an outlier that does not follow the pattern of the rest of the data. It is an outlier in the  $x$  direction but not in  $y$ . (b) The new regression equation is  $\hat{y} = 27.56 + 0.1031x$ . (c)  $r^2 = 0.052$ . This added point is influential both to the regression equation (both the intercept and slope changed substantially from  $\hat{y} = 17.38 + 0.6233x$ ) and the correlation.

2.85 (a) 36. (b) When  $x$  increases one unit,  $y$  increases by 8 (the value of the slope). (c) The intercept is 12.

2.87 IQ and GPA:  $r_1 = 0.634$ . Self-concept and GPA:  $r_2 = 0.542$ . IQ does a slightly better job.

2.89 When  $\bar{x} = \bar{x}$ ,  $\hat{y} = a + b\bar{x} = (y - b\bar{x}) + b\bar{x} = y$ .

2.91 Scatterplots and correlations were found in Exercise 2.36 and 2.54. The regression equations are  $\text{Value} = 1073.87 + (1.74 \times \text{Debt})$ , with  $r^2 = 0.5\%$ ;  $\text{Value} = -262.4 + (4.966 \times \text{Revenue})$ , with  $r^2 = 92.7\%$ ; and  $\text{Value} = -872.6 + (5.695 \times \text{Income})$ , with  $r^2 = 79.4\%$ .

2.93 The residuals sum to 0.01.

2.95 The residuals are -4.93, -5.09, 0.01, and 7.71.

2.97 (a-b)

Time	LogCount	Predicted	Residual
1	6.35957	6.332444	0.027126
3	5.75890	5.811208	-0.05231
5	5.31321	5.289972	0.023238
7	4.77068	4.768736	0.001944

**2.99 (f)** In the log scale, California is no longer an outlier anywhere, nor is it influential.

**2.101 (c)** One data point stands out in both graphs; it is West Virginia, with the largest positive residual. The next-largest positive residual belongs to Iowa. These do not seem to be influential. **(d and e)** Using the log data removes California as a potentially influential outlier. The data are more equally spread across the range. One possible disadvantage of using the log data is that explaining this to people could be difficult.

**2.103 (a)** If the line is pulled toward the influential point, the observation will not necessarily have a large residual. **(b)** High correlation is always present if there is causation. **(c)** Extrapolation is using a regression to make predictions for  $x$ -values outside the range of the data (here, using 20, for example).

**2.105** Internet use does not cause people to have fewer babies. Possible lurking variables are economic status of the country, education levels, etc.

**2.107** For example, a reasonable explanation is that the cause-and-effect relationship goes in the other direction: doing well makes students or workers feel good about themselves rather than vice versa.

**2.109** The explanatory and response variables were “consumption of herbal tea” and “cheerfulness/health.” The most important lurking variable is social interaction; many of the nursing-home residents may have been lonely before the students started visiting.

**2.111 (a)** Drawing the “best line” by eye is a very inaccurate process. But with practice, you can get better.

**2.113** The plot should show a positive association when either group of points is viewed separately and should show a large number of bachelor’s degree economists in business and graduate degree economists in academia.

**2.115** 1278 met the requirements; 751 did not meet requirements.

**2.117** Divide the cell count by the total for the table.

**2.119**  $417/974 = 0.4281$  (which rounds to 43%).

**2.121 (a)** Drivers education course (yes/no) is the explanatory variable. The number of accidents is the response. **(b)** Drivers ed would be the column ( $x$ ) variable, and number of accidents would be the row ( $y$ ) variable. **(c)** There are 6 cells. For example, the first row, first column entry could be the number who took drivers ed and had 0 accidents.

**2.123 (a)** Age is the explanatory variable. “Rejected” is the response. With the dentistry available at that time, it’s reasonable to think that as a person got older, he would have lost more teeth. **(b)**

	<20	20–25	25–30	30–35	35–40	>40
Yes	0.0002	0.0019	0.0033	0.0053	0.0086	0.0114
No	0.1761	0.2333	0.1663	0.1316	0.1423	0.1196

**(c)**

Marginal distribution of “Rejected”	
Yes	No
0.03081	0.96919

Marginal distribution of age					
<20	20–25	25–30	30–35	35–40	>40
0.1763	0.2352	0.1696	0.1369	0.1509	0.1310

(d) The conditional distribution of Rejected given Age, because we have said Age is the explanatory variable. (e) In the table, note that all columns sum to 1.

	<20	20–25	25–30	30–35	35–40	>40
Yes	0.0012	0.0082	0.0196	0.0389	0.0572	0.0868
No	0.9988	0.9918	0.9804	0.9611	0.9428	0.9132

**2.125** Students with GPAs less than 2.0 are much more likely to enroll for 11 or fewer credits (68.5%). Students with GPAs above 3.0 are most likely to enroll for 15 or more credits (66.6%).

**2.127** (a) 50.5% get enough sleep; 49.5% do not. (b) 32.2% get enough sleep; 67.8% do not. (c) Those who exercise more than the median are more likely to get enough sleep.

**2.129** 3.0% of Hospital A’s patients died, compared with 2.0% at Hospital B.

**2.131** In general, choose  $a$  to be any number from 0 to 200, and then all the other entries can be determined.

**2.133** For example, causation might be a negative association between the temperature setting on a stove and the time required to boil a pot of water (higher setting, less time). Common response might be a positive association between SAT scores and grade point average. Both of these will respond positively to a person’s IQ. An example of confounding might be a negative association between hours of TV watching and grade point average. Once again, people who are naturally smart could finish required work faster and have more time for TV; those who aren’t as smart could become frustrated and watch TV instead of doing homework.

**2.135** This is a case of confounding: the association between dietary iron and anemia is difficult to detect because malaria and helminths also affect iron levels in the body.

**2.137** For example, students who choose the online course might have more self-motivation or better computer skills.

**2.139** No; self-confidence and improving fitness could be a common response to some other personality trait, or high self-confidence could make a person more likely to join the exercise program.

**2.141** Patients suffering from more serious illnesses are more likely to go to larger hospitals and may require more time to recuperate afterward.

**2.143** People who are overweight are more likely to be on diets and so choose artificial sweeteners.

**2.145** This is an observational study—students choose their “treatment” (to take or not take the refresher sessions).

**2.147** (a) The tables are shown below.

Female <i>Titanic</i> passengers				
	Class			
	1	2	3	Total
Survived	139	94	106	339
Died	5	12	110	127
Total	144	106	216	466

Male *Titanic* passengers

	Class			Total
	1	2	3	
Survived	61	25	75	161
Died	118	146	418	682
Total	179	171	493	843

**(b)** 96.53% of first-class females survived, 88.68% of second-class females survived, and 49.07% of third-class females survived. Survival depended on class. **(c)** 34.08% survival among first class, 14.62% survival among second class, and 15.21% survival among third class. Once again, survival depended on class. **(d)** Females overall had much higher survival rates than males.

**2.149 (a)** This is a negative relationship, mostly due to two outliers. **(b)**  $r = -0.839$ . This would not be a good numerical summary for this relationship.

**2.151 (b)** In Figure 2.33, we can see that the three territories have smaller proportions of their populations over 65 than the provinces. The two areas with the largest percents of the population under 15 are Nunavut and Northwest Territories.

**2.153 (a)** The relationship is weakly increasing and linear. We almost seem to have two sets of data: five countries with high production and the rest. One country with Dwelling Permit Index approximately 225 (Canada) might be influential. **(b)** The equation is  $\text{Production} = 110.96 + 0.0732 \text{ DwellingPermit}$ . **(c)** 122.672. **(d)**  $e = -13.672$ . **(e)**  $r^2 = 2.0\%$ . Both indicate very weak relationships, but this is weaker.

**2.155** A stacked bar graph clearly shows that offering the RDC service depends on size of the bank. Larger banks are much more likely to offer the service than smaller ones.

**2.157 (a)** The marginal totals are SsBL: 1688; SME: 911; AH: 801; Ed: 319; and Other: 857. By country, the marginal totals are Canada: 176; France: 672; Germany: 218; Italy: 321; Japan: 645; U.K.: 475; U.S.: 2069. **(b)** Canada: 0.0385; France: 0.1469; Germany: 0.0476; Italy: 0.0701; Japan: 0.1410; U.K.: 0.1038; U.S.: 0.4521. **(c)** SsBL: 0.3689; SME: 0.1991; AH: 0.1750; Ed: 0.0697; Other: 0.1873.

**2.159** A school that accepts weaker students but graduates a higher-than-expected number of them would have a positive residual, while a school with a stronger incoming class but a lower-than-expected graduation rate would have a negative residual. It seems reasonable to measure school quality by how much benefit students receive from attending the school.

**2.163 (a)** The residuals are positive at the beginning and end, and negative in the middle. **(b)** The behavior of the residuals agrees with the curved relationship seen in Figure 2.34.

**2.165 (a)** The regression equation for predicting salary

from year is  $\text{Salary} = 41.253 + 3.9331 \text{ Year}$ ; for Year 25, the predicted salary is 139.58 thousand dollars, or about \$139,600. **(b)** The log salary regression equation is  $\ln \text{Salary} = 3.8675 + 0.04832 \text{ Year}$ . At Year 25 the predicted salary is  $e^{5.0755} = 160.052$ , or about \$160,050. **(c)** Although both predictions involve extrapolation, the second is more reliable because it is based on a linear fit to a linear relationship. **(d)** Interpreting relationships without a plot is risky.

**2.167 (a)** The regression equation is  $2013\text{Salary} = 6523 + 0.97291 \times (2012\text{Salary})$ . **(b)** The residuals appear rather random, but we note that the largest positive residuals are on either end of the scatterplot. The largest negative residual is for the next-to-highest 2012–2013 salaried person.

**2.169** Number of firefighters and amount of damage are common responses to the seriousness of the fire.

**2.171 (b)** The regression line  $\text{PctCollEd} = 4.033 + 0.906 \text{ FruitVeg5}$  generally describes the relationship. There is one outlier at the upper right of the scatterplot (Washington, DC). **(d)** While the scatterplot and regression support a positive association between college degrees and eating fruits and vegetables, association is not causation.

**2.173** The scatterplot of MOR against MOE shows a moderate positive linear association. The regression equation is  $\widehat{\text{MOR}} = 2653 + 0.004742\text{MOE}$ ; this regression explains  $r^2 = 0.6217$ , or about 62% of the variation in MOR. So we can use MOE to get fairly good (though not perfect) predictions of MOR.

**2.175 (a)**

	Admit	Deny
Male	490	310
Female	400	300

**(b)** Males: 61.25% admitted. Females: 57.14% admitted. **(c)** Business school: 66.67% of males, 66.67% of females. Law school: 45% of males, 50% of females. **(d)** Most male applicants apply to the business school, where admission is easier. More women apply to the law school, which is more selective.

**2.177** If we ignore “year,” Department A teaches 61.5% small classes while Department B teaches 39.6% small classes. However, in upper-level classes, 77.5% of A’s classes are small and 83.3% of B’s classes are small. Department A has 77.8% of its classes as upper-level, while only 33.96% of B’s classes are upper level.

## CHAPTER 3

**3.1** Answers will vary. One possibility is that the friend has a weak immune system.

**3.3** Answers will vary, but the individual’s denial is clearly insufficient evidence to conclude that he did not use performance enhancing drugs.

**3.5** For example, who owns the web site? Do they have data to back up this statement, and if so, what was the source of those data?

**3.7** Available data are from prior studies. They might be from either observational studies or experiments.

**3.9** This is not an experiment (running them until the batteries die is not assigning treatments) or a sample survey.

**3.11** This is an experiment. Explanatory variable: apple form (juice or whole fruit); response variable: how full the subject felt.

**3.13** The data were likely from random samples of cans of tuna.

**3.15 (a)** Anecdotal data. **(b)** This is a sample survey but likely biased. **(c)** Still a survey but random. **(d)** Answers will vary.

**3.17** In Exercise 3.14: extra milk and no extra milk. In Exercise 3.16: no pills, pills without echinacea, pills with echinacea but subjects weren’t told, and pills with echinacea that were labeled as containing echinacea.

**3.19** *Treatments* are the four coaching types that were actively assigned to the *experimental units* (subjects), who were 204 people. *Factor* is type of coaching, with four levels: increase fruit and vegetable intake and physical activity, decrease fat and sedentary leisure, decrease fat and increase physical activity, and increase fruit and vegetable intake and decrease sedentary leisure. *Response* is the measure of diet and exercise improvement after three weeks. This experiment had a very high completion rate.

**3.21** With 719 subjects, randomly assign 180 to each of the first three treatments and 179 to the last (echinacea with the labeled bottle). Afterward, compare diet and exercise improvement.

**3.23** Answers will vary due to software.

**3.25 (a)** Experimental units were the 30 students. They are human, so we can use “subjects.” **(b)** Only one “treatment,” so not comparative. One possibility is to randomly assign half to the online homework system and half to “standard” homework. **(c)** One possibility is grade on an exam over the material from that month.

**3.27 (a)** Experimental units (subjects): people who go to the web site. Treatments: description of comfort or showing discounted price. Response variable: shoe sales. **(b)** Comparative, because of two treatments. **(c)** One option to improve: randomly assign morning and afternoon treatments. **(d)** Placebo (no special description or price) could give a “baseline” sales figure.

**3.29** Starting on line 101, using 1 to 5 as morning and 6 to 0 as afternoon for comfort description, we have 19223 95034: comfort description is displayed in the afternoon on Days 2, 6, and 8 and in the morning on the other days.

**3.31** Yes; each customer (who returns) will get both treatments.

**3.33 (a)** Shopping patterns may differ on Friday and Saturday. **(b)** Responses may vary in different states. **(c)** A control is needed for comparison.

**3.35** For example, new employees should be randomly assigned to either the current program or the new one.

**3.37 (a)** Factors: calcium dose and vitamin D dose. There are nine treatments (each calcium/vitamin D combination). **(b)** Assign 20 students to each group, with 10 of each gender. **(d)** Placebo is the 0 mg calcium, 0 mg vitamin D treatment.

**3.39** There are nine treatments. Choose the number of letters in each group, and send them out at random times over several weeks.

**3.41 (a)** Population = 1 to 150, sample size = 25, then click “Reset” and “Sample.” **(b)** Without resetting, click “Sample” again. **(c)** Continue to “Sample” from those remaining.

**3.43** Design (a) is an experiment, while (b) is an observational study; with the first, any difference in colon health between the two groups could be attributed to the treatment (bee pollen or not).

**3.45 (a)** Randomly assign half the girls to get highcalcium punch; the other half will get low-calcium punch. Observe how each group processes the calcium. **(b)** Half receive high-calcium punch first; the rest get low-calcium punch first. For each subject, compute the difference in the response variable for each level. Matched pairs designs give more precise results. **(c)** The first five subjects are 35, 39, 16, 04, and 26.

**3.47** Answers will vary. For example, the trainees and experienced professionals could evaluate the same water samples.

**3.49** Population: forest owners from this region. Sample: the 348 returned questionnaires. Response rate:  $348/772 = 45\%$ . Additionally, we would like to know the sample design (among other things).

**3.51** Answers will vary depending on use of software or starting point in Table B.

**3.53 (a)** Season ticket holders. **(b)** 98 responses received. **(c)**  $98/150 = 65.3\%$ . **(d)** 34.7%. **(e)** One possibility is to offer incentives (free hotdog?).

**3.55 (a)** Answers will vary depending on use of software. **(b)** Software is usually more efficient than Table B.

**3.57 (a)** The population is all items/individuals of potential interest. **(b)** Many people probably will not realize that dihydrogen monoxide is water. **(c)** In a public setting, few people will admit to cheating.

**3.59** Population: all local businesses. Sample: the 72 returned questionnaires. Nonresponse rate: 55%.

**3.61** Note that the numbers add to 100% down the columns; that is, 39% is the percent of Fox viewers who are Republicans, *not* the percent of Republicans who watch Fox.

- 3.63** Labeled in alphabetical order, using line 126: 31 (Village Manor), 08 (Burberry), 19 (Franklin Park), 03 (Beau Jardin), and 25 (Pemberley Courts).
- 3.65** Population = 1 to 200, sample size = 20, then click “Reset” and “Sample.” Selections will vary.
- 3.67** Labeling the tracts in numerical order from 01 (block 1000) to 44 (block 3025), the selected random digits are labeled 21 (block 3002), 37 (block 3018), 18 (block 2011), 44, 23, 19, 10, 33, and 31.
- 3.69** Answers will vary. Beginning on line 110, from Group 1 (labeled 1 through 6), select 3 and 4. Continuing from there, from Group 2 (labeled 01 through 12), select 08 and 05. Continuing from there, from Group 3 (labeled 01 through 26), select 13, 09, and 04.
- 3.71** Each student has chance  $1/40$  of being selected, but the sample is not an SRS, because the only possible samples have exactly one name from the first 40, one name from the second 40, and so on.
- 3.73** Number the parcels 01 through  $n$  for each forest type. Using Table B, select Climax 1: 05, 16, 17, 40, and 20; Climax 2: 19, 45, 05, 32, 19, and 41; Climax 3: 04, 19, and 25; Secondary: 29, 20, 43, and 16.
- 3.75** Each individual has a 1-in-8 chance of being selected.
- 3.77 (a)** Households without telephones or with unlisted numbers. Such households would likely be made up of poor individuals, those who choose not to have phones, and those who do not wish to have their phone number published. Households with only cell phones are also not included. **(b)** Those with unlisted numbers. Or only cell phones.
- 3.79** The female and male students who responded are the samples. The populations are all college undergraduates (males and females) who could be judged to be similar to the respondents. This report is incomplete; a better one would give numbers about who responded, as well as the actual response rate.
- 3.81** The larger sample would have less sampling variability.
- 3.83** Answers will vary due to computer simulation. You should have a mean close to 0.5 and a standard deviation close to 0.204.
- 3.85** Answers will vary due to computer simulation. You should have a mean close to 0.5 and a standard deviation close to 0.08.
- 3.87** Answers will vary due to computer simulation. You should have a mean close to 0 and a standard deviation close to 1.
- 3.89 (a)** The larger sample size should have a smaller standard deviation (less variability).
- 3.91 (a)** Population: Students at four-year colleges in the U.S. Sample: 17,096 students. **(b)** Population: restaurant workers. Sample: 100 workers. **(c)** Population: 584 longleaf pine trees. Sample: 40 trees.
- 3.93** The histograms should be centered at about 0.6 with standard deviation about 0.1.
- 3.95** Answers will vary due to computer simulation.
- 3.97 (a)** Nonscientists might have different viewpoints and raise different concerns from those considered by scientists.
- 3.99** Answers will vary. This question calls for a reasoned opinion.
- 3.101** Answers will vary. This question calls for a reasoned opinion.
- 3.103** No. Informed consent needs clear information on what will be done.
- 3.105** Answers will vary. This question calls for a reasoned opinion.

- 3.107** Answers will vary. This question calls for a reasoned opinion.
- 3.109** The samples should be randomly ordered for analysis.
- 3.111** Interviews conducted in person cannot be anonymous.
- 3.113** Answers will vary. This question calls for a reasoned opinion.
- 3.115 (a)** Informed consent requires informing respondents about how the data will be used, how long the survey will take, etc.
- 3.117** Answers will vary. This question calls for a reasoned opinion.
- 3.119** Answers will vary. This question calls for a reasoned opinion.
- 3.121 (a)** You need information about a random selection of his games, not just the ones he chooses to talk about. **(b)** These students may have chosen to sit in the front; all students should be assigned to their seats.
- 3.123** This is an experiment because treatments are assigned. Explanatory variable: price history (steady or fluctuating). Response variable: price the subject expects to pay.
- 3.127** Randomly choose the order in which the treatments (gear and steepness combination) are tried.
- 3.129 (a)** One possibility: full-time undergraduate students in the fall term on a list provided by the registrar. **(b)** One possibility: a stratified sample with 125 students from each class rank. **(c)** Nonresponse might be higher with mailed (or emailed) questionnaires; telephone interviews exclude some students and may require repeated calling for those who do not answer; face-to-face interviews might be too costly. The topic might also be subject to response bias.
- 3.131** Use a block design: separate men and women, and randomly allocate each gender among the six treatments.
- 3.133** CASI will typically produce more honest responses to embarrassing questions.
- 3.135** Answers will vary. This question calls for a reasoned opinion.
- 3.137** Answers will vary. This question calls for a reasoned opinion.

## CHAPTER 4

- 4.1** The proportion of heads is 0.5. In this case, we did get exactly 10 heads (this will NOT happen every time).
- 4.3 (a)** This is random. We can discuss the probability (chance) that the temperature would be between 30 and 35 degrees, for example. **(b)** Depending on your school, this is not random. At my university, all student IDs begin with 900. **(c)** This is random. The probability of an ace in a single draw is  $\frac{4}{52}$  if the deck is well shuffled.
- 4.5** Answers will vary depending on your set of 25 rolls.
- 4.7** If you hear music (or talking) one time, you will almost certainly hear the same thing for several more checks after that.
- 4.9** The theoretical probability is 0.5177. What were the results of your “rolls”?
- 4.11** One possibility: from 0 to \_\_\_\_ hours (the largest number should be big enough to include all possible responses).

4.13 0.80 (add the probabilities of the other four colors and subtract from 1).

4.15 0.681.

4.17  $1/4$ , or 0.25.

4.19 (a)  $S = \{\text{Yes, No}\}$ . (b)  $S = \{0, 1, 2, \dots, n\}$  where  $n$  is large enough to include a really busy tweeter. (c)  $S = [18, 75]$  is one possibility. This is given as an interval because age is a continuous variable. (d)  $S = \{\text{Accounting, Archeology, } \dots\}$ . This list could be *very* long.

4.21 (a) Not equally likely (check the web). (b) Equally likely. (c) This could depend on the intersection; is the turn onto a one-way street? (d) Not equally likely.

4.23 (a) The probability that both of two disjoint events occur is 0. (b) Probabilities must be no more than 1. (c)  $P(A^c) = 0.65$ .

4.25 There are 6 possible outcomes:  $S = \{\text{link 1, link 2, link 3, link 4, link 5, leave}\}$ .

4.27 (a) 0.172. (b) 0.828.

4.29 (a) 0.03, so the sum equals 1. (b) 0.55.

4.31 (a) The probabilities sum to 2. (b) Legitimate (for a nonstandard deck). (c) Legitimate (for a nonstandard die).

4.33 (a) 0.28. (b) 0.88.

4.35 Take each blood type probability and multiply by 0.84 and by 0.16. For example, the probability for A-positive blood is  $(0.42)(0.84) = 0.3528$ .

4.37 (a) 0.006. (b) 0.001.

4.39 0.5160.

4.41 Observe that  $P(A \text{ and } B^c) = P(A) - P(A \text{ and } B) = P(A) - P(A)P(B)$ .

4.43 (a) Either B or O. (b)  $P(B) = 0.75$ , and  $P(O) = 0.25$ .

4.45 (a) 0.25. (b) 0.015625; 0.140625.

4.47 Possible values: 0, 1, 2. Probabilities:  $1/4, 1/2, 1/4$ .

4.49

$x$	1	2	3	4	5	6
$P(X=x)$	0.05	0.05	0.13	0.26	0.36	0.15

4.51 (a) 0.23. (b) 0.62. (c) 0.

4.53 (a) Discrete *random variables*. (b) Continuous random variables can take values from any interval. (c) Normal random variables are continuous.

4.55 (a)  $P(T) = 0.19$ . (b)  $P(TTT) = 0.0069$ ,  $P(TTT^c) = P(TT^cT) = P(T^cTT) = 0.0292$ ,  $P(TT^cT^c) = P(T^cTT^c) = P(T^cT^cT) = 0.1247$ , and  $P(T^cT^cT^c) = 0.5314$ . (c)  $P(X=3) = 0.0069$ ,  $P(X=2) = 0.0876$ ,  $P(X=1) = 0.3741$ , and  $P(X=0) = 0.5314$ .

4.57 (a) Continuous. (b) Discrete. (c) Discrete.

4.59 (a) Note that, for example, “(1, 2)” and “(2, 1)” are distinct outcomes. (b)  $1/36$ . (c) For example, four pairs add to 5, so  $P(X=5) = 4/36 = 1/9$ . (d)  $2/9$ . (e)  $5/6$ .

- 4.61 (b)**  $P(X \geq 1) = 0.9$ . **(c)** “No more than two nonword errors.”  $P(X \leq 2) = 0.7$ ;  $P(X < 2) = 0.4$ .
- 4.63 (a)** The height should be  $1/2$ . **(b)** 0.8. **(c)** 0.6. **(d)** 0.525.
- 4.65** Very close to 1.
- 4.67** Possible values: \$0 and \$5. Probabilities: 0.5 and 0.5. Mean: \$2.50.
- 4.69**  $\mu_Y = 68$ .
- 4.71**  $\sigma_{X^2} = 2.16$  and  $\sigma_X = 1.47$ .
- 4.73** As the sample size gets larger, the standard deviation decreases. The mean for 1000 will be much closer to  $\mu$  than the mean for 2 (or 100) observations.
- 4.75**  $\sigma_{X^2} = 1.45$  and  $\sigma_X = 1.204$ .
- 4.77 (a)** 202. **(b)** 198. **(c)** 60. **(d)** -20. **(e)** -140.
- 4.79** Mean = 2.2 servings.
- 4.81** 0.373 aces.
- 4.83 (a)** \$85.48. **(b)** This is larger; the negative correlation decreased the variance.
- 4.85** The exercise describes a positive correlation between calcium intake and compliance. Because of this, the variance of total calcium intake is greater than the variance we would see if there were no correlation.
- 4.87 (a)**  $\mu = \sigma = 0.5$ . **(b)**  $\mu_4 = 2$  and  $\sigma_4 = 1$ .
- 4.89 (a)** Not independent. **(b)** Independent.
- 4.91** If 1 of the 10 homes were lost, it would cost more than the collected premiums. For many policies, the average claim should be close to \$300.
- 4.93 (a)** 0.99749. **(b)** \$623.22.
- 4.95**  $1/2 = 0.5$ .
- 4.97**  $2/48 = 1/24$ .
- 4.99** The addition rule for disjoint events.
- 4.101** With 23 cards seen, there are 29 left to draw from. The four probabilities are  $45/406$ ,  $95/406$ ,  $95/406$ , and  $171/406$ .
- 4.103 (a)** 0.8. **(b)** 0.2.
- 4.105 (a)**  $5/6 = 0.833$ .
- 4.107 (a)** A = 5 to 10 years old, B = 11 to 13 years old, C = adequate calcium intake, I = inadequate calcium intake. **(b)**  $P(A) = 0.52$ ,  $P(B) = 0.48$ ,  $P(I|A) = 0.18$ ,  $P(I|B) = 0.57$ . **(c)**  $P(I) = 0.3672$ .
- 4.109** Not independent.  $P(I|A) = 0.18$ ,  $P(I|B) = 0.57$ . These are different.
- 4.111 (a)** 0.16. **(b)** 0.22. **(c)** 0.38. **(d)** For (a) and (b), use the addition rule for disjoint events; for (c), use the addition rule, and note that  $S^c$  and  $E^c = (S \text{ or } E)^c$ .
- 4.113** 0.73; use the addition rule.

- 4.115** (a) The four entries are 0.2684, 0.3416, 0.1599, 0.2301. (b) 0.5975.
- 4.117** For example, the probability of selecting a female student is 0.5717; the probability that she comes from a 4-year institution is 0.5975.
- 4.119**  $P(A|B) = 0.3142$ . If  $A$  and  $B$  were independent, then  $P(A|B)$  would equal  $P(A)$ .
- 4.121** (a)  $P(A^c) = 0.69$ . (b)  $P(A \text{ and } B) = 0.08$ .
- 4.123** 1.
- 4.125** (a)  $P(B|C) = 1/3$ .  $P(C|B) = 0.2$ .
- 4.127** (a)  $P(M) = 0.3959$ . (b)  $P(B|M) = 0.6671$ . (c)  $P(M)P(B) = 0.2521$ , so these are not independent.
- 4.129** (a) Her brother has allele type  $aa$ , and he got one allele from each parent. (b)  $P(aa) = 0.25$ ,  $P(Aa) = 0.5$ ,  $P(AA) = 0.25$ . (c)  $P(AA|\text{not } aa) = 1/3$ ,  $P(Aa|\text{not } aa) = 2/3$ .
- 4.131** 0.9333.
- 4.133** Close to  $\mu_X = 1.4$ .
- 4.135** (a) Possible values 2 and 14, with probabilities 0.4 and 0.6, respectively. (b)  $\mu_Y = 9.2$  and  $\sigma_Y = 5.8788$ . (c) There are no rules for a quadratic function of a random variable; we must use definitions.
- 4.137** (a)  $P(A) = 1/36$  and  $P(B) = 15/36$ . (b)  $P(A) = 1/36$  and  $P(B) = 15/36$ . (c)  $P(A) = 10/36$  and  $P(B) = 6/36$ . (d)  $P(A) = 10/36$  and  $P(B) = 6/36$ .
- 4.139** For example, if the point is 4 or 10, the expected gain is  $(1/3)(+20) + (2/3)(-10) = 0$ .
- 4.141** (a) All probabilities are greater than or equal to 0, and their sum is 1. (b) 0.61. (c) Both probabilities are 0.39.
- 4.143** 0.005352.
- 4.145** 0.6817.
- 4.147**  $P(\text{no point}) = 1/3$ . The probability of winning (losing) an odds bet is  $1/36$  ( $1/18$ ) on 4 or 10,  $2/45$  ( $1/15$ ) on 5 or 9,  $25/396$  ( $5/66$ ) on 6 or 8.
- 4.149** 0.1622.
- 4.151**  $P(Y < 1/3 | Y > X) = 1/9$ .

## CHAPTER 5

- 5.1** Population: iPhone users. Statistic: a median of 108 apps per device. Likely values will vary.
- 5.3**  $\mu_{\bar{x}} = 420$ ,  $\sigma_{\bar{x}} = 1$ .
- 5.5** About 95% of the time,  $\bar{x}$  is between 181 and 189.
- 5.7** (a) Each sample size has  $\mu_{\bar{x}} = 1$ . For  $n = 2$ ,  $\sigma_{\bar{x}} = 0.707$ . For  $n = 10$ ,  $\sigma_{\bar{x}} = 0.316$ . For  $n = 25$ ,  $\sigma_{\bar{x}} = 0.2$ .
- 5.9** (a) The standard deviation for  $n = 10$  will be  $\sigma_{\bar{x}} = 20/10$ . (b) Standard deviation decreases with increasing sample size. (c)  $\mu_{\bar{x}}$  always equals  $\mu$ .
- 5.11** (a)  $\mu = 125.5$ . (b) Answers will vary. (c) Answers will vary. (d) The center of the histogram represents an average of averages.

**5.13 (a)** Both populations are smartphone users. They likely are comparable. **(b)** Excluding those with no apps will increase the median because you are eliminating individuals.

**5.15 (a)** Larger. **(b)** We need  $\sigma\bar{x} \leq 0.085$ . **(c)** The smallest sample size that will fit this criterion is  $n = 213$ .

**5.17**  $\mu\bar{x} = 250$ .  $\sigma\bar{x} = 0.25$ .

**5.19 (b)** To be more than 1 ml away from the target value means the volume is less than 249 or more than 251. Using symmetry,  $P = 2P(X < 249) = 2P(z < -2) = 2(0.0228) = 0.0456$ . **(c)**  $P = 2P(X < 249) = 2P(z < -4) \approx 0$ . (Software gives 0.00006.)

**5.21 (a)**  $\bar{x}$  is not systematically higher than or lower than  $\mu$ . **(b)** With large samples,  $\bar{x}$  is more likely to be close to  $\mu$ .

**5.23 (a)**  $\mu\bar{x} = 0.3$ .  $\sigma\bar{x} = 0.08$ . **(b)** 0.0062. **(c)**  $n = 100$  is a large enough sample to be able to use the central limit theorem.

**5.25 (a)** 0.0668. **(b)** 0.0047.

**5.27** 134.5 mg/dl.

**5.29** 0.0051.

**5.31 (a)**  $N(0.5, 0.332)$ . **(b)** 0.0655. Software gives a probability of 0.0661.

**5.33 (a)**  $\bar{y}$  has a  $N(\mu_Y, \sigma_Y/m)$  distribution, and  $\bar{x}$  has a  $N(\mu_X, \sigma_X/n)$  distribution. **(b)**  $\bar{y} - \bar{x}$  has a Normal distribution with mean  $\mu_Y - \mu_X$  and standard deviation  $\sigma_Y^2 + \sigma_X^2$ .

**5.35**  $n = 1965$ .  $X = 0.48 \times 1965 = 943$ .  $\hat{p} = 0.48$ .

**5.37 (a)**  $n = 1500$ . **(b)** Answers and reasons will vary. **(c)** If the choice is “Yes,”  $X = 1025$ . **(d)** For “Yes,”  $\hat{p} = 1025/1500 = 0.683$ .

**5.39**  $B(10, 0.5)$ .

**5.41 (a)**  $P(X = 0) = 0.0467$  and  $P(X \geq 4) = 0.1792$ . **(b)**  $P(X = 6) = 0.0467$  and  $P(X \leq 2) = 0.1792$ . **(c)** The number of “failures” in the  $B(6, 0.4)$  distribution has the  $B(6, 0.6)$  distribution. With 6 trials, 0 successes is equivalent to 6 failures, and 4 or more successes is equivalent to 2 or fewer failures.

**5.43 (a)** 0.9953. **(b)** 0.8415. Using software gives 0.8422.

**5.45 (a)** 0.1563. **(b)** 0.7851.

**5.47 (a and b)** The coin is fair. The probabilities are still  $P(H) = P(T) = 0.5$ . Separate flips are independent (coins have no “memory”), so regardless of the results of the first four tosses, the fifth is equally likely to be a head or a tail. **(c)** The parameters for a binomial distribution are  $n$  and  $p$ . **(d)** This is best modeled with a Poisson distribution.

**5.49 (a)** A  $B(200, p)$  distribution seems reasonable for this setting (even though we do not know what  $p$  is). **(b)** This setting is not binomial; there is no fixed value of  $n$ . **(c)** A  $B(500, 1/12)$  distribution seems appropriate for this setting. **(d)** This is not binomial because separate cards are not independent.

**5.51 (a)** The distribution of those who say they have stolen something is  $B(10, 0.2)$ . The distribution of those who do not say they have stolen something is  $B(10, 0.8)$ . **(b)**  $X$  is the number who say they have stolen something.  $P(X \geq 4) = 1 - P(X \leq 3) = 0.1209$ .

**5.53 (a)** Stole:  $\mu = 2$ ; did not steal:  $\mu = 8$ . **(b)**  $\sigma = 1.265$ . **(c)** If  $p = 0.1$ ,  $\sigma = 0.949$ . If  $p = 0.01$ ,  $\sigma = 0.315$ . As  $p$  gets smaller, the standard deviation becomes smaller.

**5.55 (a)**  $P(X \leq 7) = 0.0172$  and  $P(X \leq 8) = 0.0566$ , so 7 is the largest value of  $m$ . **(b)**  $P(X \leq 5) = 0.0338$  and  $P(X \leq 6) = 0.0950$ , so 5 is the largest value of  $m$ . **(c)** The probability will decrease.

**5.57** The count of 5s among  $n$  random digits has a binomial distribution with  $p = 0.1$ . **(a)** 0.4686. **(b)**  $\mu = 4$ .

**5.59 (a)**  $n = 4, p = 0.7$ . **(b)**

$x$	0	1	2	3	4
$P(X=x)$	0.0081	0.0756	0.2646	0.4116	0.2401

**(c)**  $\mu = 4(0.7) = 2.8$ , and  $\sigma = 4(0.7)(1-0.7) = 0.9165$ .

**5.61 (a)** Because  $(0.7)(300) = 210$  and  $(0.3)(300) = 90$ , the approximate distribution is  $\hat{p} \sim N(0.7, (0.7)(0.3)/300 = 0.0265)$ .  $P(0.67 < \hat{p} < 0.73) = 0.7416$  (Software gives 0.7424). **(b)** If  $p = 0.9$ , the distribution of  $\hat{p}$  is approximately  $N(0.9, 0.0173)$ .  $P(0.87 < \hat{p} < 0.93) = 0.9164$  (Software gives 0.9171). **(c)** As  $p$  gets closer to 1, the probability of being within  $\pm 0.03$  of  $p$  increases.

**5.63 (a)** The mean is  $\mu = p = 0.69$ , and the standard deviation is  $\sigma = p(1-p)/n = 0.0008444$ . **(b)**  $\mu \pm 2\sigma$  gives the range 68.83% to 69.17%. **(c)** This range is considerably narrower than the historical range. In fact, 67% and 70% correspond to  $z = -23.7$  and  $z = 11.8$ .

**5.65 (a)**  $\hat{p} = 0.28$ . **(b)** 0.0934 using Table A. Software gives 0.0927 without rounding intermediate values. **(c)** Answers will vary.

**5.67 (a)**  $p = 1/4 = 0.25$ . **(b)**  $P(X \geq 10) = 0.0139$ . **(c)**  $\mu = np = 5$  and  $\sigma = 3.75 = 1.9365$  successes. **(d)** No. The trials would not be independent.

**5.69 (a)**  $X$ , the count of successes, has the  $B(900, 1/5)$  distribution, with mean  $\mu = 180$  and  $\sigma = 12$  successes. **(b)** For  $\hat{p}$ , the mean is  $\mu_{\hat{p}} = p = 0.2$  and  $\sigma_{\hat{p}} = 0.01333$ . **(c)**  $P(\hat{p} > 0.24) = 0.0013$ . **(d)** 208 or more successes.

**5.71 (a)** 0.1788. **(b)** 0.0594. **(c)** 400. **(d)** Yes.

**5.73**  $Y$  has possible values 1, 2, 3, ...  $P(\text{first } \square \text{ appears on toss } k) = (5/6)^{k-1}(1/6)$ .

**5.75 (a)**  $\mu = 50$ . **(b)** The standard deviation is  $\sigma = 50 = 7.071$ .  $P(X > 60) = 0.0793$ . Software gives 0.0786.

**5.77 (a)**  $\bar{x}$  has (approximately) an  $N(123 \text{ mg}, 0.04619 \text{ mg})$  distribution. **(b)**  $P(\bar{x} \geq 124)$  is essentially 0.

**5.79 (a)** Approximately Normal with mean  $\mu_{\bar{x}} = 2.13$  and standard deviation  $\sigma_{\bar{x}} = 0.159$ . **(b)**  $P(\bar{x} < 2) = 0.2061$ . Software gives 0.2068. **(c)** Yes, because  $n = 140$  is large.

**5.81** 0.0034.

**5.83** If the carton weighs between 755 and 830 g, then the average weight of the 12 eggs must be between  $755/12 = 62.92$  and  $830/12 = 69.17$  g. The distribution of the mean weight is  $N(66, 6/12 = 1.732)$ .  $P(62.92 < \bar{x} < 69.17) = 0.9288$ .

**5.85 (a)** He needs 14.857 (really 15) wins. **(b)**  $\mu = 13.52$  and  $\sigma = 3.629$ . **(c)** Without the continuity correction,  $P(X \geq 15) = 0.3409$ . With the continuity correction, we have  $P(X \geq 14.5) = 0.3936$ . The continuity correction is much closer.

**5.87 (a)**  $\hat{p}^F$  is approximately  $N(0.82, 0.01921)$ , and  $\hat{p}^M$  is approximately  $N(0.88, 0.01625)$ . **(b)** When we subtract two independent Normal random variables, the difference is Normal. The new mean is the difference between the two means ( $0.88 - 0.82 = 0.06$ ), and the new variance is the sum of the variances ( $0.000369 + 0.000264 = 0.000633$ ), so  $\hat{p}^M - \hat{p}^F$  is approximately  $N(0.06, 0.02516)$ . **(c)** 0.0087 (software: 0.0085).

**5.89**  $P(Y \geq 200) = P(Y/500 \geq 0.4) = P(Z \geq 4.56) = 0$ .

## CHAPTER 6

6.1  $\sigma_{\bar{x}} = \$0.40$ .

6.3 \$0.80.

6.7 The margin of error will be halved.

6.9  $n = 285$ .

6.11 The students who did not respond are (obviously) not represented in the results. They may be more (or less) likely to use credit cards.

6.13 Margins of error: 17.355, 12.272, 8.677, and 6.136; interval width decreases with increasing sample size.

6.15 (a) She did not divide the standard deviation by 500=22.361. (b) Confidence intervals concern the *population* mean. (c) 0.95 is a confidence level, not a probability. (d) The large sample size affects the distribution of the sample mean (by the central limit theorem), not the individual ratings.

6.17 (a) The margin of error is 0.244; the interval is 5.156 to 5.644. (b) The margin of error is 0.321; the interval is 5.079 to 5.721.

6.19 Margin of error, 2.29 U/l. Interval, 10.91 to 15.49 U/l.

6.21 Scenario A has a smaller margin of error; less variability in a single class rank.

6.23 (a)  $\pm 18.98$ . (b)  $\pm 18.98$ .

6.25 No; this is a range of values for the mean rent, not for individual rents.

6.27 (a) 11.03 to 11.97 hours. (b) No; this is a range of values for the mean time spent, not for individual times. (c) The sample size is large ( $n = 1200$  students surveyed).

6.29 (a) Not certain (only 95% confident). (b) We obtained the interval 86.5% to 88.5% by a method that gives a correct result 95% of the time. (c) About 0.51%. (d) No (only random sampling error).

6.31  $\bar{x} = 18.3515$  kpl; the margin of error is 0.6521 kpl.

6.33  $n = 73$ .

6.35 No; confidence interval methods can be applied only to an SRS.

6.37 (a) 0.7738. (b) 0.9774.

6.39  $H_0: \mu = 1.4 \text{ g/cm}^2$ ;  $H_a: \mu \neq 1.4 \text{ g/cm}^2$ .

6.41  $P = 0.0702$  (Software gives 0.0703).

6.43 (a) 1.645. (b)  $z > 1.645$ .

6.45 (a)  $z = 1.875$ . (b)  $P = 0.0301$  (Software gives 0.0304). (c)  $P = 0.0602$  (Software gives 0.0608).

6.47 (a) No. (b) Yes.

6.49 (a) Yes. (b) No. (c) To reject, we need  $P < \alpha$ .

6.51 (a)  $P = 0.031$  and  $P = 0.969$ . (b) We need to know whether the observed data (for example,  $\bar{x}$ ) are consistent with  $H_a$ . (If so, use the smaller  $P$ -value.)

**6.53** (a) Population mean, not sample mean. (b)  $H_0$  should be that there is no change. (c) A small  $P$ -value is needed for significance. (d) Compare  $P$ , not  $z$ , with  $\alpha$ .

**6.55** (a)  $H_0: \mu = 77$ ;  $H_a: \mu \neq 77$ . (b)  $H_0: \mu = 20$  seconds;  $H_a: \mu > 20$  seconds. (c)  $H_0: \mu = 880 \text{ ft}^2$ ;  $H_a: \mu < 880 \text{ ft}^2$ .

**6.57** (a)  $H_0: \mu = \$42,800$ ;  $H_a: \mu > \$42,800$ . (b)  $H_0: \mu = 0.4 \text{ hr}$ ;  $H_a: \mu \neq 0.4 \text{ hr}$ .

**6.59** (a)  $P = 0.9545$ . (b)  $P = 0.0455$ . (c)  $P = 0.0910$ .

**6.61**  $P = 0.09$  means there is some evidence for the wage decrease, but it is not significant at the  $\alpha = 0.05$  level.

**6.63** The difference was large enough that it would rarely arise by chance. Health issues related to alcohol use are probably discussed in the health and safety class.

**6.65** The report can be made for public school students but not for private school students. Not finding a significant increase is not the same as finding no difference.

**6.67**  $z = 4.14$ , so  $P = 0.00003$  (for a two-sided alternative).

**6.69**  $H_0: \mu = 100$ ;  $H_a: \mu \neq 100$ ;  $z = 5.75$ ; significant ( $P < 0.0001$ ).

**6.71** (a)  $z = 2.13$ ,  $P = 0.0166$ . (b) The important assumption is that this is an SRS. We also assume a Normal distribution, but this is not crucial provided there are no outliers and little skewness.

**6.73** (a)  $H_0: \mu = 0 \text{ mpg}$ ;  $H_a: \mu \neq 0 \text{ mpg}$ , where  $\mu$  is the mean difference. (b)  $z = 4.07$ , which gives a very small  $P$ -value.

**6.75** (a)  $H_0: \mu = 0.61 \text{ mg}$ ;  $H_a: \mu > 0.61 \text{ mg}$ . (b) Yes. (c) No.

**6.77**  $\bar{x} = 0.8$  is significant, but  $0.7$  is not. Smaller  $\alpha$  means that  $\bar{x}$  must be farther away.

**6.79**  $\bar{x}$  will be statistically significant. With a larger sample size,  $\bar{x}$  close to  $\mu_0$  will be significant.

**6.81** Changing to the two-sided alternative multiplies each  $P$ -value by 2.

$\bar{x}$	$P$	$\bar{x}$	$P$
0.1	0.7518	0.6	0.0578
0.2	0.5271	0.7	0.0269
0.3	0.3428	0.8	0.0114
0.4	0.2059	0.9	0.0044
0.5	0.1139	1	0.0016

**6.83** Something that occurs “fewer than 1 time in 100 repetitions” must also occur “fewer than 5 times in 100 repetitions.”

**6.85** Any  $z$  with  $2.576 < |z| < 2.807$ .

**6.87**  $P > 0.25$ .

**6.89**  $0.05 < P < 0.10$ ;  $P = 0.0602$ .

**6.91** To determine the effectiveness of alarm systems, we need to know the percent of all homes with alarm systems and the percent of burglarized homes with alarm systems.

**6.93** The first test was barely significant at  $\alpha = 0.05$ , while the second was significant at any reasonable  $\alpha$ .

**6.95** A significance test answers only question (b).

**6.97 (a)** The differences observed might occur by chance even if SES had no effect. **(b)** This tells us that the statistically insignificant test result did not occur merely because of a small sample size.

**6.99 (a)**  $P = 0.2843$ . **(b)**  $P = 0.1020$ . **(c)**  $P = 0.0023$ .

**6.101** With a larger sample, we might have significant results.

**6.107**  $n$  should be about 100,000.

**6.109** Reject the fifth ( $P = 0.002$ ) and eleventh ( $P < 0.002$ ), because the  $P$ -values are both less than  $0.05/12 = 0.0042$ .

**6.111** Larger samples give more power.

**6.113** Higher; larger differences are easier to detect.

**6.115 (a)** Power decreases. **(b)** Power decreases. **(c)** Power increases.

**6.117** Power: about 0.99.

**6.119** Power: 0.4641.

**6.121 (a)** Hypotheses: “subject should go to college” and “subject should join workforce.” Errors: recommending college for someone who is better suited for the workforce, and recommending work for someone who should go to college.

**6.123 (a)** For example, if  $\mu$  is the mean difference in scores,  $H_0: \mu = 0$ ;  $H_a: \mu \neq 0$ . **(b)** No. **(c)** For example: Was this an experiment? What was the design? How big were the samples?

**6.125 (a)** For boys:

Energy (kJ)	2399.9 to 2496.1
Protein (g)	24.00 to 25.00
Calcium (mg)	315.33 to 332.87

**(b)** For girls:

Energy (kJ)	2130.7 to 2209.3
Protein (g)	21.66 to 22.54
Calcium (mg)	257.70 to 272.30

**(c)** The confidence interval for boys is entirely above the confidence interval for girls for each food intake.

**6.129 (a)** 4.61 to 6.05 mg/dl. **(b)**  $z = 1.45$ ,  $P = 0.0735$ ; not significant.

**6.131 (b)** 26.06 to 34.74  $\mu\text{g/l}$ . **(c)**  $z = 2.44$ ,  $P = 0.0073$ .

**6.133 (a)** Under  $H_0$ ,  $\bar{x}$  has an  $N(0\%, 5.3932\%)$  distribution. **(b)**  $z = 1.28$ ,  $P = 0.1003$ . **(c)** Not significant.

**6.135** It is essentially correct.

**6.137** Find  $\bar{x}$ , then take  $\bar{x} \pm 1.96(4/12) = \bar{x} \pm 2.2632$ .

**6.139** Find  $\bar{x}$ , then compute  $z = (\bar{x} - 23)/(4/12)$ . Reject  $H_0$  based on your chosen significance level.

## CHAPTER 7

7.1 (a) \$13.75. (b) 15.

7.3 \$570.70 to \$629.30.

7.5 (a) Yes. (b) No.

7.7 4.19 to 10.14 hours per month.

7.9 Using  $t^* = 2.776$  from Table D: 0.685 to 22.515. Software gives 0.683 to 22.517.

7.11 The sample size should be sufficient to overcome any non-Normality, but the mean  $\mu$  might not be a useful summary of a bimodal distribution.

7.13 The power is about 0.9192.

7.15 The power is about 0.9452.

7.17 (a)  $t^* = 2.201$ . (b)  $t^* = 2.086$ . (c)  $t^* = 1.725$ . (d)  $t^*$  decreases with increasing sample size and increases with increasing confidence.

7.19  $t^* = 1.753$  (or  $-1.753$ ).

7.21 For the alternative  $\mu < 0$ , the answer would be the same ( $P = 0.034$ ). For the alternative  $\mu > 0$ , the answer would be  $P = 0.966$ .

7.23 (a)  $df = 26$ . (b)  $1.706 < t < 2.056$ . (c)  $0.05 < P < 0.10$ . (d)  $t = 2.01$  is not significant at either level. (e) From software,  $P = 0.0549$ .

7.25 It depends on whether  $\bar{x}$  is on the appropriate side of  $\mu_0$ .

7.27 (a)  $H_0: \mu = 4.7$ ;  $H_a: \mu \neq 4.7$ .  $t = 14.907$  with  $0.002 < P < 0.005$  (software gives  $P = 0.0045$ ). (b) 4.8968% to 5.0566%. (c) Because our confidence interval is entirely within the range of 4.7% to 5.3%, it appears that Budweiser is meeting the required standards.

7.29 (a)  $H_0: \mu = 10$ ;  $H_a: \mu < 10$ . (b)  $t = -5.26$ ,  $df = 33$ ,  $P < 0.0001$ .

7.31 (a) Distribution is not Normal; it has two peaks and one large value. (b) Maybe; we have a large sample but a small population. (c)  $27.29 \pm 5.717$ , or 21.57 to 33.01 cm. (d) One could argue for either answer.

7.33 (a) Yes; the sample size is large. (b)  $t = -2.115$ . Using Table D, we have  $0.02 < P < 0.04$ , while software gives  $P = 0.0381$ .

7.35  $H_0: \mu = 45$  versus  $H_a: \mu > 45$ .  $t = 5.457$ . Using  $df = 49$ ,  $P \approx 0$ ; with  $df = 40$ ,  $P < 0.0005$ .

7.37 (a)  $t = 5.13$ ,  $df = 15$ ,  $P < 0.001$ . (b) With 95% confidence, the mean NEAT increase is between 191.6 and 464.4 calories.

7.39 (a)  $H_0: \mu_c = \mu_d$ ;  $H_a: \mu_c \neq \mu_d$ . (b)  $t = 4.358$ ,  $P = 0.0003$ ; we reject  $H_0$ .

7.41 (a)  $H_0: \mu = 925$ ;  $H_a: \mu > 925$ .  $t = 3.27$  ( $df = 35$ ),  $P = 0.0012$ . (b)  $H_0: \mu = 935$ ;  $H_a: \mu > 935$ .  $t = 0.80$ ,  $P = 0.2146$ . (c) The confidence interval includes 935 but not 925.

7.43 (a) The differences are spread from  $-0.018$  to  $0.020$  g. (b)  $t = -0.347$ ,  $df = 7$ ,  $P = 0.7388$ . (c)  $-0.0117$  to  $0.0087$  g. (d) They may be representative of future subjects, but the results are suspect because this is not a random sample.

- 7.45** (a)  $H_0: \mu = 0$ ;  $H_a: \mu > 0$ . (b) Slightly left-skewed;  $\bar{x} = 2.5$  and  $s = 2.893$ . (c)  $t = 3.865$ ,  $df = 19$ ,  $P = 0.00052$ . (d) 1.15 to 3.85.
- 7.47** For the sign test,  $P = 0.0898$ ; not quite significant, unlike Exercise 7.38.
- 7.49**  $H_0$ : median = 0;  $H_a$ : median  $\neq$  0;  $P = 0.7266$ . This is similar to the  $t$  test  $P$ -value.
- 7.51**  $H_0$ : median = 0;  $H_a$ : median  $>$  0;  $P = 0.0013$ .
- 7.53** Reject  $H_0$  if  $|\bar{x}| \geq 0.00677$ . The power is about 7%.
- 7.55**  $n > 26$ . (The power is about 0.7999 when  $n = 26$ .)
- 7.57**  $-20.3163$  to  $0.3163$ ; do not reject  $H_0$ .
- 7.59** Using  $df = 14$ , Table D gives  $0.04 < P < 0.05$ .
- 7.61** SAS and SPSS give  $t = 2.279$ ,  $P = 0.052$ .
- 7.63** (a) Hypotheses should involve  $\mu_1$  and  $\mu_2$ . (b) The samples are not independent. (c) We need  $P$  to be small (for example, less than 0.10) to reject  $H_0$ . (d)  $t$  should be negative to reject  $H_0$  with this alternative.
- 7.65** (a) No (in fact,  $P = 0.0771$ ). (b) Yes ( $P = 0.0385$ ).
- 7.67**  $H_0: \mu_{\text{Brown}} = \mu_{\text{Blue}}$  and  $H_a: \mu_{\text{Brown}} > \mu_{\text{Blue}}$ .  $t = 2.59$ . Software gives  $P = 0.0058$ . Table D gives  $0.005 < P < 0.01$ .
- 7.69** The nonresponse is  $(3866 - 1839)/3866 = 0.5243$ , or about 52.4%. What can we say about those who do (or do not) respond despite the efforts of the researchers?
- 7.71** (a) While the distributions do not look particularly Normal, they have no extreme outliers or skewness. (b)  $\bar{x}_N = 0.5714$ ,  $S_N = 0.7300$ ,  $n_N = 14$ ;  $\bar{x}_S = 2.1176$ ,  $S_S = 1.2441$ ,  $n_S = 17$ . (c)  $H_0: \mu_N = \mu_S$ ;  $H_a: \mu_N < \mu_S$ . (d)  $t = -4.303$ , so  $P = 0.0001$  ( $df = 26.5$ ) or  $P < 0.0005$  ( $df = 13$ ). (e)  $-2.2842$  to  $-0.8082$  ( $df = 26.5$ ) or  $-2.3225$  to  $-0.7699$  ( $df = 13$ ).
- 7.73** (a) Although the data are integers, the sample sizes are large. (b) Taco Bell:  $\bar{x} = 4.1987$ ,  $s = 0.8761$ ,  $n = 307$ . McDonald's:  $\bar{x} = 3.9365$ ,  $s = 0.8768$ ,  $n = 362$ . (c)  $t = 3.85$ ,  $P = 0.0001$  ( $df = 649.4$ ) or  $P < 0.005$  ( $df = 100$ ). (d) 0.129 to 0.396 ( $df = 649.4$ ) or 0.128 to 0.391 ( $df = 306$ ) or 0.127 to 0.397 ( $df = 100$ ).
- 7.75** (a) Assuming we have SRSs from each population, this seems reasonable. (b)  $H_0: \mu_{\text{Early}} = \mu_{\text{Late}}$  and  $H_a: \mu_{\text{Early}} \neq \mu_{\text{Late}}$ . (c)  $SE_D = 1.0534$ ,  $t = 1.614$ ,  $P = 0.1075$  ( $df = 347.4$ ) or  $P = 0.1081$  ( $df = 199$ ). (d)  $-0.372$  to  $3.772$  ( $df = 347.7$ ) or  $-0.377$  to  $3.777$  ( $df = 199$ ) or  $-0.390$  to  $3.790$  ( $df = 100$ ).
- 7.77** (a) This may be near enough to an SRS if this company's working conditions were similar to those of other workers. (b) 9.99 to 13.01 mg.y/m<sup>3</sup>. (c)  $t = 15.08$ ,  $P < 0.0001$  with either  $df = 137$  or 114. (d) The sample sizes are large enough that skewness should not matter.
- 7.79** You need either sample sizes and standard deviations or degrees of freedom and a more accurate value for the  $P$ -value. The confidence interval will give us useful information about the magnitude of the difference.
- 7.81** This is a matched pairs design.
- 7.83** The next 10 employees who need screens might not be an independent group—perhaps they all come from the same department, for example.
- 7.85** (a) The north distribution (five-number summary 2.2, 10.2, 17.05, 39.1, 58.8 cm) is right-skewed, while the south distribution (2.6, 26.1, 37.70, 44.6, 52.9) is left-skewed. (b) The methods of this section

seem to be appropriate. **(c)**  $H_0: \mu_N = \mu_S$ ;  $H_a: \mu_N \neq \mu_S$ . **(d)**  $t = -2.63$  with  $df = 55.7$  ( $P = 0.011$ ) or  $df = 29$  ( $P = 0.014$ ). **(e)** Either  $-19.09$  to  $-2.57$  or  $-19.26$  to  $-2.40$  cm.

**7.87 (a)** Either  $-0.90$  to  $6.90$  units ( $df = 122.5$ ) or  $-0.95$  to  $6.95$  units ( $df = 54$ ). **(b)** Random fluctuation may account for the difference in the two averages.

**7.89 (a)**  $H_0: \mu_B = \mu_F$ ;  $H_a: \mu_B > \mu_F$ ;  $t = 1.654$ ,  $P = 0.053$  ( $df = 37.6$ ) or  $P = 0.058$  ( $df = 18$ ). **(b)**  $-0.2$  to  $2.0$ . **(c)** We need two independent SRSs from Normal populations.

**7.91**  $s_p = 0.9347$ ;  $t = -3.636$ ,  $df = 40$ ,  $P = 0.0008$ ;  $-1.6337$  to  $-0.4663$ . Both results are similar to those for Exercise 7.72.

**7.93**  $s_p = 15.96$ ;  $t = -2.629$ ,  $df = 58$ ,  $P = 0.0110$ ;  $-19.08$  to  $-2.58$  cm. All results are nearly the same as in Exercise 7.85.

**7.95**  $df = 55.725$ .

**7.97 (a)**  $df = 137.066$ . **(b)**  $s_p = 5.332$  (slightly closer to  $s_2$ , from the larger sample). **(c)** With no assumption,  $SE_1 = 0.7626$ ; with the pooled method,  $SE_2 = 0.6136$ . **(d)**  $t = 18.74$ ,  $df = 333$ ,  $P < 0.0001$ .  $t$  and  $df$  are larger, so the evidence is stronger (although it was quite strong before). **(e)**  $df = 121.503$ ;  $s_p = 1.734$ ;  $SE_1 = 0.2653$  and  $SE_2 = 0.1995$ .  $t = 24.56$ ,  $df = 333$ ,  $P < 0.0001$ .

**7.99 (a)**  $F^* = 2.25$ . **(b)** Significant at the 10% level but not at the 5% level.

**7.101** A smaller  $\sigma$  would yield more power.

**7.103**  $F = 1.106$ ,  $df = 199$  and  $201$ . Using Table D ( $df = 120$  and  $200$ ),  $P > 0.200$ . (Software gives  $P = 0.4762$ .)

**7.105**  $F = 5.263$  with  $df = 114$  and  $219$ ;  $P < 0.0001$ . The authors described the distributions as somewhat skewed, so the Normality assumption may be violated.

**7.107**  $F = 1.506$  with  $df = 29$  and  $29$ ;  $P = 0.2757$ . The stemplots in Exercise 7.85 did not appear to be Normal.

**7.109 (a)**  $F = 1.12$ ; do not reject  $H_0$ . **(b)** The critical values are  $9.60$ ,  $15.44$ ,  $39.00$ , and  $647.79$ . With small samples, these are low-power tests.

**7.111** Using a larger  $\sigma$  for planning the study is advisable because it provides a conservative (safe) estimate of the power.

**7.113**  $\bar{x} = 139.5$ ,  $s = 15.0222$ ,  $S_{\bar{x}} = 7.5111$ . We cannot consider these four scores to be an SRS.

**7.115** As  $df$  increases,  $t^*$  approaches  $1.96$ .

**7.117** Margins of error decrease with increasing sample size.

**7.119 (a)** Two independent samples. **(b)** Matched pairs. **(c)** Single sample.

**7.121 (a)**  $H_0: \mu = 1.5$ ;  $H_a: \mu < 1.5$ .  $t = -9.974$ ,  $P \approx 0$ . **(b)**  $0.697$  to  $0.962$  violations. **(d)** The sample size should be large enough to make  $t$  procedures safe.

**7.123 (a)**  $-3.008$  to  $1.302$  (Software gives  $-2.859$  to  $1.153$ ). **(b)**  $-1.761$  to  $0.055$ .

**7.125 (a)** We are looking at the average proportion for samples of  $n = 41$  and  $197$ . **(b)**  $H_0: \mu_B = \mu_W$  and  $H_a: \mu_B \neq \mu_W$ . **(c)** For First Year:  $t = 0.982$ . With  $df = 52.3$ ,  $P = 0.3305$ . For Third Year:  $t = 2.126$ ,  $df = 46.9$ ,  $P = 0.0388$ .

**7.127 (a)** Body weight: mean  $-0.7$  kg, SE  $2.298$  kg. Caloric intake: mean  $= 14$  cal, SE  $= 56.125$  cal. **(b)**  $t_1 = -0.305$  (body weight) and  $t_2 = 0.249$  (caloric intake), both  $df = 13$ , both  $P$ -values are about  $0.8$ . **(c)**  $-5.66$  to  $4.26$  kg and  $-107.23$  to  $135.23$  cal.

**7.129 (a)** At each nest, the same mockingbird responded on each day. **(b)**  $6.9774$  m. **(c)**  $t = 6.32$ ,  $P < 0.0001$ . **(d)**  $5.5968$  m. **(e)**  $t = -0.973$ ,  $P = 0.3407$ .

**7.131** How much a person eats may depend on how many people he or she is sitting with.

**7.133** No; what we have is nothing like an SRS.

**7.135**  $77.76\% \pm 13.49\%$ , or  $64.29\%$  to  $91.25\%$ .

**7.137** GPA:  $t = -0.91$ ,  $df = 74.9$  ( $P = 0.1839$ ) or  $30$  ( $0.15 < P < 0.20$ ). Confidence interval:  $-1.33$  to  $0.5$ . IQ:  $t = 1.64$ ,  $df = 56.9$  ( $P = 0.0530$ ) or  $30$  ( $0.05 < P < 0.10$ ). Confidence interval:  $-1.12$  to  $11.36$ .

**7.139**  $t = 3.65$ ,  $df = 237.0$  or  $115$ ,  $P < 0.0005$ . 95% confidence interval for the difference:  $0.78$  to  $2.60$ .

**7.141**  $t = -0.3533$ ,  $df = 179$ ,  $P = 0.3621$ .

**7.143** Basal:  $\bar{x} = 41.0455$ ,  $s = 5.6356$ . DRTA:  $\bar{x} = 46.7273$ ,  $s = 7.3884$ . Strat:  $\bar{x} = 44.2727$ ,  $s = 5.7668$ . **(a)**  $t = 2.87$ ,  $P < 0.005$ . Confidence interval for difference:  $1.7$  to  $9.7$  points. **(b)**  $t = 1.88$ ,  $P < 0.05$ . Confidence interval for difference:  $-0.24$  to  $6.7$  points.

## CHAPTER 8

**8.1 (a)**  $n = 5013$  smartphone users. **(b)**  $p$  is the proportion of smartphone users who have used the phone to search for information about a product that they purchased. **(c)**  $X = 2657$ . **(d)**  $\hat{p} = 0.530$ .

**8.3 (a)**  $0.0070$ . **(b)**  $0.530 \pm 0.014$ . **(c)**  $51.6\%$  to  $54.4\%$ .

**8.5** Shade above  $1.34$  and below  $-1.34$ .

**8.7**  $\hat{p} = 0.75$ ,  $z = 2.24$ ,  $P = 0.0250$ .

**8.9 (a)**  $z = -1.34$ ,  $P = 0.1802$  (Software gives  $P = 0.1797$ ). **(b)**  $0.1410$  to  $0.5590$ —the complement of the interval shown in Figure 8.3.

**8.11** The plot is symmetric about  $0.5$ , where it has its maximum.

**8.13 (a)**  $p$  is the proportion of students at your college who regularly eat breakfast.  $n = 200$ ,  $X = 84$ . **(b)**  $\hat{p} = 0.42$ . **(c)** We estimate that the proportion of all students at the university who eat breakfast is about  $0.42$  ( $42\%$ ).

**8.15 (a)**  $\hat{p} = 0.461$ ,  $SE_{\hat{p}} = 0.0157$ ,  $m = 0.0308$ . **(b)** Yes. **(c)**  $0.4302$  to  $0.4918$ . **(d)** We are  $95\%$  confident that between  $43\%$  and  $49.2\%$  of cell phone owners used their cell phone while in a store to call a friend or family member for advice about a purchase.

**8.17 (a)**  $\hat{p} = 0.7826$ ,  $SE_{\hat{p}} = 0.0272$ ,  $m = 0.0533$ . **(b)** This was not an SRS; they asked *all* customers in the two-week period. **(c)**  $0.7293$  to  $0.8359$ .

**8.19**  $n$  at least  $597$ .

**8.21 (a)** The confidence level cannot exceed  $100\%$ . (In practical terms, the confidence level must be *less than*  $100\%$ .) **(b)** The margin of error only accounts for random sampling error. **(c)**  $P$ -values measure the strength of the evidence against  $H_0$ , not the probability of it being true.

**8.23**  $\hat{p} = 0.6548$ ;  $0.6416$  to  $0.6680$ .

- 8.25 (a)**  $X = 934.5$ , which rounds to 935. We cannot have fractions of respondents. **(b)** Using 89%, 0.8711 to 0.9089. **(c)** 87.1% to 90.9%.
- 8.27 (a)** Values of  $\hat{p}$  outside the interval 0.1730 to 0.4720. **(b)** Values outside the interval 0.210 to 0.390.
- 8.29 (a)** About 67,179 students. **(b)** 0.4168 to 0.4232.
- 8.31** 0.4043 to 0.4557.
- 8.33 (a)**  $\pm 0.001321$ . **(b)** Other sources of error are much more significant than sampling error.
- 8.35 (a)**  $\hat{p} = 0.3275$ ; 0.3008 to 0.3541. **(b)** Speakers and listeners probably perceive sermon length differently.
- 8.37 (a)**  $H_0: p = 0.5$  versus  $H_a: p > 0.5$ ;  $\hat{p} = 0.7$ .  $z = 2.83$ ,  $P = 0.0023$ . **(c)** The test is significant at the 5% level (and the 1% level as well).
- 8.39 (a)**  $z = 1.34$ ,  $P = 0.1802$ . **(b)** 0.4969 to 0.5165.
- 8.41**  $n = 9604$ .
- 8.43** The sample sizes are 55, 97, 127, 145, 151, 145, 127, 97, and 55; take  $n = 151$ .
- 8.45** Mean =  $-0.3$ , standard deviation = 0.1360.
- 8.47 (a)** Means  $p_1$  and  $p_2$ , standard deviations  $p_1(1-p_1)/n_1$  and  $p_2(1-p_2)/n_2$ . **(b)**  $p_1 - p_2$ . **(c)**  $p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$ .
- 8.49** The interval for  $q_W - q_M$  is  $-0.0030$  to 0.2516.
- 8.51** The sample proportions support the alternative hypothesis  $p_m > p_w$ ;  $P = 0.0287$ .
- 8.53 (a)** Only 5 of 25 watched the second design for more than a minute; this does not fit the guidelines. **(b)** It is reasonable to assume that the sampled students were chosen randomly. No information was given about the size of the institution; are there more than  $20(361) = 7220$  first-year students and more than  $20(221) = 4420$  fourth-year students? There were more than 15 each “Yes” and “No” answers in each group.
- 8.55 (a)** Yes. **(b)** Yes.
- 8.57 (a)** RR (watch more than one minute) = 2.4. **(b)** RR (“Yes” answer) = 2.248.
- 8.59 (a)** Type of college is explanatory; response is requiring physical education. **(b)** *The populations are private and public colleges and universities.* **(c)**  $X_1 = 101$ ,  $n_1 = 129$ ,  $\hat{p}_1 = 0.7829$ ,  $X_2 = 60$ ,  $n_2 = 225$ ,  $p^2 = 0.2667$ . **(d)** 0.4245 to 0.6079. **(e)**  $H_0: p_1 = p_2$  and  $H_a: p_1 \neq p_2$ . We have  $\hat{p} = 60 + 101225 + 129 = 0.4548$ .  $z = 9.39$ ,  $P \approx 0$ . **(f)** All counts are greater than 15. Were these random samples?
- 8.61** 0.0363 to 0.1457.
- 8.63 (a)**  $n_1 = 1063$ ,  $\hat{p}_1 = 0.54$ ,  $n_2 = 1064$ ,  $\hat{p}_2 = 0.89$ . (We can estimate  $X_1 = 574$  and  $X_2 = 947$ .) **(b)** 0.35. **(c)** Yes; large, independent samples from two populations. **(d)** 0.3146 to 0.3854. **(e)** 35%; 31.5% to 38.5%. **(f)** A possible concern: adults were surveyed before Christmas.
- 8.65 (a)**  $n_1 = 1063$ ,  $\hat{p}_1 = 0.73$ ,  $n_2 = 1064$ ,  $\hat{p}_2 = 0.76$ . (We can estimate  $X_1 = 776$  and  $X_2 = 809$ .) **(b)** 0.03. **(c)** Yes; large, independent samples from two populations. **(d)**  $-0.0070$  to 0.0670. **(e)** 3%;  $-0.7\%$  to 6.7%. **(f)** A possible concern: adults were surveyed before Christmas.
- 8.67** No; we need independent samples from different populations.

**8.69 (a)**  $H_0$  should refer to  $p_1$  and  $p_2$ . **(b)** Only if  $n_1 = n_2$ . **(c)** Confidence intervals account for only sampling error.

**8.71 (a)**  $\hat{p}_F=0.8$ ,  $SE = 0.05164$ ;  $\hat{p}_M=0.3939$ ,  $SE = 0.04253$ . **(b)** 0.2960 to 0.5161. **(c)**  $z = 5.22$ ,  $P \approx 0$ .

**8.73 (a)**  $n = 2342$ ,  $x = 1639$ . **(b)**  $\hat{p}=0.6998$ .  $SE = 0.0095$ . **(c)** 0.6812 to 0.7184. **(d)** Yes.

**8.75** We have large samples from two independent populations (different age groups).  $\hat{p}_1=0.8161$ ,  $\hat{p}_2=0.4281$ .  $SE_D = 0.0198$ . The 95% confidence interval is 0.3492 to 0.4268.

**8.79 (a)** 1207. **(b)** 0.6483 to 0.6917. **(c)** About 64.8% to 69.2%.

**8.81** There was only one sample, not two independent samples. Many people use both.

**8.83 (a)** We have six chances to make an error. **(b)** Use  $z^* = 2.65$  (software: 2.6383). **(c)** 0.705 to 0.775, 0.684 to 0.756, 0.643 to 0.717, 0.632 to 0.708, 0.622 to 0.698, and 0.571 to 0.649.

**8.85**  $\hat{p}=0.375$ ,  $SE_D = 0.01811$ ,  $z = 6.08$ ,  $P < 0.0001$ .

**8.87** 0.6337 to 0.6813.

**8.89**  $H_0: p_F = p_M$  and  $H_a: p_F \neq p_M$ .  $X_M = 171$  and  $X_F = 150$ .  $\hat{p}=0.1600$ ,  $SE_{D_p} = 0.0164$ .  $z = 1.28$ ,  $P = 0.2009$ .

**8.93** All  $\hat{p}$ -values are greater than 0.5. Texts 3, 7, and 8 have (respectively)  $z = 0.82$ ,  $P = 0.4122$ ;  $z = 3.02$ ,  $P = 0.0025$ ; and  $z = 2.10$ ,  $P = 0.0357$ . For the other texts,  $z \geq 4.64$  and  $P < 0.00005$ .

**8.95** The difference becomes more significant as sample size increases. With  $n = 60$ ,  $P = 0.2713$ ; with  $n = 500$ ,  $P = 0.0016$ , for example.

**8.97 (a)**  $n = 534$ . **(b)**  $n = (z^*/m)^2/2$ .

**8.99 (a)**  $p_0 = 0.7911$ . **(b)**  $\hat{p}=0.3897$ ,  $z = -29.1$ ;  $P$  is tiny. **(c)**  $\hat{p}_1=0.3897$ ,  $\hat{p}_2=0.7930$ ,  $z = -29.2$ ;  $P$  is tiny.

**8.101 (a)** 0.5278 to 0.5822. **(b)** 0.5167 to 0.5713. **(c)** 0.3170 to 0.3690. **(d)** 0.5620 to 0.6160. **(e)** 0.5620 to 0.6160. **(f)** 0.6903 to 0.7397.

## CHAPTER 9

**9.1 (a)** Yes:  $47/292 = 0.161$ , No:  $245/292 = 0.839$ . **(b)** Yes:  $21/233 = 0.090$ , No:  $212/233 = 0.910$ . **(d)** Females are somewhat more likely than males to have increased the time they spend on Facebook.

**9.5** Among all three fruit consumption groups, vigorous exercise is most likely. Incidence of low exercise decreases with increasing fruit consumption.

**9.7**

Fruit	Physical Activity			Total
	Low	Medium	Vigorous	
Low	51.9	212.9	304.2	569
Medium	29.3	120.1	171.6	321
High	26.8	110.0	157.2	294
Total	108	443	633	1184

**9.9 (a)**  $df = 12$ ,  $0.05 < P < 0.10$ . **(b)**  $df = 12$ ,  $0.05 < P < 0.10$ . **(c)**  $df = 1$ ,  $0.005 < P < 0.01$ . **(d)**  $df = 1$ , 0.20

$< P < 0.25$ .

**9.11 (a)**

Explanatory variable		
Response	1	2
Yes	0.357	0.452
No	0.643	0.548
Total	1.000	1.000

(c) Explanatory variable value 1 had proportionately fewer “yes” responses.

**9.13 (a)**  $p_i$  = proportion of “Yes” responses in group  $i$ .  $H_0: p_1 = p_2$ ,  $H_a: p_1 \neq p_2$ .  $\hat{p} =$

$(75+95)/(210+210)=0.4048$ .  $z = -1.9882$ ,  $P = 0.0468$ . We fail to reject  $H_0$ . (c) The  $P$ -values agree. (d)  $z^2 = (-1.9882)^2 = 3.9529$ .

**9.15** Roundoff error.

**9.17** The contributions for the other five states are

CA	HI	IN	NV	OH
0.5820	0.0000	0.0196	0.0660	0.2264

$\chi^2 = 0.9309$ .

**9.19 (a)**  $H_0: P(\text{head}) = P(\text{tail}) = 0.5$  versus  $H_a: H_0$  is incorrect (the probabilities are not 0.5).

(b)  $\chi^2 = 1.7956$ ,  $df = 1$ ,  $P = 0.1802$ .

**9.21 (a)** Joint Distribution:

	Site 1	Site 2	Total
More than 1 min	0.24	0.10	0.34
Less than 1 min	0.26	0.40	0.66
Total	0.50	0.50	1.00

The conditional distributions are

	Site 1	Site 2	Total
More than 1 min	0.7059	0.2941	1.0000
Less than 1 min	0.3939	0.6061	1.0000

and

	Site 1	Site 2
More than 1 min	0.48	0.20
Less than 1 min	0.52	0.80
Total	1.00	1.00

(b) Joint Distribution

	1st year	4th year	Total
Yes	0.1460	0.2010	0.3471

No	0.4742	0.1787	0.6529
Total	0.6203	0.3797	1.0000

The conditional distributions are

	1st year	4th year	Total
Yes	0.4208	0.5792	1.0000
No	0.7263	0.2737	1.0000

and

	1st year	4th year
Yes	0.2355	0.5294
No	0.7645	0.4706
Total	1.0000	1.0000

**9.23** (a) Describe a relationship. (b) Describe a relationship. (c) Time of day might explain the violence content of TV programs. (d) Age would explain bad teeth.

**9.25**

Gender	Times Witnessed			Total
	Never	Once	More than once	
Girls	125.503	161.725	715.773	1003
Boys	120.497	155.275	687.227	963
Total	246	317	1403	1966

**9.27** (a)  $H_0: p_1 = p_2$  versus  $H_a: p_1 \neq p_2$ , where the proportions of interest are those for persons harassed in person.  $\hat{p}_1 = 3231/361 = 0.8892$ ,  $\hat{p}_2 = 22/641 = 0.3120$ ,  $\hat{p} = 521/1002 = 0.5200$ .  $z = 17.556$ ,  $P \approx 0$ . (b)  $H_0$ : there is no association between being harassed online and in person versus  $H_a$ : There is a relationship.  $X^2 = 308.23$ ,  $df = 1$ ,  $P \approx 0$ . (c)  $17.556^2 = 308.21$ , which agrees with  $X^2$  to within roundoff error. (d) One possibility is eliminating girls who said they have not been harassed.

**9.29** (a) The solution to Exercise 9.27 used “harassed online” as the explanatory variable. (b) Changing to use “harassed in person” for the two-proportions  $z$  test gives  $\hat{p}_1 = 0.6161$ ,  $\hat{p}_2 = 0.0832$ ,  $\hat{p} = 0.3603$ . We again compute  $z = 17.556$ ,  $P \approx 0$ . No changes will occur in the chi-square test. (c) If two variables are related, the test statistic will be the same regardless of which is viewed as explanatory.

**9.31**  $E_i = 100$  for each face of the die.

**9.33** (a) One might believe that opinion depended on the type of institution. (b) Presidents at 4-year public institutions are roughly equally divided about online courses, with presidents at 2-year public institutions slightly in favor. 4-year private school presidents are definitely not in agreement, while those at private 2-year schools seem to think online courses are equivalent to face-to-face courses.

**9.35** (a) 206. (b) We have separate samples, so the two-way table is

	Presidents	Public
Yes	206	621
No	189	1521

(c) The column totals for this table are the two sample sizes. The row totals might be seen as giving an overall opinion on the value of online courses. (d)  $H_0$ : The opinions on the value of online courses are the

same for college presidents and the general public versus  $H_a$ : The opinions are different.  $X^2 = 81.41$ ,  $df = 1$ ,  $P \approx 0$ .

**9.37** (a) For example, in the “small” stratum, 51 claims were allowed, 6 were not allowed, and the total number of claims was 57. Altogether, there were 79 claims; 67 were allowed and 12 were not. (b) 10.5% (small claims), 29.4% (medium), and 20.0% (large) were not allowed. (c) In the  $3 \times 2$  table, the expected count for large/not allowed is too small. (d) There is no relationship between claim size and whether a claim is allowed. (e)  $X^2 = 3.456$ ,  $df = 1$ ,  $P = 0.063$ .

**9.39** There is strong evidence of a change ( $X^2 = 308.3$ ,  $df = 2$ ,  $P < 0.0001$ ).

**9.41** (a) For example, among those students in trades, 320 enrolled right after high school, and 622 later. (b) In addition to the given percents, 39.4% of these students enrolled right after high school. (c)  $X^2 = 276.1$ ,  $df = 5$ ,  $P < 0.0001$ .

**9.43** (a) For example, among those students in trades, 188 relied on parents, family, or spouse, and 754 did not. (b)  $X^2 = 544.8$ ,  $df = 5$ ,  $P < 0.0001$ . (c) In addition to the given percents, 25.4% of all students relied on family support.

**9.45** (a) 57.98%. (b) 30.25%. (c) To test “There is no relationship between waking and bedtime symptoms” versus “There is a relationship,” we find  $X^2 = 2.275$ ,  $df = 1$ ,  $P = 0.132$ .

**9.47** Start by setting  $a$  equal to any number from 0 to 100.

**9.49**  $X^2 = 852.433$ ,  $df = 1$ ,  $P < 0.0005$ .

**9.55** (a) We expect each quadrant to contain one-fourth of the 100 trees. (b) *Some* random variation would not surprise us. (c)  $X^2 = 10.8$ ,  $df = 3$ ,  $P = 0.0129$ .

## CHAPTER 10

**10.1** (a) 3.1. (b) The slope of 3.1 means the *average* value of  $y$  increases by 3.1 units for each unit increase in  $x$ . (c) 82.6. (d) 72.2 to 93.0.

**10.3** (a)  $t = 1.895$ ,  $df = 25 - 2 = 23$ . From Table D, we have  $0.05 < P < 0.10$  (software gives 0.0707). (b)  $t = 2.105$ ,  $df = 25 - 2 = 23$ . From Table D,  $0.04 < P < 0.05$  (0.0464 from software). (c)  $t = 3.091$ ,  $df = 98$ . Using  $df = 80$  in Table D,  $0.002 < P < 0.005$  (0.0026 from software).

**10.5**  $m = 0.7 \text{ kg/m}^2$ . At  $x = 5.0$ , the margin of error will be larger.

**10.7** (b) The fitted line is  $\text{Spending} = -4900.5333 + 2.4667x$ . (Note: Rounding on this exercise can make a big difference in results.) (c) The residuals are (with enough decimal places in slope and intercept)  $-0.1$ ,  $0.2$ ,  $-0.1$ .  $s = 0.2449$ . (d) The model is  $y = \beta_0 + \beta_1x$ . We have estimates  $\hat{\beta}_0 = -4900.5333$  and  $\hat{\beta}_1 = 2.4677$  and  $\sigma(\epsilon_i) = 0.2449$  (e)  $s(b_1) = 0.0577$ .  $df = 1$ , so  $t^* = 12.71$ . The 95% CI is 1.733 to 3.200.

**10.9** (a)  $\beta_0$ ,  $\beta_1$ , and  $\sigma$  are the parameters. (b)  $H_0$  should refer to  $\beta_1$ . (c) The confidence interval will be narrower than the prediction interval.

**10.11** Kiplinger narrows down the number of colleges; these are an SRS from that list, not from the original 500 four-year public colleges.

**10.13** (a) \$19,591.29. (b) \$23,477.93. (c) La Crosse is farther from the center of the  $x$  distribution.

**10.15** Prediction intervals concern individuals instead of means. Departures from the Normal distribution assumption would be more severe here (in terms of how the individuals vary around the regression line).

**10.17** (a)  $H_0: \beta_1 = 0$  and  $H_a: \beta_1 > 0$ . It does not seem reasonable to believe that tuition will decrease. (b) From software,  $t = 13.94$ ,  $P < 0.0005$  ( $df = 26$ ). (c) Using  $df = 26$  from Table D,  $0.9675 \pm 2.056$  ( $0.06939$ ) =  $0.825$  to  $1.110$ . (d)  $r^2 = 88.2\%$ . (e) Inference on  $\beta_0$  would be extrapolation; there were no colleges close to \$0 tuition in 2008.

**10.19** (a) The relationship is strong (little scatter), increasing, and fairly linear; however, there may be a bit of curve at each end. (b)  $\text{OUT11} = 1075 + 1.15 \text{ OUT08}$  (or  $\hat{y} = 1075 + 1.15x$ ). (d) No overt problems are noted, even though the Normal plot wiggles around the line.

**10.21** The scatterplot shows a weak, increasing relationship between in-state and out-of-state tuition rates for 2011. Minnesota appears to be an outlier, with an in-state tuition of \$13,022 and an out-of-state tuition of \$18,022. The regression equation is  $\text{OUT11} = 17,160 + 1.017 \text{ IN11}$  (or  $\hat{y} = 17,160 + 1.017x$ ). The scatterplot of residuals against  $x$  shows no overt problems (except the low outlier for Minnesota); the Normal quantile plot also shows no problems, although we note that several schools seem to have similar residuals (slightly more than \$5000).

**10.23** (a)  $\hat{y} = -0.0127 + 0.0180x$ ,  $r^2 = 80.0\%$ . (b)  $H_0: \beta_1 = 0$ ;  $H_a: \beta_1 > 0$ ;  $t = 7.48$ ,  $P < 0.0001$ . (c) The predicted mean is  $0.07712$ ; the interval is  $0.040$  to  $0.114$ .

**10.25** (a) Both distributions are sharply right-skewed; the five-number summaries are  $0\%$ ,  $0.31\%$ ,  $1.43\%$ ,  $17.65\%$ ,  $85.01\%$  and  $0$ ,  $2.25$ ,  $6.31$ ,  $12.69$ ,  $27.88$ . (b) No;  $x$  and  $y$  do not need to be Normal. (c) There is a weak positive linear relationship. (d)  $\hat{y} = 6.247 + 0.1063x$ . (e) The residuals are right-skewed.

**10.27** (a) 17 of these 30 homes sold for more than their assessed values. (b) A moderately strong, increasing linear relationship. (c)  $\hat{y} = 66.95 + 0.6819x$ . (d) The outlier point is still an outlier in this plot; it is almost three standard deviations below its predicted value. (e) The new equation is  $\hat{y} = 37.41 + 0.8489x$ .  $s = 31.41$  decreased to  $s = 26.80$ . (f) There are no clear violations of the assumptions.

**10.29** (a) The plot could be described as increasing and roughly linear, or possibly curved; it almost looks as if there are two lines; one for years before 1980 and one after that. 2012 had an unusually low number of tornadoes, while 2004 had an unusually high number. (b)  $\text{Tornadoes} = -27,432 + 14.312 \text{ Year}$  (or  $\hat{y} = -27,432 + 14.312x$ ). The 95% confidence interval is  $14.312 \pm 2.009(1.391)$  using  $df = 50$ . (c) We see what seems to be an increasing amount of scatter in later years. (d) Based on the Normal quantile plot, we can assume that the residuals are Normally distributed. (e) After eliminating 2004 and 2012 from the data set, the new equation is  $\text{Tornadoes} = -27,458 + 14.324 \text{ Year}$ . These years are not very influential to the regression (the slope and intercept changed very little).

**10.31** (a)  $8.41\%$ . (b)  $t = 9.12$ ,  $P < 0.0001$ . (c) The students who did not answer might have different characteristics.

**10.33** (a)  $x$  (percent forested) is right-skewed;  $\bar{x} = 39.3878\%$ ,  $s_x = 32.2043\%$ .  $y$  (IBI) is left-skewed;  $\bar{y} = 65.9388$ ,  $s_y = 18.2796$ . (b) A weak positive association, with more scatter in  $y$  for small  $x$ . (c)  $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ ,  $i = 1, 2, \dots, 49$ ;  $\varepsilon_i$  are independent  $N(0, \sigma)$  variables. (d)  $H_0: \beta_1 = 0$ ;  $H_a: \beta_1 \neq 0$ ; (e)  $\text{IBI} = 59.9 + 0.153 \text{ Area}$ ;  $s = 17.79$ . For testing the hypotheses in (d),  $t = 1.92$  and  $P = 0.061$ . (f) Residual plot shows a slight curve. (g) Residuals are left-skewed.

**10.35** The first change decreases  $P$  (that is, the relationship is more significant) because it accentuates the positive association. The second change weakens the association, so  $P$  increases (the relationship is less significant).

**10.37**  $\text{Area} = 10$ ,  $\hat{y} = 57.52$ ; using forest = 63,  $\hat{y} = 69.55$ . Both predictions have a lot of uncertainty (the prediction intervals are about 70 units wide).

**10.39** (a) It appears to be quite linear. (b)  $\text{Lean} = -61.12 + 9.3187 \text{ Year}$ ;  $r^2 = 98.8\%$ . (c) 8.36 to 10.28 tenths of a millimeter/year.

**10.41** (a) 113. (b) The prediction is 991.89 mm beyond 2.9 m, or about 3.892 m. (c) Prediction interval.

**10.43**  $t = -4.16$ ,  $df = 116$ ,  $P < 0.0001$ .

**10.45**  $DFM = 1$ ,  $DFE = 18$ ,  $SSE = 3304.3$ .  $MSM = 4947.2$ ,  $MSE = 183.572$ ,  $F = 26.95$ .

**10.47** The standard error is 0.1628; the confidence interval is 0.503 to 1.187.

**10.49** For  $n = 15$ ,  $t = 2.08$  and  $P = 0.0579$ . For  $n = 25$ ,  $t = 2.77$  and  $P = 0.0109$ . Finding the same correlation with more data points is stronger evidence that the observed correlation is not just due to chance.

**10.51** (a) Strong positive linear association with one outlier (SAT 420, ACT 21). (b)  $\hat{ACT} = 1.63 + 0.0214SAT$ ,  $t = 10.78$ ,  $P < 0.0005$ . (c)  $r = 0.8167$ .

**10.53** (a)  $a_1 = 0.02617$ ,  $a_0 = -2.7522$ . (c) Mean = 21.1333 and standard deviation = 4.7137—the same as for the ACT scores.

**10.55** (a) For squared length:  $\hat{Weight} = -117.99 + 0.4970SqLen$ ,  $s = 52.76$ ,  $r^2 = 0.977$ . (b) For squared width:  $\hat{Weight} = -98.99 + 18.732SqWid$ ,  $s = 65.24$ ,  $r^2 = 0.965$ . Both scatterplots look more linear.

**10.57** IBI and area:  $r = 0.4459$ ,  $t = 3.42$ ,  $P = 0.001$  (from Exercise 10.32). IBI and percent forested:  $r = 0.2698$ ,  $t = 1.92$ ,  $P = 0.061$  (Exercise 10.33). Area and percent forested:  $r = -0.2571$ ,  $t = -1.82$ ,  $P = 0.075$ .

**10.59** The three smallest correlations (0.16 and 0.19) are the only ones that are not significant ( $P = 0.1193$  and 0.0632). The first correlation (0.28) has the smallest  $P$ -value (0.0009). The next four, and the largest correlation in the Caucasian group, have  $P < 0.001$ . The remainder have  $P < 0.01$ .

**10.61** (a) 95% confidence interval for women: 14.73 to 33.33. For men:  $-9.47$  to 42.97. These intervals overlap quite a bit. (b) For women: 22.78. For men: 16.38. The women's slope standard error is smaller in part because it is divided by a large number. (c) Choose men with a wider variety of lean body masses.

## CHAPTER 11

**11.1** (a) Second semester GPA. (b)  $n = 242$ . (c)  $p = 7$ . (d) Gender, standardized test score, perfectionism, self-esteem, fatigue, optimism, and depressive symptomatology.

**11.3** (a) Math GPA should increase when any explanatory variable increases. (b)  $DFM = 4$ ,  $DFE = 77$ . (c) All four coefficients are significantly different from 0 (although the intercept is not).

**11.5** The correlations are found in Figure 11.4. The scatterplots for the pairs with the largest correlations are easy to pick out. The whole-number scale for high school grades causes point clusters in those scatterplots.

**11.7** Using Table D: (a)  $-0.0139$  to 12.8139. (b) 0.5739 to 12.2261.

(c) 0.2372 to 9.3628. (d) 0.6336 to 8.9664. Software gives 0.6422 to 8.9578.

**11.9** (a)  $H_0$  should refer to  $\beta_2$ . (b) Squared multiple correlation. (c) Small  $P$  implies that at least one coefficient is different from 0.

**11.11** (a)  $y_i = \beta_0 + \beta_1x_{i1} + \beta_2x_{i2} + \dots + \beta_7x_{i7} + \varepsilon_i$ , where  $i = 1, 2, \dots, 142$ , and  $\varepsilon_i$  are independent  $N(0, \sigma)$  random variables. (b) The sources of variation are model ( $DFM = p = 7$ ), error ( $DFE = n - p - 1 = 134$ ), and total ( $DFT = n - 1 = 141$ ).

**11.13** (a) The fitted model is  $\hat{GPA} = -0.847 + 0.00269SATM + 0.229HSS$ . (b)  $\hat{GPA} = -0.887 + 0.00237SATM + 0.0850HSM + 0.173HSS$ . (c)  $\hat{GPA} = -1.11 + 0.00240SATM + 0.0827HSM + 0.133HSS + 0.0644HSE$ . (d)  $\hat{GPA} = 0.257 + 0.125HSM + 0.172HSS$ .

	MSE	$R^2$	$P(x_1)$	$P(x_2)$	$P(x_3)$	$P(x_4)$
(a)	0.506	25.4%	0.001	0.000		
(b)	0.501	26.6%	0.004	0.126	0.002	
(c)	0.501	27.1%	0.004	0.137	0.053	0.315
(d)	0.527	22.4%	0.024	0.003		

The “best” model is the model with SATM and HSS.

**11.15** The first variable to leave is InAfterAid ( $P$ -value = 0.465). Fitting the new model gives OutAfterAid ( $P$ -value = 0.182) as the next to leave. AvgAid ( $P$ -value = 0.184) leaves next. At that point, all variables are significant predictors. The model is  $\widehat{\text{AvgDebt}} = -9521 + 118\text{Admit} + 102\text{Yr4Grad} + 661\text{StudPerFac} + 130\text{PercBorrow}$ .

**11.17** (a) 8 and 786. (b) 7.84%; this model is not very predictive. (c) Males and Hispanics consume energy drinks more frequently. Consumption increases with risk-taking scores. (d) Within a group of students with identical (or similar) values of those other variables, energy-drink consumption increases with increasing jock identity and increasing risk taking.

**11.19** (a) Model 1: DFE = 200. Model 2: DFE = 199. (b)  $t = 3.09$ ,  $P = 0.0023$ . (c) For gene expression:  $t = 2.44$ ,  $P = 0.0153$ . For RB:  $t = 3.33$ ,  $P = 0.0010$ . (d) The relationship is still positive. When gene expression increases by 1, popularity increases by 0.204 in Model 1 and by 0.161 in Model 2 (with RB fixed).

**11.21** (a)  $\widehat{\text{BMI}} = 23.4 - 0.682(\text{PA} - 8.614) + 0.102(\text{PA} - 8.614)^2$ , (b)  $R^2 = 17.7\%$ . (c) The residuals look roughly Normal and show no obvious remaining patterns. (d)  $t = 1.83$ ,  $df = 97$ ,  $P = 0.070$ .

**11.23** (a) Budget and Opening are right-skewed; Theaters and Opinion are roughly symmetric (slightly left-skewed). Five-number summaries for Budget and Opening are appropriate; mean and standard deviation could be used for the other two variables. (b) All relationships are positive. The Budget/Theaters and Opening/Theaters relationships appear to be curved; the others are reasonably linear. The correlations between Budget, Opening, and Theaters are all greater than 0.7. Opinion is less correlated with the other three variables—about 0.4 with Budget and Opening and only 0.156 with Theaters.

**11.25** (a)  $\text{USRevenue}_i = \beta_0 + \beta_1 \text{Budget}_i + \beta_2 \text{Opening}_i + \beta_3 \text{Theaters}_i + \beta_4 \text{Opinion}_i + \varepsilon_i$ , where  $i = 1, 2, \dots, 35$ ;  $\varepsilon_i$  are independent  $N(0, \sigma)$  random variables. (b)  $\widehat{\text{USRevenue}} = -67.72 + 0.1351 \text{Budget} + 3.0165 \text{Opening} - 0.00223 \text{Theaters} + 10.262 \text{Opinion}$ . (c) *The Dark Knight* may be influential. The spread of the residuals appears to increase with Theaters. (d) 98.1%.

**11.27** (a) \$86.87 to \$154.91 million. (b) \$89.94 to \$154.99 million. (c) The intervals are very similar.

**11.29** (a) PEER is left-skewed; the other two variables are irregular. (b) PEER and FtoS are negatively correlated ( $r = -0.114$ ); FtoS and CtoF are positively correlated ( $r = 0.580$ ); the other correlation is very small.

**11.31** (a)  $\text{OVERALL}_i = \beta_0 + \beta_1 \text{PEER}_i + \beta_2 \text{FtoS}_i + \beta_3 \text{CtoF}_i + \varepsilon_i$ , where  $\varepsilon_i$  are independent  $N(0, \sigma)$  random variables. (b)  $\widehat{\text{OVERALL}} = 18.85 + 0.5746 \text{PEER} + 0.0013 \text{FtoS} + 0.1369 \text{CtoF}$ . (c) PEER: 0.4848 to 0.6644. FtoS:  $-0.0704$  to  $0.0730$ . CtoF:  $0.0572$  to  $0.2166$ . The FtoS coefficient is not significantly different from 0. (d)  $R^2 = 72.2\%$ ,  $s = 7.043$ .

**11.33** (a) For example: All distributions are skewed to varying degrees—GINI and CORRUPT to the right, the other three to the left. CORRUPT and DEMOCRACY have the most skewness. (b) GINI is negatively correlated to the other four variables (ranging from  $-0.396$  to  $-0.050$ ), while all other correlations are positive and more substantial (0.525 or more).

**11.35** (a) Refer to your regression output. (b) For example, the  $t$  statistic for the GINI coefficient grows from  $t = -0.42$  ( $P = 0.675$ ) to  $t = 4.25$  ( $P < 0.0005$ ). The DEMOCRACY  $t$  is 3.53 in the third model ( $P <$

0.0005) but drops to 0.71 ( $P = 0.479$ ) in the fourth model. **(c)** A good choice is to use GINI, LIFE, and CORRUPT: all three coefficients are significant, and  $R^2 = 77.0\%$  is nearly the same as for the fourth model from Exercise 11.34.

**11.37 (a)** Plot suggests greater variation in VO+ for large OC.  $\widehat{VO^+} = 334 + 19.5OC$ ,  $t = 4.73$ ,  $P < 0.0005$ . Plot of residuals against OC is slightly curved. **(b)**  $\widehat{VO^+} = 58 + 6.41OC + 53.9TRAP$ . Coefficient of OC is not significantly different from 0 ( $t = 1.25$ ,  $P = 0.221$ ), but coefficient of TRAP is significantly different from 0 ( $t = 3.50$ ,  $P = 0.002$ ). This is consistent with the correlations found in Exercise 11.36.

**11.39** The correlations are 0.840 (LVO+ and LVO-), 0.774 (LVO+ and LOC), and 0.755 (LVO+ and LTRAP). Regression equations,  $t$  statistics,  $R^2$ , and  $s$  for each model:  $\widehat{LVO^+} = 4.38 + 0.706LOC$ ;  $t = 6.58$ ,  $P < 0.0005$ ;  $R^2 = 0.599$ ,  $s = 0.3580$ .  $\widehat{LVO^+} = 4.26 + 0.430LOC + 0.424LTRAP$ ;  $t = 2.56$ ,  $P = 0.016$ ;  $t = 2.06$ ,  $P = 0.048$ ;  $R^2 = 0.652$ ,  $s = 0.3394$ .  $\widehat{LVO^+} = 0.872 + 0.392LOC + 0.028LTRAP + 0.672LVO^-$ ;  $t = 3.40$ ,  $P = 0.002$ ;  $t = 0.18$ ,  $P = 0.862$ ;  $t = 5.71$ ,  $P < 0.0005$ ;  $R^2 = 0.842$ ,  $s = 0.2326$ . As before, this suggests a model without LTRAP:  $\widehat{LVO^+} = 0.832 + 0.406LOC + 0.682LVO^-$ ;  $t = 4.93$ ,  $P < 0.0005$ ;  $t = 6.57$ ,  $P < 0.0005$ ;  $R^2 = 0.842$ ,  $s = 0.2286$ .

**11.41** Regression equations,  $t$  statistics,  $R^2$ , and  $s$  for each model:  $\widehat{LVO^-} = 5.21 + 0.441LOC$ ;  $t = 3.59$ ,  $P = 0.001$ ;  $R^2 = 0.308$ ,  $s = 0.4089$ .  $\widehat{LVO^-} = 5.04 + 0.057LOC + 0.590LTRAP$ ;  $t = 0.31$ ,  $P = 0.761$ ;  $t = 2.61$ ,  $P = 0.014$ ;  $R^2 = 0.443$ ,  $s = 0.3732$ .  $\widehat{LVO^+} = 1.57 - 0.293LOC + 0.245LTRAP + 0.813LVO^+$ ;  $t = -2.08$ ,  $P = 0.047$ ;  $t = 1.47$ ,  $P = 0.152$ ;  $t = 5.71$ ,  $P < 0.0005$ ;  $R^2 = 0.748$ ,  $s = 0.2558$ .  $\widehat{LVO^-} = 1.31 - 0.188LOC + 0.890LVO^+$ ;  $t = -1.52$ ,  $P = 0.140$ ;  $t = 6.57$ ,  $P < 0.0005$ ;  $R^2 = 0.728$ ,  $s = 0.2611$ .

**11.43 (a)**  $y_i = \beta_0 + \beta_1x_{i1} + \beta_2x_{i2} + \beta_3x_{i3} + \beta_4x_{i4} + \varepsilon_i$ , where  $i = 1, 2, \dots, 69$ ;  $\varepsilon_i$  are independent  $N(0, \sigma)$  random variables. **(b)**  $\widehat{PCB} = 0.94 + 11.87x_1 + 3.76x_2 + 3.88x_3 + 4.18x_4$ . All coefficients are significantly different from 0, although the constant 0.937 is not ( $t = 0.76$ ,  $P = 0.449$ ).  $R^2 = 0.989$ ,  $s = 6.382$ . **(c)** The residuals appear to be roughly Normal, but with two outliers. There are no clear patterns when plotted against the explanatory variables.

**11.45 (a)**  $\widehat{PCB} = -1.02 + 12.64PCB52 + 0.31PCB118 + 8.25PCB138$ ,  $R^2 = 0.973$ ,  $s = 9.945$ . **(b)**  $b_2 = 0.313$ ,  $P = 0.708$ . **(c)** In Exercise 11.43,  $b_2 = 3.76$ ,  $P < 0.0005$ .

**11.47** The model is  $y_i = \beta_0 + \beta_1x_{i1} + \beta_2x_{i2} + \beta_3x_{i3} + \beta_4x_{i4} + \varepsilon_i$ , where  $i = 1, 2, \dots, 69$ ;  $\varepsilon_i$  are independent  $N(0, \sigma)$  random variables. Regression gives

$\widehat{TEQ} = 1.06 - 0.097PCB52 + 0.306PCB118 + 0.106PCB138 - 0.004PCB180$  with  $R^2 = 0.677$ . Only the constant (1.06) and the PCB118 coefficient (0.306) are significantly different from 0. Residuals are slightly right-skewed and show no clear patterns when plotted with the explanatory variables.

**11.49 (a)** The correlations are all positive, ranging from 0.227 (LPCB28 and LPCB180) to 0.956 (LPCB and LPCB138). LPCB28 has one outlier (Specimen 39) when plotted with the other variables; except for that point, all scatterplots appear fairly linear. **(b)** All correlations are higher with the transformed data.

**11.51** It appears that a good model is LPCB126 and LPCB28 ( $R^2 = 0.768$ ). Adding more variables does not appreciably increase  $R^2$  or decrease  $s$ .

**11.53**  $\bar{x}$ ,  $M$ ,  $s$ , and  $IQR$  for each variable: Taste: 24.53, 20.95, 16.26, 23.9. Acetic: 5.498, 5.425, 0.571, 0.656. H2S: 5.942, 5.329, 2.127, 3.689. Lactic: 1.442, 1.450, 0.3035, 0.430. None of the variables show striking deviations from Normality. Taste and H2S are slightly right-skewed, and Acetic has two peaks. There are no outliers.

**11.55**  $\widehat{Taste} = -61.6 + 15.6Acetic$ ;  $t = 3.48$ ,  $P = 0.002$ . The residuals seem to have a Normal distribution but are positively associated with both H2S and Lactic.

**11.57**  $\widehat{Taste} = -29.9 + 37.7Lactic$ ;  $t = 5.25$ ,  $P < 0.0005$ . The residuals seem to have a Normal distribution;

there are no striking patterns for residuals against the other variables.

**11.59**  $\hat{Tast} = -26.9 + 3.80Acetic + 5.15H2S$ . For the coefficient of Acetic,  $t = 0.84$  and  $P = 0.406$ . This model is not much better than the model with H2S alone; Acetic and H2S are correlated ( $r = 0.618$ ), so Acetic does not add significant information if H2S is included.

**11.61**  $\hat{Tast} = -28.9 + 0.33Acetic + 3.91H2S + 19.7Lactic$ . The coefficient of Acetic is not significantly different from 0 ( $P = 0.942$ ). Residuals of this regression appear to be Normally distributed and show no patterns in scatterplots with the explanatory variables. It appears that the H2S/Lactic model is best.

## CHAPTER 12

**12.1** (a)  $H_0$  says the population means are all equal. (b) Experiments are best for establishing causation. (c) ANOVA is used to compare means. ANOVA *assumes* all variances are equal. (d) Multiple-comparisons procedures are used when we wish to determine which means are significantly different but have no specific relations in mind before looking at the data.

**12.3** (a) Yes:  $7/4 = 1.75 < 2$ . (b) 16, 25, and 49. (c) 31.2647. (d) 5.5915.

**12.5** (a) This is the description of *between*-group variation. (b) The *sums of squares* will add. (c)  $\sigma$  is a parameter. (d) A small  $P$  means the means are not all the same, but the distributions may still overlap.

**12.7** Assuming the  $t$  (ANOVA) test establishes that the means are different, contrasts and multiple comparisons provide no further useful information.

**12.9** (a)  $df = 3$  and 20. In Table E,  $3.10 < 3.18 < 3.86$ . (c)  $0.025 < P < 0.05$ . (d) We can conclude only that at least one mean is different from the others.

**12.11** (a)  $df$  are 3 and 60.  $F = 2.54$ .  $2.18 < F < 2.76$ , so  $0.050 < P < 0.100$ . (Software gives  $P = 0.0649$ .) (b)  $df$  are 2 and 24.  $F = 4.047$ .  $3.40 < F < 4.32$ , so  $0.025 < P < 0.050$ . (Software gives  $P = 0.0306$ .)

**12.13** (a) Response: egg cholesterol level. Populations: chickens with different diets or drugs.  $I = 3$ ,  $n_1 = n_2 = n_3 = 25$ ,  $N = 75$ . (b) Response: rating on five-point scale. Populations: the three groups of students.  $I = 3$ ,  $n_1 = 31$ ,  $n_2 = 18$ ,  $n_3 = 45$ ,  $N = 94$ . (c) Response: quiz score. Populations: students in each TA group.  $I = 3$ ,  $n_1 = n_2 = n_3 = 14$ ,  $N = 42$ .

**12.15** For all three situations, we test  $H_0: \mu_1 = \mu_2 = \mu_3$ ;  $H_a$ : at least one mean is different. (a) DFM 2, DFE 72, DFT 74.  $F(2, 72)$ . (b) DFM 2, DFE 91, DFT 93.  $F(2, 91)$ . (c) DFM 2, DFE 39, DFT 41.  $F(2, 39)$ .

**12.17** (a) This sounds like a fairly well-designed experiment, so the results should at least apply to this farmer's breed of chicken. (b) It would be good to know what proportion of the total student body falls in each of these groups—that is, is anyone overrepresented in this sample? (c) Effectiveness teaching one topic (power calculations) might not reflect overall effectiveness.

**12.19** (a)  $df = 4$  and 178. (b)  $5 + 146 = 151$  athletes were used. (c) For example, the individuals could have been outliers in terms of their ability to withstand the water bath pain. In the case of either low or high outliers, their removal would lessen the standard deviation for their sport and move that sports mean (removing a high outlier would lower the mean and removing a low outlier would raise the mean).

**12.21** (a)  $\psi = \mu_P$  and  $P - 1/4(\mu_{Text} + \mu_{Email} + \mu_{FB} + \mu_{MSN})$ . (b)  $H_0: \psi = 0$  versus  $H_a: \psi > 0$ . (c)  $t = 1.894$  with  $df = 138$ .  $P = 0.0302$ .

**12.23** (a) The table below gives the sample sizes, means, and standard deviations.

Food	$n$	$\bar{x}$	$s$
------	-----	-----------	-----

Comfort	22	4.887	0.573
Organic	20	5.584	0.594
Control	20	5.082	0.622

**(b)** Comfort food is relatively symmetric. Organic food has its most prevalent values at the extremes. Control could be called left-skewed (it does not look very symmetric).

**12.25 (a)** The means are not all equal for the three groups. Organic appears to differ from both Comfort and Control; Comfort and Control are not significantly different from each other. **(b)** The decrease in variability for the three groups and the curve in the Normal quantile plot might make us question Normality.

**12.27 (a)**  $I = 3$ ,  $N = 120$ , so  $df = 2$  and 117. **(b)** From Table E,  $P < 0.001$ . Using software,  $P = 0.0003$ . **(c)** We really shouldn't generalize these results beyond what might occur in similar shops in Mexico.

**12.29 (a)**  $F$  can be made very small (close to 0), and  $P$  close to 1. **(b)**  $F$  increases, and  $P$  decreases.

**12.31 (a)**

Group	$n$	$\bar{x}$	$s$
Control	35	-1.01	11.50
Group	34	-10.79	11.14
Individual	35	-3.71	9.08

**(b)** Yes;  $2(9.08) = 18.16 > 11.50$ . **(c)** Control is closest to a symmetric distribution; Individual seems left-skewed. However, with sample sizes at least 34 in each group, moderate departures from Normality are not a problem.

**12.33 (a)** The new group means and standard deviations will be the old means and standard deviations divided by 2.2. **(b)** Dividing by a constant will not change the Normality of the data. The test statistic is  $F = 7.77$  with  $P$ -value 0.001. These are exactly the same values obtained in Exercise 12.32.

**12.35 (a)** Based on the sample means, fiber is cheapest and cable is most expensive. **(b)** Yes; the ratio is 1.55. **(c)**  $df = 2$  and 44;  $0.025 < P < 0.050$ , or  $P = 0.0427$ .

**12.37 (a)** The variation in sample size is some cause for concern, but there can be no extreme outliers in a 1-to-7 scale, so ANOVA is probably reliable. **(b)** Yes:  $1.26/1.03 = 1.22 < 2$ . **(c)**  $F(4, 405)$ ,  $P = 0.0002$ . **(d)** Hispanic Americans are highest, Japanese are in the middle, the other three are lowest.

**12.39 (a)** Activity seems to increase with both drugs, and Drug B appears to have a greater effect. **(b)** Yes; the standard deviation ratio is 1.49.  $s_p = 3.487$ . **(c)**  $df = 4$  and 20. **(d)**  $0.05 < P < 0.10$ ; software gives  $P = 0.0642$ .

**12.41 (a)**  $\psi_1 = \mu_2 - (\mu_1 + \mu_4)/2$ . **(b)**  $\psi_2 = (\mu_1 + \mu_2 + \mu_4)/3 - \mu_3$ .

**12.43 (a)** Yes; the ratio is 1.25.  $s_p = 0.7683$ . **(b)**  $df = 2$  and 767;  $P < 0.001$ . **(c)** Compare faculty to the student average:  $\psi = \mu_2 - (\mu_1 + \mu_3)/2$ . We test  $H_0: \psi = 0$ ;  $H_a: \psi > 0$ . We find  $c = 0.585$ ,  $t = 5.99$ , and  $P < 0.0001$ .

**12.45 (a)** All three distributions show no particular skewness. Control:  $n = 15$ ,  $\bar{x} = 0.21887$ ,  $s = 0.01159$  g/cm<sup>2</sup>. Low dose:  $n = 15$ ,  $\bar{x} = 0.21593$ ,  $s = 0.01151$  g/cm<sup>2</sup>. High dose:  $n = 15$ ,  $\bar{x} = 0.23507$ ,  $s = 0.01877$  g/cm<sup>2</sup>. **(b)** All three distributions appear to be nearly Normal. **(c)**  $F = 7.72$ ,  $df = 2$  and 42,  $P = 0.001$ . **(d)** For Bonferroni,  $t^{**} = 2.49$  and  $MSD = 0.0131$ . The high-dose mean is significantly different from the

other two. (e) High doses increase bone mineral density.

**12.47** (a) Control:  $n = 10$ ,  $\bar{x} = 601.10$ ,  $s = 27.36$  mg/cm<sup>3</sup>. Low jump:  $n = 10$ ,  $\bar{x} = 612.50$ ,  $s = 19.33$  mg/cm<sup>3</sup>. High jump:  $n = 10$ ,  $\bar{x} = 638.70$ ,  $s = 16.59$  mg/cm<sup>3</sup>. Pooling is reasonable. (b)  $F = 7.98$ ,  $df = 2$  and  $27$ ,  $P = 0.002$ . We conclude that not all means are equal.

**12.49** (a)  $\psi_1 = \mu_1 - (\mu_2 + \mu_4)/2$  and  $\psi_2 = (\mu_3 - \mu_2) - (\mu_5 - \mu_4)$ . (b)  $c_1 = -3.9$ ,  $SE_{c_1} = 2.1353$ ,  $c_2 = 2.35$ , and  $SE_{c_2} = 3.487$ , (c) The first contrast is significant ( $t = -1.826$ ), but the second is not ( $t = -0.674$ ).

**12.51** (a) ECM1:  $n = 3$ ,  $\bar{x} = 65.0\%$ ,  $s = 8.66\%$ . ECM2:  $n = 3$ ,  $\bar{x} = 63.33\%$ ,  $s = 2.89\%$ . ECM3:  $n = 3$ ,  $\bar{x} = 73.33\%$ ,  $s = 2.89\%$ . MAT1:  $n = 3$ ,  $\bar{x} = 23.33\%$ ,  $s = 2.89\%$ . MAT2:  $n = 3$ ,  $\bar{x} = 6.67\%$ ,  $s = 2.89\%$ . MAT3:  $n = 3$ ,  $\bar{x} = 11.67\%$ ,  $s = 2.89\%$ . Pooling is risky because  $8.66/2.89 > 2$ . (b)  $F = 137.94$ ,  $df = 5$  and  $12$ ,  $P < 0.0005$ . We conclude that the means are not the same.

**12.53** (a)  $\psi_1 = \mu_5 - 0.25(\mu_1 + \mu_2 + \mu_3 + \mu_4)$ .  $\psi_2 = 0.5(\mu_1 + \mu_2) - 0.5(\mu_3 + \mu_4)$ .  $\psi_3 = (\mu_1 - \mu_2) - (\mu_3 - \mu_4)$ . (b) From Exercise 12.26, we have  $s_p = 18.421$ .  $c_1 = 14.65$ ,  $c_2 = 6.1$ , and  $c_3 = -0.5$ .  $s_{c_1} = 4.209$ ,  $s_{c_2} = 3.874$ , and  $s_{c_3} = 3.784$ . (c)  $t_1 = 3.48$ .  $t_2 = 1.612$ .  $t_3 = -0.132$ .  $t_{114, 0.975} = 1.980$ . Two-tailed  $P$ -values are 0.0007, 0.1097, and 0.8952. The first two contrasts are significant and the third is not.

**12.55** (a) The plot shows granularity (which varies between groups), but that should not make us question independence; it is due to the fact that the scores are all integers. (b) The ratio of the largest to the smallest standard deviations is less than 2. (c) Apart from the granularity, the quantile plots are reasonably straight. (d) Again, apart from the granularity, the quantile plots look pretty good.

**12.57** (a)  $\psi_1 = (\mu_1 + \mu_2 + \mu_3)/3 - \mu_4$ ,  $\psi_2 = (\mu_1 + \mu_2)/2 - \mu_3$ ,  $\psi_3 = \mu_1 - \mu_2$ . (b) The pooled standard deviation is  $s_p = 1.1958$ .  $SE_{c_1} = 0.2355$ ,  $SE_{c_2} = 0.1413$ ,  $SE_{c_3} = 0.1609$ . (c) Testing  $H_0: \psi_i = 0$ ;  $H_a: \psi_i \neq 0$ ; for each contrast, we find  $c_1 = -12.51$ ,  $t_1 = -53.17$ ,  $P_1 < 0.0005$ ;  $c_2 = 1.269$ ,  $t_2 = 8.98$ ,  $P_2 < 0.0005$ ;  $c_3 = 0.191$ ,  $t_3 = 1.19$ ,  $P_3 = 0.2359$ . The Placebo mean is significantly higher than the average of the other three, while the Keto mean is significantly lower than the average of the two Pyr means. The difference between the Pyr means is not significant (meaning the second application of the shampoo is of little benefit).

**12.59** The means all increase by 5%, but everything else (standard deviations, standard errors, and the ANOVA table) is unchanged.

**12.61** All distributions are reasonably Normal, and standard deviations are close enough to justify pooling. For PRE1,  $F = 1.13$ ,  $df = 2$  and  $63$ ,  $P = 0.329$ . For PRE2,  $F = 0.11$ ,  $df = 2$  and  $63$ ,  $P = 0.895$ . Neither set of pretest scores suggests a difference in means.

**12.63**  $\hat{\text{Score}} = 4.432 - 0.000102 \text{ Friends}$ . The slope is not significantly different from 0 ( $t = -0.28$ ,  $P = 0.782$ ), and the regression explains only 0.1% of the variation in score. Residuals suggest a possible curved relationship.

**12.67** (b) Answers will vary with choice of  $H_a$  and desired power. For example, with  $\mu_1 = \mu_2 = 4.4$ ,  $\mu_3 = 5$ ,  $\sigma = 1.2$ , three samples of size 75 will produce power 0.78.

**12.69** The design can be similar, although the types of music might be different. Bear in mind that spending at a casual restaurant will likely be less than at the restaurants examined in Exercise 12.40; this might also mean that the standard deviations could be smaller. Decide how big a difference in mean spending you want to detect, then do some power computations.

## CHAPTER 13

**13.1** (a) Two-way ANOVA is used when there are two factors. (b) Each level of A should occur with all three levels of B. (c) The RESIDUAL part of the model represents the error. (d)  $DFAB = (I - 1)(J - 1)$ .

**13.3** (a) Reject  $H_0$  when  $F$  is large. (b) Mean squares equal the sum of squares divided by degrees of freedom. (c) The test statistics have an  $F$  distribution. (d) If the sample sizes are not the same, the sums of squares may not add.

**13.5** (a)  $N = 36$ .  $DFA = 2$ ,  $DFB = 1$ ,  $DFAB = 2$ ,  $DFE = 30$ , so  $F$  has 2 and 30 degrees of freedom. (c)  $P > 0.10$ . (d) Interaction is not significant; the interaction plot should have roughly parallel lines.

**13.7** (a) The factors are gender ( $I = 2$ ) and age ( $J = 3$ ). The response variable is the percent of pretend play.  $N = (2)(3)(11) = 66$ . (b) The factors are time after harvest ( $I = 5$ ) and amount of water ( $J = 2$ ). The response variable is the percent of seeds germinating.  $N = 30$ . (c) The factors are mixture ( $I = 6$ ) and freezing/thawing cycles ( $J = 3$ ). The response variable is the strength of the specimen.  $N = 54$ . (d) The factors are training programs ( $I = 4$ ) and the number of days to give the training ( $J = 2$ ). The response variable is not specified but presumably is some measure of the training's effectiveness.  $N = 80$ .

**13.9** (a) The same students were tested twice. (b) The interactions plot shows a definite interaction; the control group's mean score decreased, while the expressive- writing group's mean increased. (c)  $No. 2(5.8) = 11.6 < 14.3$ .

**13.11** (a) Recall from Chapter 12 that ANOVA is robust against reasonable departures from Normality, especially when sample sizes are similar (and as large as these). (b) Yes.  $1.62/0.82 = 1.98 < 2$ . The ANOVA table is below.

Source	DF	SS	MS	$F$	$P$
Age	6	31.97	5.328	4.400	0.0003
Gender	1	44.66	44.66	36.879	0.0000
Age $\times$ Gender	6	13.22	2.203	1.819	0.0962
Error	232	280.95	1.211		
Total	245	370.80			

**13.13** (a) There appears to be an interaction; a thank-you increases repurchase intent for those with short history and decreases it for customers with long history. (b) The marginal means for history (6.245 and 7.45) convey the fact that repurchase intent is higher for customers with long history. The thank-you marginal means (6.61 and 7.085) are less useful because of the interaction.

**13.15** (a) The plot suggests a possible interaction. (b) By subjecting the same individual to all four treatments, rather than four individuals to one treatment each, we reduce the within-groups variability.

**13.17** (a) We'd expect reaction times to slow with older individuals. If bilingualism helps brain functioning, we would not expect that group to slow as much as the monolingual group. The expected interaction is seen in the plot; mean total reaction time for the older bilingual group is much less than for the older monolingual group; the lines are not parallel. (b) The interaction is just barely not significant ( $F = 3.67$ ,  $P = 0.059$ ). Both main effects are significant ( $P = 0.000$ ).

**13.19** (a) There may be an interaction; for a favorable process, a favorable outcome increases satisfaction quite a bit more than for an unfavorable process (+2.32 versus +0.24). (b) This time, the increase in satisfaction from a favorable outcome is less for a favorable process (+0.49 versus +1.32). (c) There seems to be a three-factor interaction, because the interactions in parts (a) and (b) are different.

**13.21** Humor slightly increases satisfaction (3.58 with no humor, 3.96 with humor). The process and outcome effects are greater: favorable process, 4.75; unfavorable process, 2.79; favorable outcome, 4.32; unfavorable outcome, 3.22.

**13.23** The largest-to-smallest ratio is 1.26, and the pooled standard deviation is 1.7746.

**13.25** Except for female responses to purchase intention, means decreased from Canada to the United States to France. Females had higher means than men in almost every case, except for French responses to credibility and purchase intention (a modest interaction).

**13.27 (a)** Intervention, 11.6; control, 9.967; baseline, 10.0; 3 months, 11.2; 6 months, 11.15. Overall, 10.783. The row means suggest that the intervention group showed more improvement than the control group. **(b)** Interaction means that the mean number of actions changes differently over time for the two groups.

**13.29** With  $I = 3$ ,  $J = 2$ , and 6 observations per cell, we have  $DFA = 2$ ,  $DFB = 1$ ,  $DFAB = 2$ , and  $DFE = 30$ .  $3.32 < 3.45 < 4.18$ , so  $0.025 < P_A < 0.05$  (software gives 0.0448).  $2.49 < 2.88$ , so  $P_B > 0.10$  (software gives 0.1250).  $1.14 < 2.49$ , so  $P_{AB} > 0.10$  (software gives 0.3333). The only significant effect is the main effect for factor A.

**13.31 (a)** There is little evidence of an interaction. **(b)**  $s_p = 0.1278$ . **(c)**  $\psi_1 = (\mu_{\text{new, city}} + \mu_{\text{new, hw}})/2 - (\mu_{\text{old, city}} + \mu_{\text{old, hw}})/2$ .  $\psi_2 = \mu_{\text{new, city}} - \mu_{\text{new, hw}}$ .  $\psi_3 = \mu_{\text{old, hw}} - \mu_{\text{old, city}}$ . **(d)** By subjecting the same individual to all four treatments, rather than four individuals to one treatment each, we reduce the within-groups variability.

**13.33 (b)** There seems to be a fairly large difference between the means based on how much the rats were allowed to eat but not very much difference based on the chromium level. There may be an interaction: the NM mean is lower than the LM mean, while the NR mean is higher than the LR mean. **(c)** L mean: 4.86. N mean: 4.871. M mean: 4.485. R mean: 5.246. LR minus LM: 0.63. NR minus NM: 0.892. Mean GITH levels are lower for M than for R; there is not much difference between L and N. The difference between M and R is greater among rats who had normal chromium levels in their diets (N).

**13.35 (a)**  $s_p = \$38.14$ ,  $df = 105$ . **(b)** Yes; the largest-to-smallest ratio is 1.36. **(c)** Individual sender, \$70.90; group sender, \$48.85; individual responder, \$59.75; group responder, \$60.00. **(d)** There appears to be an interaction; individuals send more money to groups, while groups send more money to individuals. **(e)**  $P = 0.0033$ ,  $P = 0.9748$ , and  $P = 0.1522$ . Only the main effect of sender is significant.

**13.37** Yes; the iron-pot means are the highest, and  $F$  for testing the effect of the pot type is very large.

**13.39 (a)** In the order listed in the table:  $\bar{x}_{11} = 25.0307$ ,  $s_{11} = 0.0011541$ ;  $\bar{x}_{12} = 25.0280$ ,  $s_{12} = 0$ ;  $\bar{x}_{13} = 25.0260$ ,  $s_{13} = 0$ ;  $\bar{x}_{21} = 25.0167$ ,  $s_{21} = 0.0011541$ ;  $\bar{x}_{22} = 25.0200$ ,  $s_{22} = 0.002000$ ;  $\bar{x}_{23} = 25.0160$ ,  $s_{23} = 0$ ;  $\bar{x}_{31} = 25.0063$ ,  $s_{31} = 0.001528$ ;  $\bar{x}_{32} = 25.0127$ ,  $s_{32} = 0.0011552$ ;  $\bar{x}_{33} = 25.0093$ ,  $s_{33} = 0.0011552$ ;  $\bar{x}_{41} = 25.0120$ ,  $s_{41} = 0$ ;  $\bar{x}_{42} = 25.0193$ ,  $s_{42} = 0.0011552$ ;  $\bar{x}_{43} = 25.0140$ ,  $s_{43} = 0.004000$ ;  $\bar{x}_{51} = 24.9973$ ,  $s_{51} = 0.001155$ ;  $\bar{x}_{52} = 25.0060$ ,  $s_{52} = 0$ ;  $\bar{x}_{53} = 25.0003$ ,  $s_{53} = 0.001528$ . **(b)** Except for Tool 1, mean diameter is highest at Time 2. Tool 1 had the highest mean diameters, followed by Tool 2, Tool 4, Tool 3, and Tool 5. **(c)**  $F_A = 412.94$ ,  $df = 4$  and  $30$ ,  $P < 0.0005$ .  $F_B = 43.60$ ,  $df = 2$  and  $30$ ,  $P < 0.0005$ .  $F_{AB} = 7.65$ ,  $df = 8$  and  $30$ ,  $P < 0.0005$ . **(d)** There is strong evidence of a difference in mean diameter among the tools (A) and among the times (B). There is also an interaction (specifically, Tool 1's mean diameters changed differently over time compared with the other tools).

**13.41 (a)** All three  $F$ -values have  $df = 1$  and  $945$ ; the  $P$ -values are  $< 0.001$ ,  $< 0.001$ , and  $0.1477$ . Gender and handedness both have significant effects on mean lifetime, but there is no interaction. **(b)** Women live about 6 years longer than men (on the average), while right-handed people average 9 more years of life than left-handed people. Handedness affects both genders in the same way, and vice versa.

**13.43 (a)** and **(b)** The first three means and standard deviations are  $\bar{x}_{1,1} = 3.2543$ ,  $s_{1,1} = 0.2287$ ;  $\bar{x}_{1,2} = 2.7636$ ,  $s_{1,2} = 0.0666$ ;  $\bar{x}_{1,3} = 2.8429$ ,  $s_{1,3} = 0.2333$ . The standard deviations range from 0.0666 to 0.3437, for a ratio of 5.16—larger than we like. **(c)** For Plant,  $F = 1301.32$ ,  $df = 3$  and  $224$ ,  $P < 0.0005$ . For Water,  $F = 9.76$ ,  $df = 6$  and  $224$ ,  $P < 0.0005$ . For interaction,  $F = 5.97$ ,  $df = 18$  and  $224$ ,  $P < 0.0005$ .

**13.45** The seven  $F$  statistics are 184.05, 115.93, 208.87, 218.37, 220.01, 174.14, and 230.17, all with  $df = 3$  and  $32$  and  $P < 0.0005$ .

**13.47** Fresh: Plant,  $F = 81.45$ ,  $df = 3$  and  $84$ ,  $P < 0.0005$ ; Water,  $F = 43.71$ ,  $df = 6$  and  $84$ ,  $P < 0.0005$ ; interaction,  $F = 1.79$ ,  $df = 18$  and  $84$ ,  $P = 0.040$ . Dry: Plant,  $F = 79.93$ ,  $df = 3$  and  $84$ ,  $P < 0.0005$ ; Water,  $F = 44.79$ ,  $df = 6$  and  $84$ ,  $P < 0.0005$ ; interaction,  $F = 2.22$ ,  $df = 18$  and  $84$ ,  $P = 0.008$ .

**13.49** The twelve  $F$  statistics are fresh biomass: 15.88, 11.81, 62.08, 10.83, 22.62, 8.20, and 10.81; dry biomass: 8.14, 26.26, 22.58, 11.86, 21.38, 14.77, and 8.66, all with  $df = 3$  and 15 and  $P < 0.003$ .

**13.51 (a)** Gender:  $df = 1$  and 174. Floral characteristic:  $df = 2$  and 174. Interaction:  $df = 2$  and 174. **(b)** Damage to males was higher for all characteristics. For males, damage was higher under characteristic level 3, while for females, the highest damage occurred at level 2. **(c)** Three of the standard deviations are at least half as large as the means. Because the response variable (leaf damage) had to be nonnegative, this suggests that these distributions are right-skewed.

**13.53** Men in CS:  $n = 39$ ,  $\bar{x} = 7.79487$ ,  $s = 1.50752$ . Men in EOS:  $n = 39$ ,  $\bar{x} = 7.48718$ ,  $s = 2.15054$ . Men in Other:  $n = 39$ ,  $\bar{x} = 7.41026$ ,  $s = 1.56807$ . Women in CS:  $n = 39$ ,  $\bar{x} = 8.84615$ ,  $s = 1.13644$ . Women in EOS:  $n = 39$ ,  $\bar{x} = 9.25641$ ,  $s = 0.75107$ . Women in Other:  $n = 39$ ,  $\bar{x} = 8.61539$ ,  $s = 1.16111$ . The means suggest that females have higher HSE grades than males. For a given gender, there is not too much difference among majors. Normal quantile plots show no great deviations from Normality, apart from the granularity of the grades (most evident among women in EO). In the ANOVA, only the effect of gender is significant ( $F = 50.32$ ,  $df = 1$  and 228,  $P < 0.0005$ ).

**13.55** Men in CS:  $n = 39$ ,  $\bar{x} = 526.949$ ,  $s = 100.937$ . Men in EOS:  $n = 39$ ,  $\bar{x} = 507.846$ ,  $s = 57.213$ . Men in Other:  $n = 39$ ,  $\bar{x} = 487.564$ ,  $s = 108.779$ . Women in CS:  $n = 39$ ,  $\bar{x} = 543.385$ ,  $s = 77.654$ . Women in EOS:  $n = 39$ ,  $\bar{x} = 538.205$ ,  $s = 102.209$ . Women in Other:  $n = 39$ ,  $\bar{x} = 465.026$ ,  $s = 82.184$ . The means suggest that students who stay in the sciences have higher mean SATV scores than those who end up in the Other group. Female CS and EO students have higher scores than males in those majors, but males have the higher mean in the Other group. Normal quantile plots suggests some right-skewness in the “Women in CS” group and also some non-Normality in the tails of the “Women in EO” group. Other groups look reasonably Normal. In the ANOVA, only the effect of major is significant ( $F = 9.32$ ,  $df = 2$  and 228,  $P < 0.0005$ ).

# FORMULAS AND KEY IDEAS CARD

## CHAPTER 1

---

- **The mean  $\bar{x}$ .** If the  $n$  observations are  $x_1, x_2, \dots, x_n$ , their mean is

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

- **The median  $M$ .** Arrange all observations in order of size, from smallest to largest. If the number of observations  $n$  is odd, the median  $M$  is the center observation in the ordered list. Find the location of the median by counting  $(n + 1)/2$  observations up from the bottom of the list. If the number of observations  $n$  is even, the median  $M$  is the mean of the two center observations in the ordered list. The location of the median is again  $(n + 1)/2$  from the bottom of the list.
- **The quartiles  $Q_1$  and  $Q_3$ .** Arrange the observations in increasing order and locate the median  $M$  in the ordered list of observations.  $Q_1$  is the median of the observations whose position in the ordered list is to the left of the location of the overall median.  $Q_3$  is the median of the observations whose position in the ordered list is to the right of the location of the overall median.
- **The five-number summary.** The smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is

$$\text{Minimum } Q_1 \ M \ Q_3 \ \text{Maximum}$$

- **A boxplot.** A graph of the five-number summary. A central box spans the quartiles  $Q_1$  and  $Q_3$ . A line in the box marks the median  $M$ . Lines extend from the box out to the smallest and largest observations.
- **The interquartile range ( $IQR$ ).** The distance between the first and third quartiles,

$$IQR = Q_3 - Q_1$$

- **The  $1.5 \times IQR$  rule for outliers.** Call an observation a suspected outlier if it falls more than  $1.5 \times IQR$  above the third quartile or below the first quartile.

- **The variance  $s^2$ .** For  $n$  observations  $x_1, x_2, \dots, x_n$ ,

$$s^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 / n - 1$$

- **The standard deviation  $s$ .** Is the square root of the variance  $s^2$ .
- **Effect of a linear transformation.** Multiplying each observation by a positive number  $b$  multiplies both measures of center (mean and median) and measures of spread (interquartile range and standard deviation) by  $b$ . Adding the same number  $a$  (either positive or negative) to each observation adds  $a$  to measures of center and to quartiles and other percentiles but does not change measures of spread.
- **Density curve.** Is always on or above the horizontal axis and has area exactly 1 underneath it.
- **The median of a density curve.** The equal-areas point, the point that divides the area under the curve in half.
- **The mean of a density curve.** The balance point at which the curve would balance if made of solid material.
- **The 68–95–99.7 rule.** In the Normal distribution with mean  $\mu$  and standard deviation  $\sigma$ , approximately **68%** of the observations fall within  $\sigma$  of the mean  $\mu$ , approximately **95%** of the observations fall within  $\sigma$  of  $\mu$ , and approximately **99.7%** of the observations fall within  $3\sigma$  of  $\mu$ .
- **Standardizing and z-scores.** If  $x$  is an observation from a distribution that has mean  $\mu$  and standard deviation  $\sigma$ ,

$$z = \frac{x - \mu}{\sigma}$$

- **The standard Normal distribution.** The Normal distribution  $N(0, 1)$  with mean 0 and standard deviation 1. If a variable  $X$  has any Normal distribution  $N(\mu, \sigma)$  with mean  $\mu$  and standard deviation  $\sigma$ , then the standardized variable

$$Z = \frac{X - \mu}{\sigma}$$

has the standard Normal distribution.

- **Use of Normal quantile plots.** If the points on a Normal quantile plot lie close to a straight line, the plot indicates that the data are Normal. Systematic deviations from a straight line indicate a non-Normal distribution. Outliers appear as points that are far away from the overall pattern of the plot.

## CHAPTER 2

---

- **Response variable, explanatory variable.** A response variable measures an outcome of a study. An explanatory variable explains or causes changes in the response variables.
- **Scatterplot.** A scatterplot shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual.
- **Positive association, negative association.** Two variables are positively associated when above-average values of one tend to accompany above-average values of the other and below-average values also tend to occur together. Two variables are negatively associated when above-average values of one tend to accompany below-average values of the other, and vice versa.
- **Correlation.** The correlation measures the direction and strength of the linear relationship between two quantitative variables. Correlation is usually written as  $r$ . Suppose that we have data on variables  $x$  and  $y$  for  $n$  individuals. The means and standard deviations of the two variables are  $\bar{x}$  and  $s_x$  for the  $x$ -values, and  $\bar{y}$  and  $s_y$  for the  $y$ -values. The correlation  $r$  between  $x$  and  $y$  is

$$r = \frac{1}{n-1} \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

- **Straight lines.** Suppose that  $y$  is a response variable (plotted on the vertical axis) and  $x$  is an explanatory variable (plotted on the horizontal axis). A straight line relating  $y$  to  $x$  has an equation of the form

$$y = b_0 + b_1 x$$

In this equation,  $b_1$  is the **slope**, the amount by which  $y$  changes when  $x$  increases by one unit. The number  $b_0$  is the **intercept**, the value of  $y$  when  $x = 0$ .

- **Equation of the least-squares regression line.** We have data on an explanatory variable  $x$  and a response variable  $y$  for  $n$  individuals. The means and standard deviations of the sample data are  $\bar{x}$  and  $s_x$  for  $x$  and  $\bar{y}$  and  $s_y$  for  $y$ , and the correlation between  $x$  and  $y$  is  $r$ . The equation of the least-squares regression line of  $y$  on  $x$  is

$$\hat{y} = b_0 + b_1 x$$

with **slope**  $b_1 = r s_y / s_x$  and **intercept**  $b_0 = \bar{y} - b_1 \bar{x}$ .

- **$r^2$  in regression.** The square of the correlation,  $r^2$ , is the fraction of the variation in the values of  $y$  that is explained by the least-squares regression of  $y$  on  $x$ .
- **Residuals.** A residual is the difference between an observed value of the response variable and the value predicted by the regression line. That is,  $y - \hat{y}$ .
- **Outliers and influential observations in regression.** An outlier is an observation that lies outside the overall pattern of the other observations. Points that are outliers in the  $y$  direction of a scatterplot have large regression residuals, but other outliers need not have large residuals. An observation is influential for a statistical calculation if removing it would markedly change the result of the calculation. Points that are outliers in the  $x$  direction of a scatterplot are often influential for the least-squares regression line.
- **Simpson's paradox.** An association or comparison that holds for all of several groups can reverse direction when the data are combined to form a single group. This reversal is called Simpson's paradox.
- **Confounding.** Two variables are confounded when their effects on a response variable cannot be distinguished from each other. The confounded variables may be either explanatory variables or lurking variables.

## CHAPTER 3

---

- **Anecdotal evidence.** Anecdotal evidence is based on haphazardly selected individual cases, which often come to our attention because they are striking in some way. These cases need not be representative of any larger group of cases.
- **Available data.** Available data are data that were produced in the past for some other purpose but that may help answer a present question.
- **Observation versus experiment.** In an observational study we observe individuals and measure variables of interest but do not attempt to influence the responses. In an experiment we deliberately impose some treatment on individuals and we observe their responses.
- **Experimental units, subjects, treatment.** The individuals on which the experiment is done are the experimental units. When the units are human beings, they are called subjects. A specific experimental condition applied to the units is called a treatment.

- **Bias.** The design of a study is biased if it systematically favors certain outcomes.
- **Principles of experimental design.** 1. Compare two or more treatments. 2. Randomize—use impersonal chance to assign experimental units to treatments. 3. Repeat each treatment on many units to reduce chance variation in the results.
- **Statistical significance.** An observed effect so large that it would rarely occur by chance is called statistically significant.
- **Random digits.** A table of random digits is a list of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 that has the following properties: The digit in any position in the list has the same chance of being any one of 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. The digits in different positions are independent in the sense that the value of one has no influence on the value of any other.
- **Block design.** A block is a group of experimental units or subjects that are known before the experiment to be similar in some way that is expected to affect the response to the treatments. In a block design, the random assignment of units to treatments is carried out separately within each block.
- **Population and sample.** The entire group of individuals that we want information about is called the population. A sample is a part of the population that we actually examine in order to gather information.
- **Voluntary response sample.** A voluntary response sample consists of people who choose themselves by responding to a general appeal. Voluntary response samples are biased because people with strong opinions, especially negative opinions, are most likely to respond.
- **Simple random sample.** A simple random sample (SRS) of size  $n$  consists of  $n$  individuals from the population chosen in such a way that every set of  $n$  individuals has an equal chance to be the sample actually selected.
- **Probability sample.** A probability sample is a sample chosen by chance. We must know what samples are possible and what chance, or probability, each possible sample has.
- **Stratified random sample.** To select a stratified random sample, first divide the population into groups of similar individuals, called strata. Then choose a separate SRS in each stratum and combine these SRSs to form the full sample.
- **Undercoverage and nonresponse.** Undercoverage occurs when some groups

in the population are left out of the process of choosing the sample. Nonresponse occurs when an individual chosen for the sample can't be contacted or does not cooperate.

- **Parameters and statistics.** A parameter is a number that describes the population. A parameter is a fixed number, but in practice we do not know its value. A statistic is a number that describes a sample. The value of a statistic is known when we have taken a sample, but it can change from sample to sample. We often use a statistic to estimate an unknown parameter.
- **Sampling distribution.** The sampling distribution of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.
- **Bias and variability.** Bias concerns the center of the sampling distribution. A statistic used to estimate a parameter is unbiased if the mean of its sampling distribution is equal to the true value of the parameter being estimated. The variability of a statistic is described by the spread of its sampling distribution. This spread is determined by the sampling design and the sample size  $n$ . Statistics from larger probability samples have smaller spreads.
- **Managing bias and variability.** To reduce bias, use random sampling. When we start with a list of the entire population, simple random sampling produces unbiased estimates—the values of a statistic computed from an SRS neither consistently overestimate nor consistently underestimate the value of the population parameter. To reduce the variability of a statistic from an SRS, use a larger sample. You can make the variability as small as you want by taking a large enough sample.
- **Population size doesn't matter.** The variability of a statistic from a random sample does not depend on the size of the population, as long as the population is at least 100 times larger than the sample.
- **Basic data ethics.** The organization that carries out the study must have an institutional review board that reviews all planned studies in advance in order to protect the subjects from possible harm. All individuals who are subjects in a study must give their informed consent before data are collected. All individual data must be kept confidential. Only statistical summaries for groups of subjects may be made public.

## CHAPTER 4

---

- **Randomness and probability.** We call a phenomenon random if individual outcomes are uncertain but there is nonetheless a regular distribution of

outcomes in a large number of repetitions. The probability of any outcome of a random phenomenon is the proportion of times the outcome would occur in a very long series of repetitions.

- **Sample space.** The sample space  $S$  of a random phenomenon is the set of all possible outcomes.
- **Event.** An event is an outcome or a set of outcomes of a random phenomenon. That is, an event is a subset of the sample space.
- **Probability rules.** Rule 1. The probability  $P(A)$  of any event  $A$  satisfies  $0 \leq P(A) \leq 1$ . Rule 2. If  $S$  is the sample space in a probability model, then  $P(S) = 1$ . Rule 3. Two events  $A$  and  $B$  are disjoint if they have no outcomes in common and so can never occur together. If  $A$  and  $B$  are disjoint,  $P(A \text{ or } B) = P(A) + P(B)$ . This is the addition rule for disjoint events. Rule 4. The complement of any event  $A$  is the event that  $A$  does not occur, written as  $A^c$ . The complement rule states that  $P(A^c) = 1 - P(A)$ .
- **Probabilities in a finite sample space.** Assign a probability to each individual outcome. These probabilities must be numbers between 0 and 1 and must have sum 1. The probability of any event is the sum of the probabilities of the outcomes making up the event.
- **Equally likely outcomes.** If a random phenomenon has  $k$  possible outcomes, all equally likely, then each individual outcome has probability  $1/k$ . The probability of any event  $A$  is  $P(A) = (\text{count of outcomes in } A)/k$ .
- **The multiplication rule for independent events.** Rule 5. Two events  $A$  and  $B$  are independent if knowing that one occurs does not change the probability that the other occurs. If  $A$  and  $B$  are independent,  $P(A \text{ and } B) = P(A)P(B)$ . This is the multiplication rule for independent events.
- **Random variable.** A random variable is a variable whose value is a numerical outcome of a random phenomenon.
- **Discrete random variable.** A discrete random variable  $X$  has a finite number of possible values. The probability distribution of  $X$  lists the values and their probabilities:

Value of $X$	$x_1$	$x_2$	$x_3$	$\dots$	$x_k$
Probability	$p_1$	$p_2$	$p_3$	$\dots$	$p_k$

The probabilities  $p_i$  must satisfy two requirements: 1. Every probability  $p_i$  is a number between 0 and 1. 2.  $p_1 + p_2 + \dots + p_k = 1$ . Find the probability of any event by adding the probabilities  $p_i$  of the particular values  $x_i$  that make up

the event.

- **Continuous random variable.** A continuous random variable  $X$  takes all values in an interval of numbers. The probability distribution of  $X$  is described by a density curve. The probability of any event is the area under the density curve and above the values of  $X$  that make up the event.
- **Mean of a discrete random variable.** Suppose that  $X$  is a discrete random variable whose distribution is

Value of $X$	$x_1$	$x_2$	$x_3$	$\dots$	$x_k$
Probability	$p_1$	$p_2$	$p_3$	$\dots$	$p_k$

To find the **mean** of  $X$ , multiply each possible value by its probability, then add all the products:

$$\mu_X = x_1p_1 + x_2p_2 + \dots + x_kp_k$$

- **Law of large numbers.** Draw independent observations at random from any population with finite mean  $\mu$ . Decide how accurately you would like to estimate  $\mu$ . As the number of observations drawn increases, the mean  $\bar{x}$  of the observed values eventually approaches the mean  $\mu$  of the population as closely as you specified and then stays that close.
- **Rules for means.** Rule 1. If  $X$  is a random variable and  $a$  and  $b$  are fixed numbers, then  $\mu_{a+bX} = a + b\mu_X$ . Rule 2. If  $X$  and  $Y$  are random variables, then  $\mu_{X+Y} = \mu_X + \mu_Y$ .
- **Variance of a discrete random variable.** Suppose that  $X$  is a discrete random variable whose distribution is

Value of $X$	$x_1$	$x_2$	$x_3$	$\dots$	$x_k$
Probability	$p_1$	$p_2$	$p_3$	$\dots$	$p_k$

and that  $\mu_X$  is the mean of  $X$ . The variance of  $X$  is

$$\sigma_X^2 = (x_1 - \mu_X)^2p_1 + (x_2 - \mu_X)^2p_2 + \dots + (x_k - \mu_X)^2p_k$$

- **Rules for variances and standard deviations.** Rule 1. If  $X$  is a random variable and  $a$  and  $b$  are fixed numbers, then  $\sigma_{a+bX}^2 = b^2\sigma_X^2$ . Rule 2. If  $X$  and  $Y$  are independent random variables, then  $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$  and  $\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2$ . This is the addition rule for variances of independent random variables. Rule 3. If  $X$  and  $Y$  have correlation  $\rho$ , then  $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y$  and  $\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y$ . This is the general addition rule for variances of random variables. To find the standard

deviation, take the square root of the variance.

- **Rules of probability.** Rule 1.  $0 \leq P(A) \leq 1$  for any event  $A$ . Rule 2.  $P(S) = 1$ . Rule 3. Addition rule: If  $A$  and  $B$  are disjoint events, then  $P(A \text{ or } B) = P(A) + P(B)$ . Rule 4. Complement rule: For any event  $A$ ,  $P(A^c) = 1 - P(A)$ . Rule 5. Multiplication rule: If  $A$  and  $B$  are independent events, then  $P(A \text{ and } B) = P(A)P(B)$ .
- **Union.** The union of any collection of events is the event that at least one of the collection occurs.
- **Addition rule for disjoint events.** If events  $A$ ,  $B$ , and  $C$  are disjoint in the sense that no two have any outcomes in common, then  $P(\text{one or more of } A, B, C) = P(A) + P(B) + P(C)$ . This rule extends to any number of disjoint events.
- **General addition rule for unions of two events.** For any two events  $A$  and  $B$ ,  $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ .
- **Multiplication rule.** The probability that both of two events  $A$  and  $B$  happen together can be found by  $P(A \text{ and } B) = P(A)P(B | A)$ . Here  $P(B | A)$  is the conditional probability that  $B$  occurs, given the information that  $A$  occurs.
- **Definition of conditional probability.** When  $P(A) > 0$ , the conditional probability of  $B$  given  $A$  is  $P(B | A) = P(A \text{ and } B) / P(A)$ .
- **Intersection.** The intersection of any collection of events is the event that all of the events occur.
- **Bayes's rule.** Suppose that  $A_1, A_2, \dots, A_k$  are disjoint events whose probabilities are not 0 and add to exactly 1. That is, any outcome is in exactly one of these events. Then if  $C$  is any other event whose probability is not 0 or 1,

$$P(A_i | C) = \frac{P(C | A_i)P(A_i)}{P(C | A_1)P(A_1) + \dots + P(C | A_k)P(A_k)}$$

- **Independent events.** Two events  $A$  and  $B$  that both have positive probability are independent if  $P(B | A) = P(B)$ .

## CHAPTER 5

---

- **The sample mean  $\bar{x}$**  of an SRS of size  $n$  drawn from a large population with mean  $\mu$  and standard deviation  $\sigma$  has a sampling distribution with mean  $\mu_{\bar{x}} = \mu$  and standard deviation  $\sigma_{\bar{x}} = \sigma / \sqrt{n}$ .

- **Linear combinations of independent Normal random variables** have Normal distributions. In particular, if the population has a Normal distribution, so does  $\bar{x}$ .
- **The central limit theorem** states that for large  $n$  the sampling distribution of  $\bar{x}$  is approximately  $N(\mu, \sigma/\sqrt{n})$  for any population with mean  $\mu$  and finite standard deviation  $\sigma$ . This includes populations of both continuous and discrete random variables.
- **The binomial distribution.** A count  $X$  of successes has the binomial distribution  $B(n, p)$  when there are  $n$  trials, all independent, each resulting in a success or a failure, and each having the same probability  $p$  of a success. The mean of  $X$  is  $\mu_X = np$  and the standard deviation is  $\sigma_X = \sqrt{np(1-p)}$ .
- **The sample proportion of success**  $\hat{p} = X/n$  has mean  $\mu_{\hat{p}} = p$  and standard deviation  $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$ . It is an unbiased estimator of the population proportion  $p$ .
- **The sampling distribution of the count of successes.** The  $B(n, p)$  distribution is a good approximation to the sampling distribution of the count of successes in an SRS of size  $n$  from a large population containing proportion  $p$  of successes. We will use this approximation when the population is at least 20 times larger than the sample.
- **The sampling distribution of the sample proportion.** The sampling distribution of  $\hat{p}$  is not binomial but the  $B(n, p)$  distribution can be used to do probability calculations about  $\hat{p}$  by restating them in terms of the count  $X$ . We will use the  $B(n, p)$  distribution when the population is at least 20 times larger than the sample.
- **The Normal approximation to the binomial distribution** says that if  $X$  is a count having the  $B(n, p)$  distribution, then when  $n$  is large,  $X$  is approximately  $N(np, \sqrt{np(1-p)})$ . In addition, the sample proportion  $\hat{p} = X/n$  is  $N(p, \sqrt{p(1-p)/n})$ . We will use these approximations when  $np \geq 10$  and  $n(1-p) \geq 10$ . The continuity correction improves the accuracy of the Normal approximations.

## CHAPTER 6

---

- **Confidence interval.** The purpose of a confidence interval is to estimate an unknown parameter with an indication of how accurate the estimate is and of how confident we are that the result is correct. Any confidence interval has two parts: an interval computed from the data and a confidence level. The interval often has the form estimate  $\pm$  margin of error.

- **Confidence level.** The confidence level states the probability that the method will give a correct answer. That is, if you use 95% confidence intervals, in the long run 95% of your intervals will contain the true parameter value. When you apply the method once, you do not know if your interval gave a correct value (this happens 95% of the time) or not (this happens 5% of the time).
- **Confidence interval for the mean  $\mu$ .** For a Normal population with known standard deviation  $\sigma$ , a level  $C$  confidence interval for the mean  $\mu$  is given by  $\bar{x} \pm m$ , where the margin of error  $m = z^* \sigma / n$ . Here  $z^*$  is obtained from the standard Normal distribution such that the probability is  $C$  that a standard Normal random variable takes a value between  $-z^*$  and  $z^*$ .
- **Margin of error.** Other things being equal, the margin of error of a confidence interval decreases as the confidence level  $C$  decreases, the sample size  $n$  increases, and the population standard deviation  $\sigma$  decreases. The sample size  $n$  required to obtain a confidence interval of specified margin of error  $m$  for a Normal mean is  $n = (z^* \sigma / m)^2$ , where  $z^*$  is the critical point for the desired level of confidence.
- **A test of significance** is intended to assess the evidence provided by data against a null hypothesis  $H_0$  in favor of an alternative hypothesis  $H_a$ . The hypotheses are stated in terms of population parameters. Usually  $H_0$  is a statement that no effect or no difference is present, and  $H_a$  says that there is an effect or difference. The difference can be in a specific direction (one-sided alternative) or in either direction (two-sided alternative).
- **The test statistic and  $P$ -value.** The test of significance is based on a test statistic. The  $P$ -value is the probability, computed assuming that  $H_0$  is true, that the test statistic will take a value at least as extreme as that actually observed. Small  $P$ -values indicate strong evidence against  $H_0$ . Calculating  $P$ -values requires knowledge of the sampling distribution of the test statistic when  $H_0$  is true. If the  $P$ -value is as small or smaller than a specified value  $\alpha$ , the data are statistically significant at significance level  $\alpha$ .
- **Significance test concerning an unknown mean  $\mu$ .** Significance tests for the hypothesis  $H_0: \mu = \mu_0$  are based on the  $z$  statistic,  $z = (\bar{x} - \mu_0) / (\sigma / n)$ . This  $z$  test assumes an SRS of size  $n$ , known population standard deviation  $\sigma$ , and either a Normal population or a large sample.

## CHAPTER 7

---

- **Standard error.** When the standard deviation of a statistic is estimated from the data, the result is called the standard error of the statistic. The standard

error of the sample mean  $\bar{x}$  is  $SE_{\bar{x}} = \sigma/\sqrt{n}$ .

- **The  $t$  distributions.** Suppose that an SRS of size  $n$  is drawn from an  $N(\mu, \sigma)$  population. The one-sample  $t$  statistic  $t = (\bar{x} - \mu) / (s/\sqrt{n})$  has the  $t$  distribution with  $n - 1$  degrees of freedom.
- **The one-sample  $t$  confidence interval.** Consider an SRS of size  $n$  drawn from a population having unknown mean  $\mu$ . A level  $C$  confidence interval for  $\mu$  is  $\bar{x} \pm t^* s/\sqrt{n}$ , where  $t^*$  is the value for the  $t(n - 1)$  density curve with area  $C$  between  $-t^*$  and  $t^*$ . The quantity  $t^* s/\sqrt{n}$  is the margin of error. This interval is exact when the population distribution is Normal and is approximately correct for large  $n$  in other cases.
- **The one-sample  $t$  test.** Suppose that an SRS of size  $n$  is drawn from a population having unknown mean  $\mu$ . To test the hypothesis  $H_0: \mu = \mu_0$ , compute the one-sample  $t$  statistic  $t = (\bar{x} - \mu_0) / (s/\sqrt{n})$ .  $P$ -values or fixed significance levels are computed from the  $t(n - 1)$  distribution.
- **Matched pairs  $t$  procedures.** These one-sample procedures are used to analyze matched pairs data by first taking the differences within each matched pair to produce a single sample.
- **Robustness of  $t$  procedures.** The  $t$  procedures are relatively robust against non-Normal populations. The  $t$  procedures are useful for non-Normal data when  $15 \leq n < 40$  unless the data show outliers or strong skewness. When  $n \geq 40$ , the  $t$  procedures can be used even for clearly skewed distributions.
- **Power of a  $t$  test.** The power of the  $t$  test is calculated like that of the  $z$  test, using an approximate value for both  $\sigma$  and  $s$ .
- **Sign test.** The sign test is a distribution-free test because it uses probability calculations that are correct for a wide range of population distributions. The sign test for “no treatment effect” in matched pairs counts the number of positive differences. The  $P$ -value is computed from the  $B(n, 1/2)$  distribution, where  $n$  is the number of non-0 differences. The sign test is less powerful than the  $t$  test in cases where use of the  $t$  test is justified.
- **The two-sample  $t$  test.** Suppose that an SRS of size  $n_1$  is drawn from a Normal population with unknown mean  $\mu_1$  and that an independent SRS of size  $n_2$  is drawn from another Normal population with unknown mean  $\mu_2$ . To test the hypothesis  $H_0: \mu_1 = \mu_2$ , compute the two-sample  $t$  statistic  $t = (\bar{x}_1 - \bar{x}_2) / (s_p \sqrt{1/n_1 + 1/n_2})$  and use  $P$ -values or critical values for the  $t(k)$  distribution, where the degrees of freedom  $k$  either are approximated by software or are the smaller of  $n_1 - 1$  and  $n_2 - 1$ .

- **The two-sample  $t$  test.** Suppose that an SRS of size  $n_1$  is drawn from a Normal population with unknown mean  $\mu_1$  and that an independent SRS of size  $n_2$  is drawn from another Normal population with unknown mean  $\mu_2$ . The confidence interval for  $\mu_1 - \mu_2$  is given by  $(\bar{x}_1 - \bar{x}_2) \pm t^* s_p \sqrt{1/n_1 + 1/n_2}$ . This interval has confidence level at least  $C$  no matter what the population standard deviations may be. Here,  $t^*$  is the value for the  $t(k)$  density curve with area  $C$  between  $-t^*$  and  $t^*$ , where the degrees of freedom  $k$  either are approximated by software or are the smaller of  $n_1 - 1$  and  $n_2 - 1$ .
- **Pooled two-sample  $t$  procedures.** If we can assume that the two populations have equal variances, pooled two-sample  $t$  procedures can be used. These are based on the pooled estimator  $s_p^2 = ((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2) / (n_1 + n_2 - 2)$  of the unknown common variance and the  $t(n_1 + n_2 - 2)$  distribution.

## CHAPTER 8

---

- **Large-sample confidence interval for a population proportion.** Choose an SRS of size  $n$  from a large population with an unknown proportion  $p$  of successes. The sample proportion is  $\hat{p} = X/n$ , where  $X$  is the number of successes. The standard error of  $\hat{p}$  is  $SE_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n}$  and the margin of error for confidence level  $C$  is  $m = z^* SE_{\hat{p}}$ , where the critical value  $z^*$  is the value for the standard Normal density curve with area  $C$  between  $-z^*$  and  $z^*$ . An approximate level  $C$  confidence interval for  $p$  is  $\hat{p} \pm m$ . Use this interval for 90%, 95%, or 99% confidence when the number of successes and the number of failures are both at least 10.
- **Large-sample significance test for a population proportion.** Draw an SRS of size  $n$  from a large population with an unknown proportion  $p$  of successes. To test the hypothesis  $H_0: p = p_0$ , compute the  $z$  statistic,  $z = (\hat{p} - p_0) / \sqrt{p_0(1-p_0)/n}$ . In terms of a standard Normal random variable  $Z$ , the approximate  $P$ -value for a test of  $H_0$  against  $H_a: p > p_0$  is  $P(Z \geq z)$ ,  $H_a: p < p_0$  is  $P(Z \leq z)$ , and  $H_a: p \neq p_0$  is  $2P(Z \geq |z|)$ .
- **Sample size for desired margin of error.** The level  $C$  confidence interval for a proportion  $p$  will have a margin of error approximately equal to a specified value  $m$  when the sample size satisfies  $n = (z^*/m)^2 p^*(1-p^*)$ . Here  $z^*$  is the critical value for confidence  $C$ , and  $p^*$  is a guessed value for the proportion of successes in the future sample. The margin of error will be less than or equal to  $m$  if  $p^*$  is chosen to be 0.5. The sample size required when  $p^* = 0.5$  is  $n = (1/4)(z^*/m)^2$ .

- Large-sample confidence interval for comparing two proportions.** Choose an SRS of size  $n_1$  from a large population having proportion  $p_1$  of successes and an independent SRS of size  $n_2$  from another population having proportion  $p_2$  of successes. The estimate of the difference in the population proportions is  $D = \hat{p}_1 - \hat{p}_2$ . The standard error of  $D$  is  $SED = \sqrt{(\hat{p}_1(1-\hat{p}_1)/n_1) + (\hat{p}_2(1-\hat{p}_2)/n_2)}$  and the margin of error for confidence level  $C$  is  $m = z^* SE_D$ , where the critical value  $z^*$  is the value for the standard Normal density curve with area  $C$  between  $-z^*$  and  $z^*$ . An approximate level  $C$  confidence interval for  $p_1 - p_2$  is  $D \pm m$ . Use this method for 90%, 95%, or 99% confidence when the number of successes and the number of failures in each sample are both at least 10.
- Significance test for comparing two proportions.** To test the hypothesis  $H_0: p_1 = p_2$  compute the  $z$  statistic  $z = (\hat{p}_1 - \hat{p}_2) / SED_p$  where the pooled standard error is  $SED_p = \sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}$  and where  $\hat{p} = (X_1 + X_2) / (n_1 + n_2)$ . In terms of a standard Normal random variable  $Z$ , the  $P$ -value for a test of  $H_0$  against  $H_a: p_1 > p_2$  is  $P(Z \geq z)$ ,  $H_a: p_1 < p_2$  is  $P(Z \leq z)$ , and  $H_a: p_1 \neq p_2$  is  $2P(Z \geq |z|)$ .

## CHAPTER 9

---

- Chi-square statistic.** The chi-square statistic is a measure of how much the observed cell counts in a two-way table diverge from the expected cell counts. The formula for the statistic is

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

where “observed” represents an observed cell count, “expected” represents the expected count for the same cell, and the sum is over all  $r \times c$  cells in the table.

- Chi-square test for two-way tables.** The null hypothesis  $H_0$  is that there is no association between the row and column variables in a two-way table. The alternative is that these variables are related. If  $H_0$  is true, the chi-square statistic  $\chi^2$  has approximately a  $\chi^2$  distribution with  $(r - 1)(c - 1)$  degrees of freedom. The  $P$ -value for the chi-square test is  $P(\chi^2 \geq X^2)$ , where  $\chi^2$  is a random variable having the  $\chi^2$ (df) distribution with  $df = (r - 1)(c - 1)$ .
- Expected cell counts.** Expected count = (row total  $\times$  column total) /  $n$ .
- The chi-square goodness of fit test.** Data for  $n$  observations of a categorical variable with  $n$  possible outcomes are summarized as observed counts,  $n_1, n_2,$

... ,  $n_k$ , in  $k$  cells. A null hypothesis specifies probabilities  $p_1, p_2, \dots, p_k$  for the possible outcomes. For each cell, multiply the total number of observations  $n$  by the specified probability to determine the expected counts: expected count =  $np_i$ . The chi-square statistic measures how much the observed cell counts differ from the expected cell counts. The formula for the statistic is

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

The degrees of freedom are  $k - 1$ , and  $P$ -values are computed from the chi-square distribution.

## CHAPTER 10

---

- **Simple linear regression.** The statistical model for simple linear regression assumes the means of the response variable  $y$  fall on a line when plotted against  $x$ , with the observed  $y$ 's varying Normally about these means. For  $n$  observations, this model can be written  $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ , where  $i = 1, 2, \dots, n$ , and the  $\varepsilon_i$  are assumed to be independent and Normally distributed with mean 0 and standard deviation  $\sigma$ . Here  $\beta_0 + \beta_1 x_i$  is the mean response when  $x = x_i$ . The parameters of the model are  $\beta_0, \beta_1$ , and  $\sigma$ .
- **Estimation of model parameters.** The population regression line intercept and slope,  $\beta_0$  and  $\beta_1$ , are estimated by the intercept and slope of the least-squares regression line,  $b_0$  and  $b_1$ . The parameter  $\sigma$  is estimated by  $s = \sqrt{\sum e_i^2 / (n - 2)}$ , where the  $e_i$  are the residuals  $e_i = y_i - \hat{y}_i$ .
- **Confidence interval and significance test for  $\beta_1$ .** A level  $C$  confidence interval for population slope  $\beta_1$  is  $b_1 \pm t^* SE_{b_1}$  where  $t^*$  is the value for the  $t(n - 2)$  density curve with area  $C$  between  $-t^*$  and  $t^*$ . The test of the hypothesis  $H_0: \beta_1 = 0$  is based on the  $t$  statistic  $t = b_1 / SE_{b_1}$  and the  $t(n - 2)$  distribution. This tests whether there is a straight-line relationship between  $y$  and  $x$ . There are similar formulas for confidence intervals and tests for  $\beta_0$ , but these are meaningful only in special cases.
- **Confidence interval for the mean response.** The estimated mean response for the subpopulation corresponding to the value  $x^*$  of the explanatory variable is  $\mu_{\hat{y}} = b_0 + b_1 x^*$ . A level  $C$  confidence interval for the mean response is  $\mu_{\hat{y}} \pm t^* SE_{\mu_{\hat{y}}}$  where  $t^*$  is the value for the  $t(n - 2)$  density curve with area  $C$  between  $-t^*$  and  $t^*$ .
- **Prediction interval for the estimated response.** The estimated value of the

response variable  $y$  for a future observation from the subpopulation corresponding to the value  $x^*$  of the explanatory variable is  $\hat{y} = b_0 + b_1 x^*$ . A level  $C$  prediction interval for the estimated response is  $\hat{y} \pm t^* SE_{\hat{y}}$  where  $t^*$  is the value for the  $t(n - 2)$  density curve with area  $C$  between  $-t^*$  and  $t^*$ . The standard error for the prediction interval is larger than that for the confidence interval because it also includes the variability of the future observation around its subpopulation mean.

## CHAPTER 11

---

- Multiple linear regression.** The statistical model for multiple linear regression with response variable  $y$  and  $p$  explanatory variables  $x_1, x_2, \dots, x_p$  is  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i$  where  $i = 1, 2, \dots, n$ . The  $\varepsilon_i$  are assumed to be independent and Normally distributed with mean 0 and standard deviation  $\sigma$ . The parameters of the model are  $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ , and  $\sigma$
- Estimation of model parameters.** The multiple regression equation predicts the response variable by a linear relationship with all the explanatory variables:  $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p$ . The  $\beta$ 's are estimated by  $b_0, b_1, b_2, \dots, b_p$ , which are obtained by the method of least squares. The parameter  $\sigma$  is estimated by  $s = \text{MSE} = \sum e_i^2 / (n - p - 1)$  where the  $e_i$  are the residuals,  $e_i = y_i - \hat{y}_i$ .
- Confidence interval for  $\beta_j$ .** A level  $C$  confidence interval for  $\beta_j$  is  $b_j \pm t^* SE_{b_j}$  where  $t^*$  is the value for the  $t(n - p - 1)$  density curve with area  $C$  between  $-t^*$  and  $t^*$ . The test of the hypothesis  $H_0: \beta_j = 0$  is based on the  $t$  statistic  $t = b_j / SE_{b_j}$  and the  $t(n - p - 1)$  distribution. The estimate  $b_j$  of  $\beta_j$  and the test and confidence interval for  $\beta_j$  are all based on a specific multiple linear regression model. The results of all of these procedures change if other explanatory variables are added to or deleted from the model.
- The ANOVA  $F$  test.** The ANOVA table for a multiple linear regression gives the degrees of freedom, sum of squares, and mean squares for the model, error, and total sources of variation. The ANOVA  $F$  statistic is the ratio  $\text{MSM} / \text{MSE}$  and is used to test the null hypothesis  $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$ . If  $H_0$  is true, this statistic has an  $F(p, n - p - 1)$  distribution.
- Squared multiple correlation.** The squared multiple correlation is given by the expression  $R^2 = \text{SSM} / \text{SST}$  and is interpreted as the proportion of the variability in the response variable  $y$  that is explained by the explanatory variables  $x_1, x_2, \dots, x_p$  in the multiple linear regression.

## CHAPTER 12

---

- **One-way analysis of variance (ANOVA)** is used to compare several population means based on independent SRSs from each population. The populations are assumed to be Normal with possibly different means and the same standard deviation. To do an analysis of variance, first compute sample means and standard deviations for all groups. Side-by-side boxplots give an overview of the data. Examine Normal quantile plots (either for each group separately or for the residuals) to detect outliers or extreme deviations from Normality. Compute the ratio of the largest to the smallest sample standard deviation. If this ratio is less than 2 and the Normal quantile plots are satisfactory, ANOVA can be performed.
- **ANOVA  $F$  test.** An analysis of variance table organizes the ANOVA calculations. Degrees of freedom, sums of squares, and mean squares appear in the table. The  $F$  statistic is the ratio  $MSG/MSE$  and is used to test the null hypothesis that the population means are *all* equal. The alternative hypothesis is true if there are *any* differences among the population means. The  $F(I - 1, N - I)$  distribution is used to compute the  $P$ -value.
- **Contrasts.** Specific questions formulated before examination of the data can be expressed as contrasts. A contrast is a combination of population means of the form  $\psi = \sum a_i \mu_i$  where the coefficients  $a_i$  sum to 0. The corresponding sample contrast is  $c = \sum a_i \bar{x}_i$ . The standard error of  $c$  is  $SE_c = sp \sqrt{\sum a_i^2 / n_i}$ . Tests and confidence intervals for contrasts provide answers to these specific questions.
- **Multiple comparisons.** To perform a multiple-comparisons procedure, compute  $t$  statistics for all pairs of means using the formula  $t_{ij} = (\bar{x}_i - \bar{x}_j) / (sp \sqrt{1/n_i + 1/n_j})$ . If  $|t_{ij}| \geq t^{**}$  we declare that the population means  $\mu_i$  and  $\mu_j$  are different. Otherwise, we conclude that the data do not distinguish between them. The value of  $t^{**}$  depends upon which multiple-comparisons procedure we choose.

## CHAPTER 13

---

- **Two-way analysis of variance** is used to compare population means when populations are classified according to two factors. ANOVA assumes that the populations are Normal with possibly different means and the same standard deviation and that independent SRSs are drawn from each population. As with one-way ANOVA, preliminary analysis includes examination of means, standard deviations, and Normal quantile plots.

- **ANOVA table and F tests.** ANOVA separates the total variation into parts for the model and error. The model variation is separated into parts for each of the main effects and the interaction. These calculations are organized into an ANOVA table. Pooling is used to estimate the within-group variance.  $F$  statistics and  $P$ -values are used to test hypotheses about the main effects and the interaction.
- **Marginal means** are calculated by taking averages of the cell means across rows and columns. Careful inspection of the cell means is necessary to interpret statistically significant main effects and interactions. Plots are a useful aid.

# NOTES AND DATA SOURCES

## CHAPTER 1

1. See **census.gov**.
2. From *State of Drunk Driving Fatalities in America 2010*, available at **centurycouncil.org**.
3. James P. Purdy, “Why first-year college students select online research sources as their favorite,” *First Monday*, 17, No. 9 (September 3, 2012). See **firstmonday.org**.
4. Data collected in the lab of Connie Weaver, Department of Foods and Nutrition, Purdue University, and provided by Linda McCabe.
5. Haipeng Shen, “Nonparametric regression for problems involving lognormal distributions,” PhD dissertation, University of Pennsylvania, 2003. Thanks to Haipeng Shen and Larry Brown for sharing the data.
6. From the Digest of Education Statistics at the website of the National Center for Education Statistics, **nces.ed.gov/programs/digest**.
7. See Note 4.
8. Based on Barbara Ernst et al., “Seasonal variation in the deficiency of 25-hydroxyvitamin D<sub>3</sub> in mildly to extremely obese subjects,” *Obesity Surgery*, 19 (2009), pp. 180–183.
9. More information about the *Titanic* can be found at the website for the Titanic Project in Belfast, Ireland, at **titanicbelfast.com/Home.aspx**.
10. Data describing the passengers on the *Titanic* can be found at **lib.stat.cmu.edu/S/Harrell/data/descriptions/titanic.html**.
11. See **semiocast.com/publications/2012\_01\_31\_Brazil\_becomes\_2nd\_country\_on\_T**
12. Data for 2011 from Table 1.1 in the U.S. Energy Information Administration’s *December 2012 Monthly Energy Review*, available at **eia.gov/totalenergy/data/monthly/pdf/mer.pdf**.
13. From the Color Assignment website of Joe Hallock, **joehallock.com/edu/COM498/index.html**.
14. U.S. Environmental Protection Agency, *Municipal Solid Waste Generation, Recycling, and Disposal in the United States: Tables and Figures for 2010*.

15. November 2012 report from [marketshare.hitslink.com](http://marketshare.hitslink.com).
16. Color popularity for 2011 from the Dupont Automotive Color report; see [dupont.com/Media\\_center/en\\_US/color\\_popularity](http://dupont.com/Media_center/en_US/color_popularity).
17. Data for November 2012, from [internetworldstats.com/facebook.htm](http://internetworldstats.com/facebook.htm).
18. See previous note.
19. Data provided by Darlene Gordon, Purdue University.
20. Data for 1980 to 2012 are available from the World Bank at [data.worldbank.org/indicator/IC.REG.DURS](http://data.worldbank.org/indicator/IC.REG.DURS). Data for 2012 were used for this example.
21. See, for example, <http://www.nacubo.org/Research>.
22. The data were provided by James Kaufman. The study is described in James C. Kaufman, "The cost of the muse: Poets die young," *Death Studies*, 27 (2003), pp. 813–821. The quote from Yeats appears in this article.
23. See, for example, the bibliographic entry for Gosset in the School of Mathematics and Statistics of the University of St. Andrews, Scotland, MacTutor History of Mathematics archive at [www.history.mcs.st-andrews.ac.uk/Biographies/Gosset.html](http://www.history.mcs.st-andrews.ac.uk/Biographies/Gosset.html).
24. These and other data that were collected and used by Gosset can be found in the Guinness Archives in Dublin. See [guinness-storehouse.com/en/Archive.aspx](http://guinness-storehouse.com/en/Archive.aspx).
25. These data were provided by Krista Nichols, Department of Biological Sciences, Purdue University.
26. From the Interbrand website; see [interbrand.com/en/best-global-brands](http://interbrand.com/en/best-global-brands).
27. From [beer100.com/beercalories.htm](http://beer100.com/beercalories.htm) on January 4, 2013.
28. See Noel Cressie, *Statistics for Spatial Data*, Wiley, 1993.
29. Data provided by Francisco Rosales of the Department of Nutritional Sciences, Pennsylvania State University.
30. Data provided by Betsy Hoza, Department of Psychological Sciences, University of Vermont.
31. Net worth for 2010 from the *Federal Reserve Bulletin*, 98, No. 2 (2012), p. 17.
32. For more information about earthquakes, see the U.S. Geological Service website at [usgs.gov](http://usgs.gov).
33. We thank Ethan J. Temeles of Amherst College for providing the data. His work is described in Ethan J. Temeles and W. John Kress, "Adaptation in a plant-hummingbird association," *Science*, 300 (2003), pp. 630–633.

34. The National Assessment of Educational Progress (NAEP) is conducted by the National Center for Education Statistics (NCES). The NAEP is a large assessment of student knowledge in a variety of subjects. See [nces.ed.gov/nationsreportcard/naepdata](http://nces.ed.gov/nationsreportcard/naepdata).
35. See the NCAA Eligibility Center Quick Reference Sheet, available at [fs.ncaa.org/Docs/eligibility\\_center/Quick\\_Reference\\_Sheet.pdf](http://fs.ncaa.org/Docs/eligibility_center/Quick_Reference_Sheet.pdf).
36. Distributions for SAT scores can be found at the College Board website, [research.collegeboard.org/content/sat-data-tables](http://research.collegeboard.org/content/sat-data-tables).
37. See previous note.
38. See [stubhub.com](http://stubhub.com).
39. From Matthias R. Mehl et al., “Are women really more talkative than men?,” *Science*, 317, No. 5834 (2007), p. 82. The raw data were provided by Matthias Mehl.
40. From the American Heart Association website, [americanheart.org](http://americanheart.org).
41. From [fueleconomy.gov](http://fueleconomy.gov).
42. From [cdc.gov/brfss](http://cdc.gov/brfss). The data were collected in 2011, with the exception of the fruits and vegetables variable, which is from 2009, the most recent year when this variable was included in the survey.
43. See Note 16.
44. See [worldbank.org](http://worldbank.org). These data are among the files available under “Data,” “Indicators.”
45. Data for 2013 were downloaded from [isp-review.toptenreviews.com](http://isp-review.toptenreviews.com).
46. See previous note.
47. The Institute of Medicine website, [www.iom.edu](http://www.iom.edu), provides links to reports related to dietary reference intakes as well as other health and nutrition topics.
48. *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium and Carotenoids*, National Academy of Sciences, 2000.
49. See previous note.

## Chapter 2

1. Hannah G. Lund et al., “Sleep patterns and predictors of disturbed sleep in a large population of college students,” *Adolescent Health*, 46, No. 2 (2010), pp. 97–99.
2. See previous note.
3. See [cfs.purdue.edu/FN/campcalcium/public.htm](http://cfs.purdue.edu/FN/campcalcium/public.htm) for information about the

2010 camp.

4. See [consumersunion.org/about](http://consumersunion.org/about).

5. “Best laundry detergents,” *Consumer Reports*, November 2011, pp. 8–9.

6. OECD StatExtracts, Organisation for Economic Co-operation and Development, downloaded on January 8, 2013, from [stats.oecd.org/wbos](http://stats.oecd.org/wbos).

7. These studies were conducted by Connie Weaver, Department of Nutrition Science, Purdue University, over the past 20 years. The data for this example were provided by Linda McCabe. More details concerning this particular study and references to other related studies are given in Lu Wu, “Calcium requirements and metabolism in Chinese-American boys and girls,” *Journal of Bone Mineral Research*, 25, No. 8 (2010), pp. 1842–1849.

8. A sophisticated treatment of improvements and additions to scatterplots is W. S. Cleveland and R. McGill, “The many faces of a scatterplot,” *Journal of the American Statistical Association*, 79 (1984), pp. 807–822.

9. Stewart Warden et al., “Throwing induces substantial torsional adaption within the midshaft humerus of male baseball players,” *Bone*, 45 (2009), pp. 931–941. The data were provided by Stewart Warden, Department of Physical Therapy, School of Health and Rehabilitation Sciences, Indiana University.

10. See [spectrumtechniques.com/isotope\\_generator.htm](http://spectrumtechniques.com/isotope_generator.htm).

11. These data were collected under the supervision of Zach Grigsby, Science Express Coordinator, College of Science, Purdue University.

12. See [beer100.com/beercalories.htm](http://beer100.com/beercalories.htm).

13. See [worldbank.org](http://worldbank.org).

14. James T. Fleming, “The measurement of children’s perception of difficulty in reading materials,” *Research in the Teaching of English*, 1 (1967), pp. 136–156.

15. Data for 2012 from [forbes.com/nfl-valuations/](http://forbes.com/nfl-valuations/).

16. From [en.wikipedia.org/wiki/10000\\_metres](http://en.wikipedia.org/wiki/10000_metres).

17. A careful study of this phenomenon is W. S. Cleveland, P. Diaconis, and R. McGill, “Variables on scatterplots look more highly correlated when the scales are increased,” *Science*, 216 (1982), pp. 1138–1141.

18. Data from a plot in James A. Levine, Norman L. Eberhardt, and Michael D. Jensen, “Role of nonexercise activity thermogenesis in resistance to fat gain in humans,” *Science*, 283 (1999), pp. 212–214.

19. Frank J. Anscombe, “Graphs in statistical analysis,” *American Statistician*, 27 (1973), pp. 17–21.

20. From the website of the National Center for Education Statistics,

**nces.ed.gov.**

21. Debora L. Arsenau, “Comparison of diet management instruction for patients with non-insulin dependent diabetes mellitus: Learning activity package vs. group instruction,” master’s thesis, Purdue University, 1993.
22. The facts in Exercise 2.100 come from Nancy W. Burton and Leonard Ramist, *Predicting Success in College: Classes Graduating since 1980*, Research Report No. 2001-2, The College Board, 2001.
23. See Note 19.
24. See **iom.edu**.
25. Based on a study described in Corby C. Martin et al., “Children in school cafeterias select foods containing more saturated fat and energy than the Institute of Medicine recommendations,” *Journal of Nutrition*, 140 (2010), pp. 1653–1660.
26. You can find a clear and comprehensive discussion of numerical measures of association for categorical data in Chapter 2 of Alan Agresti, *Categorical Data Analysis*, 2nd ed., Wiley, 2002.
27. Edward Bumgardner, “Loss of teeth as a disqualification for military service,” *Transactions of the Kansas Academy of Science*, 18 (1903), pp. 217–219.
28. Based on reports prepared by Andy Zehner, vice president for Student Affairs, Purdue University.
29. Data are from the NOAA Satellite and Information Service at **ncdc.noaa.gov/special-reports/groundhog-day.php**.
30. From M.-Y. Chen et al., “Adequate sleep among adolescents is positively associated with health status and health-related behaviors,” *BMC Public Health*, 6, No. 59 (2006); available from **biomedicalcentral.com/1471-2458/6/59**.
31. M. S. Linet et al., “Residential exposure to magnetic fields and acute lymphoblastic leukemia in children,” *New England Journal of Medicine*, 337 (1997), pp. 1–7.
32. *The Health Consequences of Smoking: 1983*, U.S. Public Health Service, 1983.
33. Dennis Bristow et al., “Thirty games out and sold out for months! An empirical examination of fan loyalty to two Major League Baseball teams,” *Journal of Management Research*, 2, No. 1 (2010), E2; available at **macrothink.org/jmr**.
34. See **www12.statcan.ca/english/census06/analysis/agesex/ProvTerr1.cfm**.

35. OECD StatExtracts, Organisation for Economic Co-operation and Development, downloaded on June 29, 2008, from [stats.oecd.org/wbos](http://stats.oecd.org/wbos).
36. For an overview of remote deposit capture, see [remotedepositcapture.com/overview/rdc.overview.aspx](http://remotedepositcapture.com/overview/rdc.overview.aspx).
37. From the “Community Bank Competitiveness Survey,” 2008, *ABA Banking Journal*. The survey is available at [nxtbook.com/nxtbooks/sb/ababj-compsurv08/index.php](http://nxtbook.com/nxtbooks/sb/ababj-compsurv08/index.php).
38. The counts reported were calculated using counts of the numbers of banks in the different regions and the percents given in the ABA report.
39. *Education Indicators: An International Perspective*, Institute of Education Studies, National Center for Education Statistics; see [nces.ed.gov/surveys/international](http://nces.ed.gov/surveys/international).
40. Information about this procedure was provided by Samuel Flanigan of *U.S. News & World Report*. See [usnews.com/usnews/rankguide/rghome.htm](http://usnews.com/usnews/rankguide/rghome.htm) for a description of the variables used to construct the ranks and for the most recent ranks.
41. From the Social Security website, [ssa.gov/OACT/babynames](http://ssa.gov/OACT/babynames).
42. See [cdc.gov/brfss/](http://cdc.gov/brfss/). The data file BRFSS contains several variables from this source.
43. We thank Zhiyong Cai of Texas A&M University for providing the data. The data are from work performed in connection with his PhD dissertation in the Department of Forestry and Natural Resources, Purdue University.
44. Although these data are fictitious, similar though less simple situations occur. See P. J. Bickel and J. W. O’Connell, “Is there a sex bias in graduate admissions?,” *Science*, 187 (1975), pp. 398–404.

### Chapter 3

1. See [norc.uchicago.edu](http://norc.uchicago.edu).
2. Stewart Warden et al., “Throwing induces substantial torsional adaption within the midshaft humerus of male baseball players,” *Bone*, 45 (2009), pp. 931–941.
3. Corby C. Martin et al., “Children in school cafeterias select foods containing more saturated fat and energy than the Institute of Medicine recommendations,” *Journal of Nutrition*, 140 (2010), pp. 1653–1660.
4. Based on “Look, no hands: Automatic soap dispensers,” *Consumer Reports*, February 2013, p. 11.
5. From “Did you know,” *Consumer Reports*, February 2013, p. 10.

6. Bruce Barrett et al., “Echinacea for treating the common cold,” *Annals of Internal Medicine*,153 (2010), pp. 769–777.
7. For a full description of the STAR program and its follow-up studies, go to **heros-inc.org/star.htm**.
8. See Note 6.
9. Bonnie Spring et al., “Multiple behavior changes in diet and activity,” *Archives of Internal Medicine*,172, No. 10 (2012), pp. 789–796.
10. Based on Gerardo Ramirez and Sian L. Beilock, “Writing about testing worries boosts exam performance in the classroom,” *Science*,331 (2011), p. 2011. Although we describe the experiment as not including a control group, the researchers who conducted this study did, in fact, use one.
11. A general discussion of failures of blinding is Dean Ferguson et al., “Turning a blind eye: The success of blinding reported in a random sample of randomised, placebo controlled trials,” *British Medical Journal*,328 (2004), p. 432.
12. Based on a study conducted by Sandra Simonis under the direction of Professor Jon Harbor from the Purdue University Department of Earth, Atmospheric Sciences, and Planetary.
13. Based on a study conducted by Tammy Younts directed by Professor Deb Bennett of the Purdue University Department of Educational Studies. For more information about Reading Recovery, see **readingrecovery.org/**.
14. Based on a study conducted by Rajendra Chaini under the direction of Professor Bill Hoover of the Purdue University Department of Forestry and Natural Resources.
15. From the Hot Ringtones list at **billboard.com/**on January 28, 2013.
16. From the Rock Songs list at **billboard.com/**on January 28, 2013.
17. From the online version of the Bureau of Labor Statistics, *Handbook of Methods*,modified April 17, 2003, at **bls.gov**.The details of the design are more complicated than we describe.
18. For more detail on the material of this section and complete references, see P. E. Converse and M. W. Traugott, “Assessing the accuracy of polls and surveys,” *Science*,234 (1986), pp. 1094–1098.
19. From **census.gov/cps/methodology/nonresponse.html**on January 29, 2013.
20. From **www3.norc.Org/GSSWebsite/FAQs/**on January 29, 2013.
21. See **pewresearch.org/about**.
22. See “Assessing the representativeness of public opinion surveys,” May 15,

2012, from [people-press.org/2012/05/15](http://people-press.org/2012/05/15).

23. Sex: Tom W. Smith, “The *JAMA* controversy and the meaning of sex,” *Public Opinion Quarterly*, 63 (1999), pp. 385–400. Welfare: from a *New York Times*/CBS News Poll reported in the *New York Times*, July 5, 1992. Scotland: “All set for independence?” *Economist*, September 12, 1998. Many other examples appear in T. W. Smith, “That which we call welfare by any other name would smell sweeter,” *Public Opinion Quarterly*, 51 (1987), pp. 75–83.

24. From [gallup.com](http://gallup.com) on November 10, 2009.

25. From [pewresearch.org](http://pewresearch.org) on November 10, 2009.

26. From [thefuturescompany.com](http://thefuturescompany.com) on January 29, 2013.

27. From [aauw.org/act/laf/library/harassment\\_stats.cf](http://aauw.org/act/laf/library/harassment_stats.cf) on January 30, 2013.

28. John C. Bailar III, “The real threats to the integrity of science,” *Chronicle of Higher Education*, April 21, 1995, pp. B1–B2.

29. The difficulties of interpreting guidelines for informed consent and for the work of institutional review boards in medical research are a main theme of Beverly Woodward, “Challenges to human subject protections in U.S. medical research,” *Journal of the American Medical Association*, 282 (1999), pp. 1947–1952. The references in this paper point to other discussions.

30. Quotation from the *Report of the Tuskegee Syphilis Study Legacy Committee*, May 20, 1996. A detailed history is James H. Jones, *Bad Blood: The Tuskegee Syphilis Experiment*, Free Press, 1993.

31. Dr. Hennekens’s words are from an interview in the Annenberg/Corporation for Public Broadcasting video series *Against All Odds: Inside Statistics*.

32. See [ftc.gov/opa/2009/04/kellogg.shtm](http://ftc.gov/opa/2009/04/kellogg.shtm).

33. On February 12, 2012, the CBS show *60 Minutes* reported the latest news on this study, which was published in the *Journal of Clinical Oncology* in 2007. See [cbsnews.com/video/watch/?id=7398476n](http://cbsnews.com/video/watch/?id=7398476n).

34. From Randi Zlotnik Shaul et al., “Legal liabilities in research: Early lessons from North America,” *BMJ Medical Ethics*, 6, No. 4 (2005), pp. 1–4.

35. See previous note.

36. The report was issued in February 2009 and is available from [ftc.gov/os/2009/02/P085400behavadreport.pdf](http://ftc.gov/os/2009/02/P085400behavadreport.pdf).

## Chapter 4

1. An informative and entertaining account of the origins of probability theory is Florence N. David, *Games, Gods and Gambling*, Charles Griffin, London, 1962.
2. Color popularity for 2011 from the Dupont Automotive Color report; see [http://dupont.com/Media\\_center/en\\_US/color\\_popularity](http://dupont.com/Media_center/en_US/color_popularity).
3. You can find a mathematical explanation of Benford's law in Ted Hill, "The first-digit phenomenon," *American Scientist*, 86 (1996), pp. 358–363; and Ted Hill, "The difficulty of faking data," *Chance*, 12, No. 3 (1999), pp. 27–31. Applications in fraud detection are discussed in the second paper by Hill and in Mark A. Nigrini, "I've got your number," *Journal of Accountancy*, May 1999, available online at [aicpa.org/pubs/jofa/joaiss.htm](http://aicpa.org/pubs/jofa/joaiss.htm).
4. Royal Statistical Society news release, "Royal Statistical Society concerned by issues raised in Sally Clark case," October 23, 2001, at [www.rss.org.uk](http://www.rss.org.uk). For background, see an editorial and article in the *Economist*, January 22, 2004. The editorial is entitled "The probability of injustice."
5. See [cdc.gov/mmwr/preview/mmwrhtml/mm57e618a1.htm](http://cdc.gov/mmwr/preview/mmwrhtml/mm57e618a1.htm).
6. See the previous note.
7. From [funtonia.com/top\\_ringtones\\_chart.asp](http://funtonia.com/top_ringtones_chart.asp). This website gives popularity scores based on download activity on the Internet. These scores were converted to probabilities for this exercise by dividing each popularity score by the sum of the scores for the top ten ringtones.
8. See [bloodbook.com/world-abo.html](http://bloodbook.com/world-abo.html) for the distribution of blood types for various groups of people.
9. From Statistics Canada, [www.statcan.ca](http://www.statcan.ca).
10. We use  $\bar{x}$  both for the random variable, which takes different values in repeated sampling, and for the numerical value of the random variable in a particular sample. Similarly,  $s$  and  $\hat{p}$  stand both for random variables and for specific values. This notation is mathematically imprecise but statistically convenient.
11. We will consider only the case in which  $X$  takes a finite number of possible values. The same ideas, implemented with more advanced mathematics, apply to random variables with an infinite but still countable collection of values.
12. Based on a Pew Internet report, "Teens and distracted driving," available from [pewinternet.org/Reports/2009/Teens-and-Distracted-Driving.aspx](http://pewinternet.org/Reports/2009/Teens-and-Distracted-Driving.aspx).
13. See [pewinternet.org/Reports/2009/17-Twitter-and-Status-Updating-Fall-2009.aspx](http://pewinternet.org/Reports/2009/17-Twitter-and-Status-Updating-Fall-2009.aspx).
14. The mean of a continuous random variable  $X$  with density function  $f(x)$  can be found by integration:

$$\mu_X = \int xf(x) dx$$

This integral is a kind of weighted average, analogous to the discrete-case mean

$$\mu_X = \sum xP(X=x)$$

The variance of a continuous random variable  $X$  is the average squared deviation of the values of  $X$  from their mean, found by the integral

$$\sigma_X^2 = \int (x-\mu)^2 f(x) dx$$

15. See A. Tversky and D. Kahneman, "Belief in the law of small numbers," *Psychological Bulletin*, 76 (1971), pp. 105–110, and other writings of these authors for a full account of our misperception of randomness.

16. Probabilities involving runs can be quite difficult to compute. That the probability of a run of three or more heads in 10 independent tosses of a fair coin is  $(1/2) + (1/128) = 0.508$  can be found by clever counting. A general treatment using advanced methods appears in Section XIII.7 of William Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd ed., Wiley, 1968.

17. R. Vallone and A. Tversky, "The hot hand in basketball: On the misperception of random sequences," *Cognitive Psychology*, 17 (1985), pp. 295–314. A later series of articles that debate the independence question is A. Tversky and T. Gilovich, "The cold facts about the 'hot hand' in basketball," *Chance*, 2, No. 1 (1989), pp. 16–21; P. D. Larkey, R. A. Smith, and J. B. Kadane, "It's OK to believe in the 'hot hand,'" *Chance*, 2, No. 4 (1989), pp. 22–30; and A. Tversky and T. Gilovich, "The 'hot hand': Statistical reality or cognitive illusion?" *Chance*, 2, No. 4 (1989), pp. 31–34.

18. Based on a study discussed in S. Atkinson, G. McCabe, C. Weaver, S. Abrams, and K O'Brien, "Are current calcium recommendations for adolescents higher than needed to achieve optimal peak bone mass? The controversy," *Journal of Nutrition*, 138, No. 6 (2008), pp. 1182–1186.

19. Based on a study described in Corby C. Martin et al., "Children in school cafeterias select foods containing more saturated fat and energy than the Institute of Medicine recommendations," *Journal of Nutrition*, 140 (2010), pp. 1653–1660.

20. Based on *The Ethics of American Youth--2008*, available from the Josephson Institute, [charactercounts.org/programs/reportcard](http://charactercounts.org/programs/reportcard).

21. See [nces.ed.gov/programs/digest](http://nces.ed.gov/programs/digest). Data are from the 2012 *Digest of Education Statistics*.

22. From the 2012 *Statistical Abstract of the United States*, Table 299.

23. *Ibid.*, Table 278.

## Chapter 5

1. K. M. Orzech et al., “The state of sleep among college students at a large public university,” *Journal of American College Health*, 59 (2011), pp. 612–619.
2. The description of the 2011 survey and results can be found at **[blog.appsfire.com/infographic-ios-apps-vs-web-apps](http://blog.appsfire.com/infographic-ios-apps-vs-web-apps)**.
3. Haipeng Shen, “Nonparametric regression for problems involving lognormal distributions,” PhD dissertation, University of Pennsylvania, 2003. Thanks to Haipeng Shen and Larry Brown for sharing the data.
4. Findings are from the *Time* Mobility Poll run between June 29 and July 28, 2012. The results were published in the August 27, 2012, issue of *Time*.
5. Statistical methods for dealing with time-to-failure data, including the Weibull model, are presented in Wayne Nelson, *Applied Life Data Analysis*, Wiley, 1982.
6. Findings are from Nielsen’s “State of the Appnation— a year of change and growth in U.S. Smartphones,” posted May 16, 2012, on **[blog.nielsen.com/nielsenwire/](http://blog.nielsen.com/nielsenwire/)**.
7. Statistics regarding Facebook usage can be found at **[www.facebook.com/notes/facebook-data-team/anatomy-of-facebook/10150388519243859](http://www.facebook.com/notes/facebook-data-team/anatomy-of-facebook/10150388519243859)**.
8. From the grade distribution database of the Indiana University Office of the Registrar, **[gradedistribution.registrar.indiana.edu](http://gradedistribution.registrar.indiana.edu)**.
9. Karel Kleisner et al., “Trustworthy-looking face meets brown eyes,” *PLoS ONE*, 8, No. 1 (2013), e53285, doi:10.1371/journal.pone.0053285.
10. Diane M. Dellavalle and Jere D. Haas, “Iron status is associated with endurance performance and training in female rowers,” *Medicine and Science in Sports and Exercise*, 44, No. 8 (2012), pp. 1552–1559.
11. Results of this and other questions from this 2011 survey can be found at **<http://www.mumsnet.com/surveys/pressure-on-children-and-parents>**.
12. *Crossing the Line: Sexual Harassment at School*, a report from the American Association of University Women Educational Foundation published in 2011. See **[www.aauw.org/](http://www.aauw.org/)**.
13. S. A. Rahimtoola, “Outcomes 15 years after valve replacement with a mechanical vs. a prosthetic valve: Final report of the Veterans Administration randomized trial,” American College of Cardiology, **[www.acc.org/education/online/trials/acc2000/15yr.htm](http://www.acc.org/education/online/trials/acc2000/15yr.htm)**.
14. The full online clothing store ratings are featured in the December 2008

issue of *Consumer Reports* and online at [www.ConsumerReports.org](http://www.ConsumerReports.org).

15. The results of this 2012 survey can be found at [www.theaa.com/newsroom/news-2012/streetwatch-october-2012-fewer-potholes.html](http://www.theaa.com/newsroom/news-2012/streetwatch-october-2012-fewer-potholes.html).

16. The results of this 2012 survey can be found at [josephsoninstitute.org](http://josephsoninstitute.org).

17. A description and summary of this 2012 survey can be found at [www.ipsos-na.com/news-polls/pressrelease.aspx?id=5537](http://www.ipsos-na.com/news-polls/pressrelease.aspx?id=5537).

18. This 2011 survey was performed by Ipsos MediaCT right before a new Copyright Amendment Act went into effect in New Zealand. Results of this survey can be found at [www.copyright.co.nz/News/2195/](http://www.copyright.co.nz/News/2195/).

19. Lydia Saad, “Americans’ preference for smaller families edges higher,” Gallup Poll press release, June 30, 2011, [www.gallup.com](http://www.gallup.com).

20. A summary of Larry Wright’s study can be found at [www.nytimes.com/2009/03/04/sports/basketball/04freethrow.html](http://www.nytimes.com/2009/03/04/sports/basketball/04freethrow.html).

21. Barbara Means et al., “Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies,” U.S. Department of Education, Office of Planning, Evaluation, and Policy Development, 2010.

22. Dafna Kanny et al., “Vital signs: Binge drinking among women and high school girls—United States, 2011,” *Morbidity and Mortality Weekly Report*, January 8, 2013.

23. Information was obtained from “Price comparisons of wireline, wireless and internet services in Canada and with foreign jurisdictions,” Canadian Radio-Television and Telecommunications Commission, April 6, 2012.

24. This information can be found at [www.census.gov/genealogy/names/dist.all.last](http://www.census.gov/genealogy/names/dist.all.last).

## Chapter 6

1. Noel Cressie, *Statistics for Spatial Data*, Wiley, 1993. The significance test result that we report is one of several that could be used to address this question. See pp. 607–609 of the Cressie book for more details.

2. The 2010–2011 statistics for California were obtained from the California Department of Education website, [dq.cde.ca.gov](http://dq.cde.ca.gov).

3. Based on information reported in “How America pays for college 2012,” found online at [www.siena.edu/uploadedfiles/home/SallieMaeHowAmericaPays2012.pdf](http://www.siena.edu/uploadedfiles/home/SallieMaeHowAmericaPays2012.pdf).

4. See Note 3. This total amount includes grants, scholarships, loans, and

assistance from friends and family.

5. Average starting salary taken from the September 2012 salary survey by the National Association of Colleges and Employers.

6. The standard reference here is Bradley Efron and Robert J. Tibshirani, *An Introduction to the Bootstrap*, Chapman Hall, 1993. A less technical overview is in Bradley Efron and Robert J. Tibshirani, "Statistical data analysis in the computer age," *Science*, 253 (1991), pp. 390–395.

7. See [www.thekaraokechannel.com/online/#](http://www.thekaraokechannel.com/online/#).

8. These annual surveys can be found at [www.apa.org/news/press/releases/stress/index.aspx](http://www.apa.org/news/press/releases/stress/index.aspx).

9. C. M. Weaver et al., "Quantification of biochemical markers of bone turnover by kinetic measures of bone formation and resorption in young healthy females," *Journal of Bone and Mineral Research*, 12 (1997), pp. 1714–1720.

10. See Note 5.

11. Euna Hand and Lisa M. Powell, "Consumption patterns of sugar-sweetened beverages in the United States," *Journal of the Academy of Nutrition and Dietetics*, 113, No. 1 (2013), pp. 43–53.

12. See the 2012 press release from the *Student Monitor*, at [www.studentmonitor.com](http://www.studentmonitor.com).

13. Elizabeth Mendes, "U.S. job satisfaction struggles to recover to 2008 levels," Gallup News Service, May 31, 2011. Found at [www.gallup.com/poll/](http://www.gallup.com/poll/).

14. The vehicle is a 2002 Toyota Prius.

15. Regional cost-of-living rates are often computed using the Department of Labor, Bureau of Labor Statistics, metropolitan-area consumer price indexes. These can be found at [www.bls.gov/cpi](http://www.bls.gov/cpi).

16. See Note 11.

17. M. Garaulet et al., "Timing of food intake predicts weight loss effectiveness," *International Journal of Obesity*, 1 (2013), pp. 1–8.

18. Giacomo DeGiorgi et al., "Be as careful of the company you keep as of the books you read: Peer effects in education and on the labor market," National Bureau of Economic Research, working paper 14948 (2009).

19. Seung-Ok Kim, "Burials, pigs, and political prestige in neolithic China," *Current Anthropology*, 35 (1994), pp. 119–141.

20. These data were collected in connection with the Purdue Police Alcohol Student Awareness Program run by Police Officer D. A. Larson.

21. National Assessment of Educational Progress, *The Nation's Report Card*,

Mathematics 2011.

22. Matthew A. Lapierre et al., “Background television in the homes of U.S. children,” *Pediatrics*, 130, No. 5 (2012), pp. 839–846.
23. Sogol Javaheri et al., “Sleep quality and elevated blood pressure in adolescents,” *Circulation*, 118 (2008), pp. 1034–1040.
24. Victor Lun et al., “Evaluation of nutritional intake in Canadian high-performance athletes,” *Clinical Journal of Sports Medicine*, 19, No. 5 (2009), pp. 405–411.
25. R. A. Fisher, “The arrangement of field experiments,” *Journal of the Ministry of Agriculture of Great Britain*, 33 (1926), p. 504, quoted in Leonard J. Savage, “On rereading R. A. Fisher,” *Annals of Statistics*, 4 (1976), p. 471. Fisher’s work is described in a biography by his daughter: Joan Fisher Box, *R. A. Fisher: The Life of a Scientist*, Wiley, 1978.
26. The editorial was written by Phil Anderson. See *British Medical Journal*, 328 (2004), pp. 476–477. A letter to the editor on this topic by Doug Altman and J. Martin Bland appeared shortly after. See “Confidence intervals illuminate absence of evidence,” *British Medical Journal*, 328 (2004), pp. 1016–1017.
27. A. Kamali et al., “Syndromic management of sexually-transmitted infections and behavior change interventions on transmission of HIV-1 in rural Uganda: A community randomised trial,” *Lancet*, 361 (2003), pp. 645–652.
28. T. D. Sterling, “Publication decisions and their possible effects on inferences drawn from tests of significance— or vice versa,” *Journal of the American Statistical Association*, 54 (1959), pp. 30–34. Related comments appear in J. K. Skipper, A. L. Guenther, and G. Nass, “The sacredness of 0.05: A note concerning the uses of statistical levels of significance in social science,” *American Sociologist*, 1 (1967), pp. 16–18.
29. For a good overview of these issues, see Bruce A. Craig, Michael A. Black, and Rebecca W. Doerge, “Gene expression data: The technology and statistical analysis,” *Journal of Agricultural, Biological, and Environmental Statistics*, 8 (2003), pp. 1–28.
30. Erick H. Turner et al., “Selective publication of antidepressant trials and its influence on apparent efficacy,” *New England Journal of Medicine*, 358 (2008), pp. 252–260.
31. Robert J. Schiller, “The volatility of stock market prices,” *Science*, 235 (1987), pp. 33–36.
32. Padmaja Ayyagari and Jody L. Sindelar, “The impact of job stress on smoking and quitting: Evidence from the HRS,” National Bureau of Economic Research, working paper 15232 (2009).

33. Corby K. Martin et al., “Children in school cafeterias select foods containing more saturated fat and energy than the Institute of Medicine recommendations,” *Journal of Nutrition*, 140 (2010), pp. 1653–1660.
34. Data from Joan M. Susic, “Dietary phosphorus intakes, urinary and peritoneal phosphate excretion and clearance in continuous ambulatory peritoneal dialysis patients,” MS thesis, Purdue University, 1985.
35. Mugdha Gore and Joseph Thomas, “Store image as a predictor of store patronage for nonprescription medication purchases: A multiattribute model approach,” *Journal of Pharmaceutical Marketing & Management*, 10 (1996), pp. 45–68.
36. Greg L. Stewart et al., “Exploring the handshake in employment interviews,” *Journal of Applied Psychology*, 93 (2008), pp. 1139–1146.

## Chapter 7

1. Average hours per month obtained from “The Cross-Platform Report, 3rd Quarter 2012,” Nielsen Company (2013).
2. C. Don Wiggins, “The legal perils of ‘underdiversification’— a case study,” *Personal Financial Planning*, 1, No. 6 (1999), pp. 16–18.
3. These data were collected as part of a larger study of dementia patients conducted by Nancy Edwards, School of Nursing, and Alan Beck, School of Veterinary Medicine, Purdue University.
4. These recommendations are based on extensive computer work. See, for example, Harry O. Posten, “The robustness of the one-sample *t*-test over the Pearson system,” *Journal of Statistical Computation and Simulation*, 9 (1979), pp. 133–149; and E. S. Pearson and N. W. Please, “Relation between the shape of population distribution and the robustness of four simple test statistics,” *Biometrika*, 62 (1975), pp. 223–241.
5. The data were obtained on August 24, 2006, from an iPod owned by George McCabe, Jr.
6. The method is described in Xiao-Hua Zhou and Sujuan Gao, “Confidence intervals for the log-normal mean,” *Statistics in Medicine*, 16 (1997), pp. 783–790.
7. You can find a practical discussion of distribution-free inference in Myles Hollander and Douglas A. Wolfe, *Nonparametric Statistical Methods*, 2nd ed., Wiley, 1999.
8. Statistics regarding Facebook usage can be found at [www.facebook.com/notes/facebook-data-team/anatomy-of-facebook/10150388519243859](http://www.facebook.com/notes/facebook-data-team/anatomy-of-facebook/10150388519243859).

9. A description of the lawsuit can be found at [www.cnn.com/2013/02/26/business/california-anheuser-busch-lawsuit/index.html](http://www.cnn.com/2013/02/26/business/california-anheuser-busch-lawsuit/index.html).
10. See Note 1.
11. Christine L. Porath and Amir Erez, "Overlooked but not untouched: How rudeness reduces onlookers' performance on routine and creative tasks," *Organizational Behavior and Human Decision Processes*, 109 (2009), pp. 29–44.
12. The vehicle is a 2002 Toyota Prius owned by the third author.
13. Niels van de Ven et al., "The return trip effect: Why the return trip often seems to take less time," *Psychonomic Bulletin and Review*, 18, No. 5 (2011), pp. 827–832.
14. Sujata Sethi et al., "Study of level of stress in the parents of children with attention-deficit/hyperactivity disorder," *Journal of Indian Association for Child and Adolescent Mental Health*, 8, No. 2 (2012), pp. 25–37.
15. James A. Levine, Norman L. Eberhardt, and Michael D. Jensen, "Role of nonexercise activity thermogenesis in resistance to fat gain in humans," *Science*, 283 (1999), pp. 212–214. Data for this study are available from the *Science* website, [www.sciencemag.org](http://www.sciencemag.org).
16. These data were collected in connection with a bone health study at Purdue University and were provided by Linda McCabe.
17. Data provided by Joseph A. Wipf, Department of Foreign Languages and Literatures, Purdue University.
18. Data from Wayne Nelson, *Applied Life Data Analysis*, Wiley, 1982, p. 471.
19. Summary information can be found at the National Center for Health Statistics website, [www.cdc.gov/nchs/nhanes.htm](http://www.cdc.gov/nchs/nhanes.htm).
20. Detailed information about the conservative  $t$  procedures can be found in Paul Leaverton and John J. Birch, "Small sample power curves for the two sample location problem," *Technometrics*, 11 (1969), pp. 299–307; in Henry Scheffé, "Practical solutions of the Behrens-Fisher problem," *Journal of the American Statistical Association*, 65 (1970), pp. 1501–1508; and in D. J. Best and J. C. W. Rayner, "Welch's approximate solution for the Behrens-Fisher problem," *Technometrics*, 29 (1987), pp. 205–210.
21. This example is adapted from Maribeth C. Schmitt, "The effects of an elaborated directed reading activity on the metacomprehension skills of third graders," PhD dissertation, Purdue University, 1987.
22. See the extensive simulation studies in Harry O. Posten, "The robustness of the two-sample  $t$  test over the Pearson system," *Journal of Statistical*

*Computation and Simulation*, 6 (1978), pp. 295–311.

23. M. Garaulet et al., “Timing of food intake predicts weight loss effectiveness,” *International Journal of Obesity*, advance online publication, January 29, 2013, doi:10.1038/ijo.2012.229.
24. This study is reported in Roseann M. Lyle et al., “Blood pressure and metabolic effects of calcium supplementation in normotensive white and black men,” *Journal of the American Medical Association*, 257 (1987), pp. 1772–1776. The individual measurements in Table 7.5 were provided by Dr. Lyle.
25. Karel Kleisner et al., “Trustworthy-looking face meets brown eyes,” *PLoS ONE*, 8, No. 1 (2013), e53285, doi:10.1371/journal.pone.0053285.
26. Reynol Junco, “Too much face and not enough books: The relationship between multiple indices of Facebook use and academic performance,” *Computers in Human Behavior*, 28, No. 1 (2011), pp. 187–198.
27. C. E. Cryfer et al., “Misery is not miserly: Sad and self-focused individuals spend more,” *Psychological Science*, 19 (2008), pp. 525–530.
28. A. A. Labroo et al., “Of frog wines and frowning watches: Semantic priming, perceptual fluency, and brand evaluation,” *Journal of Consumer Research*, 34 (2008), pp. 819–831.
29. The 2012 study can be found at [www.qsrmagazine.com/content/2012-drive-thru-](http://www.qsrmagazine.com/content/2012-drive-thru-).
30. Grant D. Brinkworth et al., “Long-term effects of a very low-carbohydrate diet and a low-fat diet on mood and cognitive function,” *Archives of Internal Medicine*, 169 (2009), pp. 1873–1880.
31. B. Bakke et al., “Cumulative exposure to dust and gases as determinants of lung function decline in tunnel construction workers,” *Occupational Environmental Medicine*, 61 (2004), pp. 262–269.
32. Samara Joy Nielsen and Barry M. Popkin, “Patterns and trends in food portion sizes, 1977–1998,” *Journal of the American Medical Association*, 289 (2003), pp. 450–453.
33. Gordana Mrdjenovic and David A. Levitsky, “Nutritional and energetic consequences of sweetened drink consumption in 6- to 13-year-old children,” *Journal of Pediatrics*, 142 (2003), pp. 604–610.
34. David Han-Kuen Chu, “A test of corporate advertising using the elaboration likelihood model,” MS thesis, Purdue University, 1993.
35. M. F. Picciano and R. H. Deering, “The influence of feeding regimens on iron status during infancy,” *American Journal of Clinical Nutrition*, 33 (1980), pp. 746–753.
36. The problem of comparing spreads is difficult even with advanced

methods. Common distribution-free procedures do not offer a satisfactory alternative to the  $F$  test, because they are sensitive to unequal shapes when comparing two distributions. A good introduction to the available methods is W. J. Conover, M. E. Johnson, and M. M. Johnson, “A comparative study of tests for homogeneity of variances, with applications to outer continental shelf bidding data,” *Technometrics*, 23 (1981), pp. 351–361. Modern resampling procedures often work well. See Dennis D. Boos and Colin Brownie, “Bootstrap methods for testing homogeneity of variances,” *Technometrics*, 31 (1989), pp. 69–82.

37. G. E. P. Box, “Non-normality and tests on variances,” *Biometrika*, 40 (1953), pp. 318–335. The quote appears on page 333.

38. This city’s restaurant inspection data can be found at [www.jsonline.com/watchdog/dataondemand/](http://www.jsonline.com/watchdog/dataondemand/).

39. Braz Camargo et al., “Interracial friendships in college,” *Journal of Labor Economics*, 28 (2010), pp. 861–892.

40. Based on Loren Cordain et al., “Influence of moderate daily wine consumption on body weight regulation and metabolism in healthy free-living males,” *Journal of the American College of Nutrition*, 16 (1997), pp. 134–139.

41. G. E. Smith et al., “A cognitive training program based on principles of brain plasticity: Results from the Improvement in Memory with Plasticity-Based Adaptive Cognitive Training (IMPACT) study,” *Journal of the American Geriatrics Society*, epub (2009) 57, No. 4, pp. 594–603.

42. Douglas J. Levey et al., “Urban mockingbirds quickly learn to identify individual humans,” *Proceedings of the National Academy of Sciences*, 106 (2009), pp. 8959–8962.

43. B. Wansink et al., “Fine as North Dakota wine: Sensory expectations and the intake of companion foods,” *Physiology & Behavior*, 90 (2007), pp. 712–716.

44. Anne Z. Hoch et al., “Prevalence of the female athlete triad in high school athletes and sedentary students,” *Clinical Journal of Sports Medicine*, 19 (2009), pp. 421–428.

45. This exercise is based on events that are real. The data and details have been altered to protect the privacy of the individuals involved.

46. Based loosely on D. R. Black et al., “Minimal interventions for weight control: A cost-effective alternative,” *Addictive Behaviors*, 9 (1984), pp. 279–285.

47. These data were provided by Professor Sebastian Heath, School of Veterinary Medicine, Purdue University.

48. J. W. Marr and J. A. Heady, “Within- and between-person variation in

dietary surveys: Number of days needed to classify individuals,” *Human Nutrition: Applied Nutrition*, 40A (1986), pp. 347–364.

## Chapter 8

1. The actual distribution of  $X$  based on an SRS from a finite population is the *hypergeometric distribution*. Details regarding this distribution can be found in Sheldon M. Ross, *A First Course in Probability*, 8th ed., Prentice Hall, 2010.
2. From [pewinternet.org/Reports/2013/Coming-and-going-on-facebook.aspx](http://pewinternet.org/Reports/2013/Coming-and-going-on-facebook.aspx), February 5, 2013.
3. Results of the survey are available at [slideshare.net/duckofdoom/google-research-about-mobile-internet-in-2011](http://slideshare.net/duckofdoom/google-research-about-mobile-internet-in-2011).
4. Details of exact binomial procedures can be found in Myles Hollander and Douglas Wolfe, *Nonparametric Statistical Methods*, 2nd ed., Wiley, 1999.
5. See A. Agresti and B. A. Coull, “Approximate is better than ‘exact’ for interval estimation of binomial proportions,” *American Statistician*, 52 (1998), pp. 119–126. A detailed theoretical study is Lawrence D. Brown, Tony Cai, and Anirban DasGupta, “Confidence intervals for a binomial proportion and asymptotic expansions,” *Annals of Statistics*, 30 (2002), pp. 160–201.
6. See, for example, [pilatesmethodalliance.org](http://pilatesmethodalliance.org).
7. See [pewinternet.org/Reports/2013/in-store-mobile-commerce.aspx](http://pewinternet.org/Reports/2013/in-store-mobile-commerce.aspx).
8. Heather Tait, *Aboriginal Peoples Survey, 2006: Inuit Health and Social Conditions*, Social and Aboriginal Statistics Division, Statistics Canada, 2008. Available from [statcan.gc.ca/pub](http://statcan.gc.ca/pub).
9. See [southerncross.co.nz/about-the-group/media-releases/2013.aspx](http://southerncross.co.nz/about-the-group/media-releases/2013.aspx).
10. See [commonsensemedia.org/sites/default/files/full\\_cap-csm\\_report\\_results-1-7-13.pdf](http://commonsensemedia.org/sites/default/files/full_cap-csm_report_results-1-7-13.pdf).
11. See “National Survey of Student Engagement, the College Student Report 2009,” available online at [nsse.iub.edu/index.cfm](http://nsse.iub.edu/index.cfm).
12. This survey and others that study issues related to college students can be found at [nelliemae.com](http://nelliemae.com).
13. See Note 11.
14. Information about the survey can be found online at [saint-denis.library.arizona.edu/natcong](http://saint-denis.library.arizona.edu/natcong).
15. See Note 2.
16. See Alan Agresti and Brian Caffo, “Simple and effective confidence intervals for proportions and differences of proportions result from adding two

successes and two failures,” *American Statistician*, 45 (2000), pp. 280–288. The plus four interval is a bit conservative (true coverage probability is higher than the confidence level) when  $p_1$  and  $p_2$  are equal and close to 0 or 1, but the traditional interval is much less accurate and has the fatal flaw that the true coverage probability is *less* than the confidence level.

17. J. M. Tanner, “Physical growth and development,” in J. O. Forfar and G. C. Arneil, *Textbook of Paediatrics*, 3rd ed., Churchill Livingstone, 1984, pp. 1–292.

18. Based on T. A. Brighton et al., “Low-dose aspirin for preventing recurrent venous thromboembolism,” *New England Journal of Medicine*, 367, No. 21 (2012), pp. 1979–1987. The analysis in the published manuscript used a slightly more complicated summary, called the hazard ratio, to compare the treatments.

19. Edward Bumfardner, “Loss of teeth as a disqualification for military service,” *Transactions of the Kansas Academy of Science*, 18 (1903), pp. 217–219.

20. B. J. Bradley et al., “Historical perspective and current status of the physical education requirement at American 4-year colleges and universities,” *Research Quarterly for Exercise and Sport*, 83, No. 4 (2012), pp. 503–512.

21. Erin K. O’Loughlin et al., “Prevalence and correlates of exergaming in youth,” *Pediatrics*, 130 (2012), pp. 806–814.

22. From a Pew Internet Project Data Memo by Amanda Lenhart et al., dated December 2008. Available at [pewinternet.org](http://pewinternet.org).

23. From Monica Macaulay and Colleen Brice, “Don’t touch my projectile: Gender bias and stereotyping in syntactic examples,” *Language*, 73, No. 4 (1997), pp. 798–825. The first part of the title is a direct quote from one of the texts.

24. The report, dated May 18, 2012, is available from [pewinternet.org/Reports/2012/Future-of-Gamification/Overview.aspx](http://pewinternet.org/Reports/2012/Future-of-Gamification/Overview.aspx).

25. From the Pew Research Center’s Project for Excellence in Journalism, *The State of the News Media 2012*, available from [stateofthemedias.org/?src=prc-headline](http://stateofthemedias.org/?src=prc-headline).

26. See [iom.edu](http://iom.edu).

27. Based on a study described in Corby C. Martin et al., “Children in school cafeterias select foods containing more saturated fat and energy than the Institute of Medicine recommendations,” *Journal of Nutrition*, 140 (2010), pp. 1653–1660.

28. Data are from the NOAA Satellite and Information Service at [ncdc.noaa.gov/special-reports/groundhog-day.php](http://ncdc.noaa.gov/special-reports/groundhog-day.php).

29. From [pewinternet.org/~media/Files/Reports/2013/PIP\\_SocialMediaUsers.pdf](http://pewinternet.org/~media/Files/Reports/2013/PIP_SocialMediaUsers.pdf).
30. From [forbes.com/sites/ericsavitz/2013/01/11/totally-pwned-2012-u-s-video-game-retail-sales-tumble-22](http://forbes.com/sites/ericsavitz/2013/01/11/totally-pwned-2012-u-s-video-game-retail-sales-tumble-22).
31. From the Entertainment Software Association website at [theesa.com/facts](http://theesa.com/facts).
32. See Note 12.
33. See S. W. Lagakos, B. J. Wessen, and M. Zelen, “An analysis of contaminated well water and health effects in Woburn, Massachusetts,” *Journal of the American Statistical Association*, 81 (1986), pp. 583–596, and the following discussion. This case is the basis for the movie *A Civil Action*.
34. This case is discussed in D. H. Kaye and M. Aickin (eds.), *Statistical Methods in Discrimination Litigation*, Marcel Dekker, 1986; and D. C. Baldus and J. W. L. Cole, *Statistical Proof of Discrimination*, McGraw-Hill, 1980.
35. See Note 12.

## Chapter 9

1. From J. Cantor, “Long-term memories of frightening media often include lingering trauma symptoms,” poster paper presented at the Association for Psychological Science Convention, New York, May 26, 2006.
2. When the expected cell counts are small, it is best to use a test based on the exact distribution rather than the chi-square approximation, particularly for  $2 \times 2$  tables. Many statistical software systems offer an “exact” test as well as the chi-square test for  $2 \times 2$  tables.
3. From E. Y. Peck, “Gender differences in film-induced fear as a function of type of emotion measure and stimulus content: A meta-analysis and laboratory study,” PhD dissertation, University of Wisconsin–Madison.
4. D.-C. Seo et al., “Relations between physical activity and behavioral and perceptual correlates among midwestern college students,” *Journal of Americal College Health*, 56, No. 2 (2007), pp. 187–197.
5. See, for example, Alan Agresti, *Categorical Data Analysis*, 2nd ed., Wiley, 2007.
6. From P. Strazzullo et al., “Salt intake, stroke, and cardiovascular disease: a meta-analysis of prospective studies,” *British Medical Journal*, 339 (2009), pp. 1–9. The meta-analysis combined data from 14 study cohorts taken from 10 different studies.
7. N. R. Cook et al., “Long term effects of dietary sodium reduction on cardiovascular disease outcomes: Observational follow-up of the trials of the

hypertension prevention (TOHP),” *British Medical Journal*, 334 (2007), pp. 1–8.

8. The sampling procedure was designed by George McCabe. It was carried out by Amy Conklin, an undergraduate honors student in the Department of Foods and Nutrition at Purdue University.

9. The analysis could also be performed by using a two-way table to compare the states of the selected and not-selected students. Since the selected students are a relatively small percent of the total sample, the results will be approximately the same.

10. See the M&M Mars website at [us.mms.com/us/about/products](http://us.mms.com/us/about/products) for this and other information.

11. Catherine Hill and Holly Kearl, *Crossing the Line: Sexual Harassment at School*, American Association of University Women, Washington, DC, 2011.

12. Based on [pewsocialtrends.org/files/2011/08/online-learning.pdf](http://pewsocialtrends.org/files/2011/08/online-learning.pdf).

13. For an overview of remote deposit capture, see [remotedepositcapture.com/overview/rdc.overview.aspx](http://remotedepositcapture.com/overview/rdc.overview.aspx).

14. From the Community Bank Competitiveness Survey, 2008, *ABA Banking Journal*. The survey is available at [nxtbook.com/nxtbooks/sb/ababj-compsurv08/index.php](http://nxtbook.com/nxtbooks/sb/ababj-compsurv08/index.php).

15. See [nhcaa.org](http://nhcaa.org).

16. These data are a composite based on several actual audits of this type.

17. Data provided by Professor Marcy Towns of the Purdue University Department of Chemistry.

18. Based on *The Ethics of American Youth–2008*, available from the Josephson Institute at [charactercounts.org/programs/reportcard](http://charactercounts.org/programs/reportcard).

19. From the Survey of Canadian Career College Students Phase II: In-School Student Survey, 2008. This report is available from [hrsdc.gc.ca/eng/publications\\_resources](http://hrsdc.gc.ca/eng/publications_resources).

## Chapter 10

1. Data based on Michael L. Mestek et al., “The relationship between pedometer-determined and self-reported physical activity and body composition variables in college-aged men and women,” *Journal of American College Health*, 57 (2008), pp. 39–44.

2. The vehicle is a Pontiac transport van.

3. Information regarding bone health can be found in “Osteoporosis: Peak bone

mass in women,” last reviewed in January 2012 and available at [www.niams.nih.gov/Health\\_Info/Bone/Osteoporosis/bone\\_mass.asp](http://www.niams.nih.gov/Health_Info/Bone/Osteoporosis/bone_mass.asp).

4. The data were provided by Linda McCabe and were collected as part of a large study of women’s bone health and another study of calcium kinetics, both directed by Professor Connie Weaver of the Department of Foods and Nutrition, Purdue University.

5. These data were provided by Professor Wayne Campbell of the Purdue University Department of Foods and Nutrition.

6. For more information about nutrient requirements, see the Institute of Medicine publications on Dietary Reference Intakes available at [www.nap.edu](http://www.nap.edu).

7. The method is described in Chapter 2 of M. Kutner et al., *Applied Linear Statistical Models*, 5th ed., Irwin, 2004.

8. National Science Foundation, Division of Science Resources Statistics, *Academic Research and Development Expenditures: Fiscal Year 2009*, Detailed Statistical Tables NSF 11-313, Arlington, VA, 2011. Available at [www.nsf.gov/statistics/nsf11313/](http://www.nsf.gov/statistics/nsf11313/).

9. This annual report can be found at [www.kiplinger.com](http://www.kiplinger.com).

10. Tuition rates for 2008 and 2011 were obtained from [www.findthebest.com](http://www.findthebest.com).

11. These are part of the data from the EESEE story “Blood Alcohol Content,” found on the text website, [www.whfreeman.com/ips8e](http://www.whfreeman.com/ips8e).

12. M. Mondello and J. Maxcy, “The impact of salary dispersion and performance bonuses in NFL organizations,” *Management Decision*, 47 (2009), pp. 110–123. These data were collected from [www.cbssports.com/nfl/playerrankings/regularseason/](http://www.cbssports.com/nfl/playerrankings/regularseason/) and [content.usatoday.com/sports/football/nfl/salaries/](http://content.usatoday.com/sports/football/nfl/salaries/).

13. Selling price and assessment value available at [php.jconline.com/propertysales/propertysales.php](http://php.jconline.com/propertysales/propertysales.php).

14. Data available at [www.ncdc.noaa.gov](http://www.ncdc.noaa.gov).

15. Matthew P. Martens et al., “The co-occurrence of alcohol use and gambling activities in first-year college students,” *Journal of American College Health*, 57 (2009), pp. 597–602.

16. Based on Dan Dauwalter’s master’s thesis in the Department of Forestry and Natural Resources at Purdue University. More information is available in Daniel C. Dauwalter et al., “An index of biotic integrity for fish assemblages in Ozark Highland streams of Arkansas,” *Southeastern Naturalist*, 2 (2003), pp. 447–468. These data were provided by Emmanuel Frimpong.

17. G. Geri and B. Palla, “Considerazioni sulle più recenti osservazioni ottiche alla Torre Pendente di Pisa,” *Estratto dal Bollettino della Società Italiana di Topografia e Fotogrammetria*, 2 (1988), pp. 121–135. Professor Julia Mortera of the University of Rome provided valuable assistance with the translation.
18. M. Kuo et al., “The marketing of alcohol to college students: The role of low prices and special promotions,” *American Journal of Preventive Medicine*, 25, No. 3 (2003), pp. 204–211.
19. Rates can be found in “Annual Return of Key Indices (1993–2012),” available at [www.lazardnet.com](http://www.lazardnet.com).
20. These data can be found in the report titled “Grade inflation at American colleges and universities,” at [www.gradeinflation.com](http://www.gradeinflation.com).
21. Toben F. Nelson et al., “The state sets the rate: The relationship among state-specific college binge drinking, state binge drinking rates, and selected state alcohol control policies,” *American Journal of Public Health*, 95, No. 3 (2005), pp. 441–446.
22. Data on a sample of 12 of 56 perch in a data set contributed to the *Journal of Statistics Education* data archive [www.amstat.org/publications/jse/](http://www.amstat.org/publications/jse/) by Juha Puranen of the University of Helsinki.
23. L. Cooke et al., “Relationship between parental report of food neophobia and everyday food consumption in 2–6-year-old children,” *Appetite*, 41 (2003), pp. 205–206.
24. Alexandra Burt, “A mechanistic explanation of popularity: Genes, rule breaking, and evocative gene-environment correlations,” *Journal of Personality and Social Psychology*, 96 (2009), pp. 783–794.

## Chapter 11

1. This data set is similar to those used at Purdue University to assess academic success.
2. Mary E. Pritchard and Gregory S. Wilson, “Predicting academic success in undergraduates,” *Academic Exchange Quarterly*, 11 (2007), pp. 201–206.
3. R. M. Smith and P. A. Schumacher, “Predicting success for actuarial students in undergraduate mathematics courses,” *College Student Journal*, 39, No. 1 (2005), pp. 165–177.
4. Based on Leigh J. Maynard and Malvern Mupandawana, “Tipping behavior in Canadian restaurants,” *International Journal of Hospitality Management*, 28 (2009), pp. 597–603.
5. Kathleen E. Miller, “Wired: Energy drinks, jock identity, masculine norms, and risk taking,” *Journal of American College Health*, 56 (2008), pp. 481–489.

6. From a table entitled “Largest Indianapolis-area architectural firms,” *Indianapolis Business Journal*, December 16, 2003.
7. The data were obtained from the Internet Movie Database (IMDb), available at [www.imdb.com](http://www.imdb.com) on April 20, 2010.
8. The 2009 table of 200 top universities can be found at [www.timeshighereducation.co.uk](http://www.timeshighereducation.co.uk).
9. The results were published in C. M. Weaver et al., “Quantification of biochemical markers of bone turnover by kinetic measures of bone formation and resorption in young healthy females,” *Journal of Bone and Mineral Research*, 12 (1997), pp. 1714–1720. The data were provided by Linda McCabe.
10. This data set was provided by Joanne Lasrado of the Purdue University Department of Foods and Nutrition.
11. These data are based on experiments performed by G. T. Lloyd and E. H. Ramshaw of the CSIRO Division of Food Research, Victoria, Australia. Some results of the statistical analyses of these data are given in G. P. McCabe, L. McCabe, and A. Miller, “Analysis of taste and chemical composition of cheddar cheese, 1982–83 experiments,” CSIRO Division of Mathematics and Statistics Consulting Report VT85/6; and in I. Barlow et al., “Correlations and changes in flavour and chemical parameters of cheddar cheeses during maturation,” *Australian Journal of Dairy Technology*, 44 (1989), pp. 7–18.

## Chapter 12

1. R. Kanai et al., “Online social network size is reflected in human brain structure,” *Proceedings of the Royal Society—Biological Sciences*, 297 (2012), pp. 1327–1334.
2. Based on Stephanie T. Tong et al., “Too much of a good thing? The relationship between number of friends and interpersonal impressions on Facebook,” *Journal of Computer-Mediated Communication*, 13 (2008), pp. 531–549.
3. This rule is intended to provide a general guideline for deciding when serious errors may result by applying ANOVA procedures. When the sample sizes in each group are very small, this rule may be a little too conservative. For unequal sample sizes, particular difficulties can arise when a relatively small sample size is associated with a population having a relatively large standard deviation.
4. Penny M. Simpson et al., “The eyes have it, or do they? The effects of model eye color and eye gaze on consumer ad response,” *Journal of Applied Business and Economics*, 8 (2008), pp. 60–71.

5. Several different definitions for the noncentrality parameter of the noncentral  $F$  distribution are in use. When  $I = 2$ , the  $\gamma$  defined here is equal to the square of the noncentrality parameter  $\delta$  that we used for the two-sample  $t$  test in Chapter 7. Many authors prefer  $=\lambda/I$ . We have chosen to use  $\gamma$  because it is the form needed for the SAS function PROBF.
6. Bryan Raudenbush et al., "Pain threshold and tolerance differences among intercollegiate athletes: Implication of past sports injuries and willingness to compete among sports teams," *North American Journal of Psychology*, 14 (2012), pp. 85–94.
7. Eileen Wood et al., "Examining the impact of off-task multi-tasking with technology on real-time classroom learning," *Computers & Education*, 58 (2012), pp. 365–374.
8. Kendall J. Eskine, "Wholesome foods and wholesome morals? Organic foods reduce prosocial behavior and harshen moral judgments," *Social Psychological and Personality Science*, 2012, doi: 10.1177/1948550612447114.
9. Adam I. Perlman et al., "Massage therapy for osteoarthritis of the knee: A randomized dose-finding trial," *PLoS ONE*, 7, No. 2 (2012), e30248, doi:10.1371/journal.pone.0030248.
10. Jesus Tanguma et al., "Shopping and bargaining in Mexico: The role of women," *Journal of Applied Business and Economics*, 9 (2009), pp. 34–40.
11. Jeffrey T. Kullgren et al., "Individual- versus group-based financial incentives for weight loss," *Annals of Internal Medicine*, 158, No. 7 (2013), pp. 505–514.
12. Corinne M. Kodama and Angela Ebreo, "Do labels matter? Attitudinal and behavioral correlates of ethnic and racial identity choices among Asian American undergraduates," *College Student Affairs Journal*, 27, No. 2 (2009), pp. 155–175.
13. Sangwon Lee and Seonmi Lee, "Multiple play strategy in global telecommunication markets: An empirical analysis," *International Journal of Mobile Marketing*, 3 (2008), pp. 44–53.
14. Christie N. Scollon et al., "Emotions across cultures and methods," *Journal of Cross-cultural Psychology*, 35 (2004), pp. 304–326.
15. Adrian C. North et al., "The effect of musical style on restaurant consumers' spending," *Environment and Behavior*, 35 (2003), pp. 712–718.
16. Woo Gon Kim et al., "Influence of institutional DINESERV on customer satisfaction, return intention, and word-of-mouth," *International Journal of Hospitality Management*, 28 (2009), pp. 10–17.
17. The experiment was performed in Connie Weaver's lab in the Purdue

University Department of Foods and Nutrition. The data were provided by Berdine Martin and Yong Jiang.

18. The data were provided by James Kaufman. The study is described in James C. Kaufman, “The cost of the muse: Poets die young,” *Death Studies*, 27 (2003), pp. 813–821. The quote from Yeats appears in this article.

19. Data provided by Jo Welch of the Purdue University Department of Foods and Nutrition.

20. Steve Badylak et al., “Marrow-derived cells populate scaffolds composed of xenogeneic extracellular matrix,” *Experimental Hematology*, 29 (2001), pp. 1310–1318.

21. This exercise is based on data provided from a study conducted by Jim Baumann and Leah Jones of the Purdue University School of Education.

## Chapter 13

1. See [www.who.int/topics/malaria/en/](http://www.who.int/topics/malaria/en/) for more information about malaria.

2. This example is based on a 2009 study described at [clinicaltrials.gov/ct2/show/NCT00623857](http://clinicaltrials.gov/ct2/show/NCT00623857).

3. We present the two-way ANOVA model and analysis for the general case in which the sample sizes may be unequal. If the sample sizes vary a great deal, serious complications can arise. There is no longer a single standard ANOVA analysis. Most computer packages offer several options for the computation of the ANOVA table when cell counts are unequal. When the counts are approximately equal, all methods give essentially the same results.

4. Euna Hand and Lisa M. Powell, “Consumption patterns of sugar-sweetened beverages in the United States,” *Journal of the Academy of Nutrition and Dietetics*, 113, No. 1 (2013), pp. 43–53.

5. Rick Bell and Patricia L. Pliner, “Time to eat: The relationship between the number of people eating and meal duration in three lunch settings,” *Appetite*, 41 (2003), pp. 215–218.

6. Karolyn Drake and Jamel Ben El Hine, “Synchronizing with music: Intercultural differences,” *Annals of the New York Academy of Sciences*, 99 (2003), pp. 429–437.

7. Example 13.10 is based on a study described in P. D. Wood et al., “Plasma lipoprotein distributions in male and female runners,” in P. Milvey (ed.), *The Marathon: Physiological, Medical, Epidemiological, and Psychological Studies*, New York Academy of Sciences, 1977.

8. Gerardo Ramirez and Sian L. Beilock, “Writing about testing worries boosts exam performance in the classroom,” *Science*, 331 (2011), pp. 211–213.

9. Felix Javier Jimenez-Jimenez et al., "Influence of age and gender in motor performance in healthy adults," *Journal of the Neurological Sciences*, 302 (2011), pp. 72–80.
10. Tomas Brodin et al., "Dilute concentrations of a psychiatric drug alter behavior of fish from natural populations," *Science*, 339 (2013), pp. 814–815.
11. Vincent P. Magnini and Kiran Karande, "The influences of transaction history and thank you statements in service recovery," *International Journal of Hospitality Management*, 28 (2009), pp. 540–546.
12. Brian Wansink et al., "The office candy dish: Proximity's influence on estimated and actual consumption," *International Journal of Obesity*, 30 (2006), pp. 871–875.
13. Data based on Brian T. Gold et al., "Lifelong bilingualism maintains neural efficiency for cognitive control in aging," *Journal of Neuroscience*, 33, No. 2 (2013), pp. 387–396.
14. Annette N. Senitko et al., "Influence of endurance exercise training status and gender on postexercise hypotension," *Journal of Applied Physiology*, 92 (2002), pp. 2368–2374.
15. Willemijn M. van Dolen, Ko de Ruyter, and Sandra Streukens, "The effect of humor in electronic service encounters," *Journal of Economic Psychology*, 29 (2008), pp. 160–179.
16. Jane Kolodinsky et al., "Sex and cultural differences in the acceptance of functional foods: A comparison of American, Canadian, and French college students," *Journal of American College Health*, 57 (2008), pp. 143–149.
17. Judith McFarlane et al., "An intervention to increase safety behaviors of abused women," *Nursing Research*, 51 (2002), pp. 347–354.
18. Gad Saad and John G. Vongas, "The effect of conspicuous consumption on men's testosterone levels," *Organizational Behavior and Human Decision Processes*, 110 (2009), pp. 80–92.
19. Klaus Boehnke et al., "On the interrelation of peer climate and school performance in mathematics: A German-Canadian-Israeli comparison of 14-year-old school students," in B. N. Setiadi, A. Supratiknya, W. J. Lonner, and Y. H. Poortinga (eds.), *Ongoing Themes in Psychology and Culture* (Online Ed.), International Association for Cross-Cultural Psychology.
20. Data provided by Julie Hendricks and V. J. K. Liu of the Department of Foods and Nutrition, Purdue University.
21. Lijia Lin et al., "Animated agents and learning: Does the type of verbal feedback they provide matter?" *Computers and Education*, 2013, doi: 10.1016/j.compedu.2013.04.017.

22. Tamar Kugler et al., “Trust between individuals and groups: Groups are less trusting than individuals but just as trustworthy,” *Journal of Economic Psychology*, 28 (2007), pp. 646–657.
23. Based on A. A. Adish et al., “Effect of consumption of food cooked in iron pots on iron status and growth of young children: A randomised trial,” *Lancet*, 353 (1999), pp. 712–716.
24. Based on a problem from Renée A. Jones and Regina P. Becker, Department of Statistics, Purdue University.
25. For a summary of this study and other research in this area, see Stanley Coren and Diane F. Halpern, “Left-handedness: A marker for decreased survival fitness,” *Psychological Bulletin*, 109 (1991), pp. 90–106.
26. Data provided by Neil Zimmerman of the Purdue University School of Health Sciences.
27. See I. C. Feller et al., “Sex-biased herbivory in Jack-in-the-pulpit (*Arisaema triphyllum*) by a specialist thrips (*Heterothrips arisaemae*),” in *Proceedings of the 7th International Thysanoptera Conference*, Reggio Callabrio, Italy, pp. 163–172.

## PHOTO CREDITS

### CHAPTER 1

<b>PAGE 1</b>	Jordan Siemens/Getty Images
<b>PAGE 4</b>	Alamy
<b>PAGE 10</b>	© Carl Skepper/Alamy
<b>PAGE 63</b>	Mitchell Layton/Getty Images

### CHAPTER 2

<b>PAGE 81</b>	Sam Edwards/Getty Images
<b>PAGE 82</b>	Alamy
<b>PAGE 87</b>	© Kristoffer Tripplaar/Alamy
<b>PAGE 154</b>	Alamy

### CHAPTER 3

<b>PAGE 167</b>	Thinkstock
<b>PAGE 170</b>	U.S. Department of Education Institute of Education Sciences National Center for Education Statistics
<b>PAGE 176</b>	Alamy
<b>PAGE 180</b>	© Alex Segre/Alamy
<b>PAGE 192</b>	hartcreations/iStockphoto
<b>PAGE 195</b>	© Ann E Parry/Alamy
<b>PAGE 199</b>	GSS
<b>PAGE 214</b>	© blickwinkel/Alamy
<b>PAGE 222</b>	National Archives and Records Administration NARA

## **CHAPTER 4**

<b>PAGE 231</b>	Jgroup/Dreamstime.com
<b>PAGE 239</b>	© MBI/Alamy
<b>PAGE 241</b>	Norlito/iStockphoto
<b>PAGE 246</b>	Profimedia.CZ a.s./Alamy
<b>PAGE 272</b>	skynesher/iStockphoto
<b>PAGE 285</b>	© Randy Faris/Corbis

## **CHAPTER 5**

<b>PAGE 301</b>	Digital Vision/Thinkstock
<b>PAGE 306</b>	Jacob Wackerhausen/iStockphoto
<b>PAGE 321</b>	Istockphoto/Thinkstock
<b>PAGE 324</b>	D. Hurst/Alamy
<b>PAGE 330</b>	NetPhotos/Alamy

## **CHAPTER 6**

<b>PAGE 351</b>	Alamy
<b>PAGE 354</b>	© Syracuse Newspapers/Caroline Chen/The Image Works
<b>PAGE 376</b>	Alamy
<b>PAGE 383</b>	Joe Raedle/Getty Images
<b>PAGE 386</b>	Olivier Voisin/Photo Researchers
<b>PAGE 408</b>	Photo by The Photo Works

## **CHAPTER 7**

<b>PAGE 417</b>	Getty Images/Blend Images RM
<b>PAGE 429</b>	Richard Kail/Photo Researchers, Inc.
<b>PAGE 437</b>	© Oramstock/Alamy
<b>PAGE 449</b>	Robert Warren/Getty Images
<b>PAGE 452</b>	Getty Images/Photo Researchers
<b>PAGE 456</b>	Serge Krouglikoff/Getty Images

## **CHAPTER 8**

<b>PAGE 487</b>	iStockphoto/Thinkstock
<b>PAGE 494</b>	Photolibrary
<b>PAGE 496</b>	Alamy

**PAGE 501** iStockphoto

## **CHAPTER 9**

**PAGE 529** © Image Source/Alamy

**PAGE 530** © Pixellover RM 6/Alamy

**PAGE 541** Alamy

**PAGE 549** Alamy

## **CHAPTER 10**

**PAGE 563** Jack Hollingsworth/Photodisc/Getty

**PAGE 566** Ruth Jenkinson/Getty Images

**PAGE 581** © Drive Images/Alamy

**PAGE 583** Doncaster and Bassetlaw Hospitals/Science Source

## **CHAPTER 11**

**PAGE 611** Barry Austin Photography/Getty Images

**PAGE 631** © Radius Images/Alamy

## **CHAPTER 12**

**PAGE 643** © Monkey Business Images Ltd/ Dreamstime.com

**PAGE 645** © Ingram Publishing/Alamy

**PAGE 673** Thinkstock

## **CHAPTER 13**

**PAGE 691** © Jupiter Images/Getty Images

**PAGE 695** Professor Pietro M. Motta/Photo Researchers

**PAGE 698** © Patti McConville/Alamy

**PAGE 700** © Banana Stock/Agfotostock

## **CHAPTER 14**

**PAGE 14-1** © Blend Images/Alamy

**PAGE 14-12** Nigel Cattlin/Alamy

**PAGE 14-17** © ZUMA Press, Inc./Alamy

## **CHAPTER 15**

- PAGE 15-1** Steven King/Icon SMI 258/Steven King/Icon SMI/Newscom  
**PAGE 15-4** DWC/Alamy  
**PAGE 15-11** © Jeff Greenberg/Alamy  
**PAGE 15-18** © Iofoto/Dreamstime  
**PAGE 15-23** Photo by Jason Barnette; courtesy of Purdue University

## **CHAPTER 16**

- PAGE 16-1** Digital Vision/Thinkstock  
**PAGE 16-14** Digital Vision/Thinkstock  
**PAGE 16-50** istockphoto

## **CHAPTER 17**

- PAGE 17-1** Pressmaster/Shutterstock  
**PAGE 17-4** Michael Rosenfeld/Getty Images  
**PAGE 17-8** Alamy  
**PAGE 17-9** George Frey/Bloomberg via Getty Images  
**PAGE 17-12** © Jeff Greenberg/The Image Works

## INDEX

---

- Acceptance sampling, 406
- Alternative hypothesis. *See* Hypothesis, alternative
- ACT college entrance examination, 75, 318, 608–609
- Adequate Calcium Today (ACT) study, 551
- Analysis of variance (ANOVA)
  - one-way, 644–677
  - regression, 586–589, 617–618
  - two-way, 692–706, 708
- Analysis of variance table
  - one-way, 657–662
  - regression, 589, 617
  - two-way, 702–703
- Anonymity, 221–222
- Aggregation, 148
- Applet
  - Central Limit Theorem, 309, 310, 311
  - Confidence Interval, 357, 358, 371, 413
  - Correlation and Regression, 105, 108, 109, 126, 138
  - Law of Large Numbers, 236, 269
  - Mean and Median, 34, 51, 52
  - Normal Approximation to Binomial, 333
  - Normal Curve, 65, 74
  - One-variable statistical calculator, 17
  - One-Way ANOVA, 660, 682
  - Probability, 232, 236
  - Simple Random Sample, 191, 195, 203, 217
  - Statistical Power, 411, 412
  - Statistical Significance, 393, 394
  - Two-variable statistical calculator, 126
    - Probability, 217, 232, 236, 346
- AppsFire, 303, 317
- Association, 83–84, 536

- and causation, 134, 136, 152–153
  - negative, 90, 98
  - positive, 90, 98
- Attention deficit hyperactivity disorder (ADHD), 444
- Available data, 169, 174
  
- Bar graph, 11, 25
- Bayes's rule, 292–293
- Behavioral and social science experiments, 224–226
- Benford's law, 242–243, 254
- Bias *see also* Unbiased estimator
  - in a sample, 194, 198, 200, 201, 210, 211, 212, 215
  - in an experiment, 179, 188, 207
  - of a statistic, 210–211, 215
- Binomial coefficient, 337, 343
- Binomial distribution. *See* Distribution, binomial
- Binomial setting, 322, 343, 14-1
- Block, 187, 188
- Bonferroni procedure, 402, 670–671
- Bootstrap, 367, *See also* Chapter 16
- Boston Marathon, 30, 17-40
- Boxplot, 37–38, 48
  - modified, 41, 48
  - side-by-side, 41, 48, 643, 649
- Buffon, Count, 234
  
- Canadian Internet Use Survey (CIUS), 14-25
- Capability, 17-34
- Capture-recapture sampling, 214
- Case, 2, 8, 613
- Categorical data. *See* Variable, categorical
- Causation, 134, 136, 152–155, 156, 176
- Cause-and-effect diagram, 17-4
- Cell, 140, 693
- Census, 171, 174, 636
- Census Bureau, 9, 203, 204, 349, 391
- Center of a distribution, 20, 25, 47
- Centers for Disease Control and Prevention, 163, 166, 230, 347, 608

Central limit theorem, 3, 307–313, 314, 316  
 Chi-square distribution. *See* Distribution, chi-square  
 Chi-square statistic, 538, 550  
     and the  $z$  statistic, 544–545  
     goodness of fit test, 552–553  
 Clinical trials, 222  
 Clusters in data, 50, 95  
 Coefficient of determination, 662. *See* Correlation, multiple  
 Coin tossing, 233, 323, 331, 335, 339  
 Column variable. *See* Variable, row and column  
 Common response, 153, 156  
 Complement of an event. *See* Event, complement  
*Condé Nast Traveler magazine*, 15-4, 15-20  
 Conditional distribution. *See* Distribution, conditional  
 Conditional probability. *See* Probability, conditional  
 Confidence interval, 356–358, 368  
     bootstrap, 16-14–16-16, 16-32–16-38  
     cautions, 365–367  
     for multiple comparisons, 672  
     for odds ratio, 14-10, 14-19  
     for slope in a logistic regression, 14-10, 14-19  
     relation to two-sided tests, 387–388  
      $t$  for a contrast, 665  
      $t$  for difference of means, 451–454, 467  
         pooled, 462  
      $t$  for matched pairs, 431  
      $t$  for mean response in regression, 576–578, 585  
      $t$  for one mean, 420–421, 441  
      $t$  for regression parameters, 574, 584, 616, 633  
      $z$  for one mean, 358–361  
      $z$  for one proportion  
         large sample, 490, 503  
         plus four, 493, 503  
      $z$  for difference of proportions  
         large sample, 510, 522  
         plus four, 514, 522  
     simultaneous, 672  
 Confidence level, 356, 368

Confidentiality, 220–221, 226

Confounding, 153–154, 156, 173, 174, 429

Consumer Behavior Report 14-25–14-26

Consumer Report on Eating Share Trend (CREST) 631, 14-24

Consumer Reports National Research Center, 330

Consumers Union, 87, 16-37

Continuity correction, 335–336, 343, 15-7

Contrast, 650, 663–668, 678

Control chart, 17-7, 17-17

- individuals chart, 17-41
- $p$  chart, 17-52–17-57
- $R$  chart, 17-23, 17-35
- $s$  chart, 17-12–17-17
- $\bar{x}$  chart, 17-8–17-12, 17-14, 17-17

Control group, 172, 178, 179, 188, 320, 397, 400, 686

Correlation, 103–104, 107, 275

- and regression, 119, 121
- based on averaged data, 134, 136
- between random variables, 275
- bootstrap confidence interval, 16-36–16-38
- cautions about, 126–134
- multiple, 618. *See* Coefficient of determination
- nonsense, 134
- inference for, 597–599
- population, 597
- properties, 105
- squared, 119–120, 121, 588, 600
- test for, 597, 600

Count, 10, 487 *see also* Frequency

- distribution of, 320–325, 339–342, 343

Critical value, 389, 390

- of chi-square distribution, 539, Table F
- of  $F$  distribution, 474–476, Table E
- of standard normal distribution, 65, 359, 389, Table A
- of  $t$  distribution, 419–420, Table D

Cumulative proportion, 63–65, 72

- standard normal, 63, Table A

- Data, 2
  - Anecdotal, 168, 174
  - Available, 169, 174
- Data mining, 135
- Decision analysis, 406–411
- Degree of Reading Power, 452–455, 16-43–16-46
- Degrees of freedom, 44
  - approximation for, 451, 460, 467
  - of chi-square distribution, 538
  - of chi-square test, 539
  - of  $F$  distribution, 474
  - of one-way ANOVA, 658–659
  - of  $t$  distribution, 419, 441
  - of two-way ANOVA, 697–698, 702–703
  - of regression ANOVA, 587, 589, 600
  - of regression  $t$ , 574, 577, 579, 584
  - of regression  $s^2$ , 569
  - of two-sample  $t$ , 451, 460, 462
- Deming, W. Edwards, 17-40
- Density curve, 54–56, 72, 257–259, 261
- Density estimation, 71
- Department of Transportation, 17-57
- Design, 183. *see also* Experiments
  - block, 187–188
  - experimental, 179
  - repeated-measures, 701, 709, 711, 713, 714
  - sampling, 192–201
- Direction of a relationship, 89, 98
- Disjoint events. *See* Event, disjoint
- Distribution, 3, 25
  - bimodal, 72
  - binomial, 322–327, 343, 14-2, Table C
    - formula, 336–338, 343
    - Normal approximation, 331–335, 343
    - use in the sign test, 438–439
  - bootstrap, 16-24–16-30
  - of categorical variable, 10

chi-square, 538, 550, Table F  
 conditional, 144, 148, 533  
 describing, 20  
 examining, 20  
 exponential, 309  
 geometric, 348  
 $F$ , 474–475, Table E  
 joint, 141–142, 148  
 jointly Normal, 597  
 marginal, 142–143, 148  
 noncentral  $F$ , 676  
 noncentral  $t$ , 477  
 Normal, 58–59, 72, 257–261  
     for probabilities, 257–261  
     standard, 63, 72, Table A  
 Poisson, 339–342  
 population, 302  
 probability. *See* Probability distribution  
 of quantitative variable, 13–18  
 sampling. *See* Sampling distribution  
 skewed, 2, 25  
 symmetric, 20, 25  
 $t$ , 419–420 Table D  
 trimodal, 72  
 tails, 19  
 uniform, 257–258  
 unimodal, 20  
 Weibull, 315–316  
 Distribution-free procedure, 436. *See also* Chapter 15  
 Double-blind experiment, 185, 188  
 Dual X-ray absorptiometry scanner, 375, 445–446, 447, 17-38–17-39  
 Estimation, 267–268  
 Ethics, 167, 217–226  
 Excel, 4, 5, 91, 158, 184, 196, 427, 459, 491, 511, 571, 613, 630, 674, 707  
 Expected value, 265 *see also* Mean of a random variable  
 Expected cell count, 537, 550, 552, 556  
 Experiment, 171, 172, 174

- block design, 187, 188
- cautions about, 185–186
- comparative, 178, 188
- completely randomized, 184
- matched pairs, 186, 188
- principles, 181
- units, 175, 188

Explanatory variable. *See* Variable, explanatory

Exploratory data analysis, 9, 25, 167

Extrapolation, 113, 121

Event, 239, 248

- complement of, 240, 249, 283, 294
- disjoint, 240, 249, 240, 248, 283
- empty, 285
- independent, 240, 249
- intersection, 290
- union, 283

*F* distribution. *See* Distribution, *F*

*F* test

- one-way ANOVA, 660
- regression ANOVA, 588, 618
- for collection of regression coefficients, 630–631, 637
- for standard deviations, 474
- two-way ANOVA, 703

Facebook, 28, 317, 318, 442, 465, 468, 488–493, 510–511, 518–520, 525, 530–531, 648–662, 663–673, 680, 687, 689, 14-2–14-4, 14-6–14-8, 15-26, 15-33

Factor, experimental, 175, 188, 643, 692–696

Federal Aviation Administration (FAA), 319

Fisher, Sir R. A., 396, 411, 475

Fitting a line, 110–111

Five-number summary, 37–38, 48

Flowchart, 17-4–17-5

Form of a relationship, 89, 98

Frequency, 16, 25

Frequency table, 16

Gallup-Healthways Well-Being Index, 370

Gallup Poll, 202, 345, 346  
Genetic counseling, 297  
Genomics, 399  
Goodness of fit, 551–556  
Gosset, William, 48, 420, 16-11

Health and Retirement Study (HRS), 413  
Histogram, 15–18, 25  
Hypothesis  
    alternative, 374–375, 390  
        one-sided, 375, 390  
        two-sided, 375, 390  
    null, 374, 390  
Hypothesis testing, 410–411. *See also* Significance test

Independence, 235  
    in two-way tables, 547–548, 550  
    of events, 244–245, 293, 294  
    of random variables, 275  
Indicator variable, 635–636, 710, 14-4  
Inference, statistical. *See* Statistical inference  
Influential observation, 129–130, 136, 573, 625  
Informed consent, 220, 226  
Institutional review board (IRB), 219, 226  
Instrument, 6  
Interaction, 696, 697–701  
Intercept of a line, 111  
    of least-squares line, 115, 121, 565  
Internet Movie Database (IMDb), 637  
Intervention, 173  
Intersection of events, 290, 294  
Interquartile range (IQR), 39–40, 48  
iPod, 437  
iTunes, 2–3

JMP, 118, 146, 459, 491, 497, 511, 519, 532, 14-14

Karaoke Channel, 369  
Kerrich, John, 234

Key characteristics of a data set, 4, 8  
 Key characteristics of data for relationships, 85  
 Kruskal-Wallis test, 15-28–15-33  
  
 Label, 2, 3, 8  
 Law of large numbers, 267–270, 279  
 Law School Admission Test (LSAT), 401, 481  
 Leaf, in a stemplot, 13, 25  
 Leaning Tower of Pisa, 607  
 Least significant difference, 670  
 Least squares, 114, 615–616  
 Least squares regression line, 113–115, 121, 563, 584  
 Level of a factor, 175, 188, 653, 675, 692–697  
 Line, equation of, 111
 

- least-squares, 114, 568

 Linear relationship, 89, 98  
 Linear transformation. *See* Transformation, linear  
 Logarithm transformation. *See* Transformation, logarithm  
 Logistic regression, 631–632. *See also* Chapter 14  
 Logit, 14-5  
 Lurking variable. *See* Variable, lurking  
  
 Main effect, 696, 697–701  
 Major League Baseball, 15-3  
 Mann-Whitney test, 15-5  
 Margin of error, 211, 215, 356, 362, 368
 

- for a difference in two means, 451, 467
- for a difference in two proportions, 510, 521
- for a single mean, 359–360, 368, 420–421, 441
- for a single proportion, 490, 503

 Marginal means, 699, 708  
 Matched pairs design, 186, 188
 

- inference for, 429–432, 438–440, 15-20, 15-25

 Mean, 31
 

- of binomial distribution, 328, 343
- of density curve, 56–57, 72
- of difference of sample means, 449
- of difference of sample proportions, 508

- of normal distribution, 58–59
- of random variable, 264–265, 279
- rules for, 271–271, 279
- of sample mean, 305–306, 316
- of sample proportion, 489
- trimmed, 53
- versus median, 34

Mean square

- in one-way ANOVA, 659–661, 678
- in two-way ANOVA, 702–703
- in multiple linear regression, 617, 634
- in simple linear regression, 587, 589

Median, 33

- inference for, 438–440
- of density curve, 56–57, 72

Mendel, Gregor, 246–247

Meta-analysis, 548–549, 551

Minitab, 116, 145, 325, 426, 427, 438, 492, 497, 512, 519, 533, 536, 554, 557, 570, 608, 622, 629, 674, 675, 681, 707, 14-4, 14-11, 14-14, 14-16, 14-18, 14-21, 14-22, 15-9, 15-21, 15-26, 15-31, 15-33

Mode, 20, 25

Motorola, 17-2

Multiple comparisons, 668–673

National AIDS Behavioral Surveys, 345

National Assessment of Educational Progress (NAEP), 61, 73, 392

National Association of Colleges and Employers (NACE), 365, 370

National Congregations Study, 506

National Crime Victimization Survey, 713

National Football League, 102, 604

National Health and Nutrition Examination Survey (NHANES), 383, 698

National Oceanic and Atmospheric Administration (NOAA), 605

National Science Foundation (NSF), 599

National Student Loan Survey, 528

New Jersey Pick-It Lottery, 16-20–16-22

Neyman, Jerzy, 410

Nielsen Company, 421, 443

Noncentrality parameter

- for  $t$ , 478
- for  $F$ , 676
- Nonparametric procedure, 436, 438–440. *See also* Chapter 15
- Nonresponse, 198, 201
- Normal distribution. *See* Distribution, Normal
- Normal distribution calculations, 63–68, 72
- Normal probability plot. *See* Normal quantile plot
- Normal scores, 69
- Normal quantile plot, 68–70, 72
- Null hypothesis. *See* Hypothesis, null
  
- Observational study, 172
- Odds, 632, 14-2, 14-5, 14-19
- Odds ratio, 632, 14-7, 14-19
- Outcomes, 175, 188
- Out-of-control rules, 17-24–17-26
- Outliers, 20, 21, 25, 15-1
  - $1.5 \times IQR$  criterion, 39–40, 48
  - regression, 130–131
  
- Parameter, 206, 215
- Pareto chart, 17-18, 17-54, 17-83
- Pearson, Egon, 410
- Pearson, Karl, 234
- Percent, 10, 487
- Percentile, 35
- Permutation tests, 16-42–16-52
- Pew survey 488, 504, 510, 523, 525, 559, 14-2, 14-20, 15-15, 16-55, 16-57
- Pie chart, 12, 25
- Placebo effect, 178
- Plug-in principle, 16-9, 16-10
- Pooled estimator
  - of population proportion, 517
  - of ANOVA variance, 654, 659
  - of variance in two samples, 461
- Population, 171, 192, 201, 206
- Population distribution. *See* Distribution, population
- Power, 402–404, 411

- and Type II error, 410
- increasing, 405–406
- of one-way ANOVA,
- of  $t$  test
  - one-sample, 434–435
  - two-sample, 477–479
- of  $z$  test, 402–405
- Prediction, 110, 112, 121
- Prediction interval, 578–580, 585, 617
- Probability, 233, 235
  - conditional, 286–288, 294
  - equally likely outcomes, 243–244
  - finite sample space, 242
- Probability distribution, 253, 259, 260
  - mean of, 264–265, 272
  - standard deviation of, 274, 275–276
  - variance of, 272–276
- Probability histogram, 254
- Probability model, 237, 248
- Probability rules, 240
  - addition, 240, 249, 283, 285, 294
  - complement, 240, 249, 283, 294
  - general, 282–286, 294
  - multiplication, 244–245, 248, 283, 294
- Probability sample. *See* Sample, probability
- Process capability indexes, 17-41–17-47
- Proportion, sample, 321, 343, 488, 503
  - distribution of, 330–332, 343, 489
  - inference for a single proportion, 488–503
  - inference for comparing two proportions, 508–521
- Punxsutawney Phil, 149–150, 525
- $P$ -value, 377
- Quartiles, 35
  - of a density curve, 57
- R, 340, 16-10, 16-11, 16-12, 16-15, 16-19, 16-35, 16-39, 16-46
- Randomization

- consequences of, 213
- experimental, 179, 188
- how to, 181–184

Random digits, 182, 188, Table B

Random number generator, 386

Random phenomenon, 233, 235

Random variable, 252–253, 260

- continuous, 256–257, 261
- discrete, 253, 260
- mean of, 264–265
- standard deviation of, 274
- variance of, 273–274

Randomized comparative experiment, 181

Randomized response survey, 299–300

Ranks, 15-4, 15-15

Rate, 6

Regression, 109–121

- and correlation, 119, 121
- cautions about, 126–136
- deviations, 567, 568, 586, 614
- interpretation, 116
- least-squares, 113–115, 121, 568, 615
- multiple, 612–618
- nonlinear, 582–584
- simple linear, 109–121, 564–599

Regression equation, population, 612

Regression line, 110, 121

- population, 565, 584

Relative risk, 520, 522

Reliability, 323

Resample, 367. *See also* Chapter 16

Residual, 126–127, 136, 569, 584, 615, 633, 652, 653

- plots, 128, 136, 572–573, 625

Resistant measure, 32, 48

Response bias, 200, 201

Response rate, 193

Response variable. *See* Variable, response

Ringtone, 196

Robustness, 32, 432–433, 455–456, 477, 15-1

Roundoff error, 26

Row variable. *See* Variable, row and column

Sallie Mae, 360

Sample, 171, 192, 201, 206

- cautions about, 198–200
- design of, 192–201
- multistage, 197–198
- probability, 196–201
- proportion, 321, 488
- simple random (SRS), 194, 201
- stratified, 196–197, 201
- systematic, 204

Sample size, choosing

- confidence interval for a mean, 363–365
- confidence interval for a proportion, 500
- one-way ANOVA, 675–677
- $t$  test, one-sample, 434–435
- $t$  test, two-sample, 477–479

Sample space, 237, 248

- finite, 242

Sample survey, 171, 174, 192, 201

Sampling distribution, 208–209, 215

- of difference of means, 449
- of regression estimators, 574
- of sample count, 325, 332, 339, 343–344
- of sample mean, 307, 316, 433
- of sample proportion, 332, 343

Sampling variability, 207

SAS, 458, 478, 492, 498, 512, 520, 571, 601, 674, 676, 704, 705, 14-17, 15-7, 15-13, 15-16, 15-22, 15-32

SAT college entrance examination, 67, 75–76, 354, 608–609, 619–622, 627–631, 635, 718

Scatterplot, 87–89, 97

- adding categorical variables to, 94–95
- smoothing, 96

Shape of a distribution, 20, 25

Shewhart, Walter, 17-7, 17-32  
 Sign test, 438–440  
 Significance level, 379, 395–398  
 Significance, statistical, 378–382  
     and Type I error, 409  
 Significance test, 372–390  
     chi-square for two-way table, 539, 550  
         relation to  $z$  test, 544  
     chi-square for goodness of fit, 552–553, 556  
      $F$  test in one-way ANOVA, 660–662  
      $F$  test in regression, 588–590, 600, 618  
      $F$  test for a collection of regression coefficients, 630–631, 637  
      $F$  test for standard deviations, 474–476  
      $F$  tests in two-way ANOVA, 703  
     Kruskal-Wallis test, 15-28–15-23  
     relationship to confidence intervals, 386–388  
      $t$  test for a contrast, 665  
      $t$  test for correlation, 597–599, 600  
      $t$  test for one mean, 422–424  
      $t$  test for matched pairs, 429–431  
      $t$  test for two means, 454, 466–467  
         pooled, 462  
      $t$  test for regression coefficients, 574, 584  
      $t$  tests for multiple comparisons, 670  
     use and abuse, 394–400  
     Wilcoxon rank sum test, 15-5  
     Wilcoxon signed rank test, 15-20  
      $z$  test for one mean, 383, 390  
      $z$  test for one proportion, 495, 504  
      $z$  test for logistic regression slope, 14-10, 14-19  
      $z$  test for two proportions, 517, 522  
      $z$  test for two means, 450, 466  
 Simple random sample. *See* Sample, simple random  
 Simpson's paradox, 146–147, 148  
 Simulation, 207  
 Simultaneous confidence intervals, 672 68–95–99.7 rule, 59–60, 72  
 Skewed distribution. *See* Distribution, skewed  
 Slope of a line, 111

- of least-squares line, 115, 121, 568
- Small numbers, law of, 269–270
- Spread of a distribution, 20, 25, 35, 42, 47
- Spreadsheet, 5. *see also* Excel
- SPSS, 117, 123, 124, 145, 426, 427, 460, 534, 554, 570, 589, 598, 656, 657, 668, 671, 14-14, 14-18, 15-9, 15-21, 15-32
- Standard & Poor's 500-Stock Index, 425–426
- Standard deviation, 42, 48 *See also* Variance
  - of binomial distribution, 329, 343
  - of density curve, 57–58, 72
  - of difference between sample means, 449–450
  - of difference between sample proportions, 509
  - of Normal distribution, 58
  - of Poisson distribution, 339, 344
  - of regression intercept and slope, 593
  - pooled
    - for two samples, 462
    - in ANOVA, 654
  - properties, 44
  - of random variable, 274, 279
  - rules for, 275–276, 279
  - of sample mean, 306, 316
  - of sample proportion, 418, 489
- Standard error, 418
  - bootstrap, 16-6, 16-8–16-9
  - of a contrast, 665
  - of a difference of sample proportions, 510, 521
  - for regression prediction, 595, 600
  - of regression intercept and slope, 593, 600
  - of mean regression response, 595, 600
  - of a sample mean, 418, 440
  - of a sample proportion, 418, 489, 503
- Standard Normal distribution. *See* Distribution, standard Normal
- Standardized observation, 61, 72
- Statistic, 206, 215
- Statistical inference, 167, 205–213, 352–353
  - for Nonnormal populations, 436–440. *See also* Chapter 15
  - for small samples, 457–460

Statistical process control, Chapter 17  
Statistical significance. *See* Significance, statistical  
Stem-and-leaf plot. *See* Stemplot  
Stemplot, 13, 25  
    back-to-back, 14  
    splitting stems, 14  
    trimming, 14  
Strata, 197, 201. *See also* Sample, stratified  
Strength of a relationship, 89, 98. *See also* Correlation  
StubHub! 71–72, 16-12, 16-23  
Student Monitor, 370  
Subjects, experimental, 175, 188  
Subpopulation, 565, 612–613  
Sums of squares  
    in one-way ANOVA, 658–659  
    in two-way ANOVA, 702–703  
    in multiple linear regression, 617  
    in simple linear regression, 586–587  
Survey of Study Habits and Attitudes (SSHA), 393  
Systematically larger, 15-10  
Symmetric distribution. *See* Distribution, symmetric  
  
*t* distribution. *See* Distribution, *t*  
*t* inference procedures  
    for contrasts, 665  
    for correlation, 597  
    for matched pairs, 429–431  
    for multiple comparisons, 670  
    for one mean, 421, 423  
    for two means, 450–454  
    for two means, pooled, 461–462  
    for regression coefficients, 574, 616  
    for regression mean response, 577  
    for regression prediction, 579  
    robustness of, 432–433, 455–456  
Tails of a distribution. *See* Distribution, tails  
Test of significance. *See* Significance test  
Test statistic, 375–376

Testing hypotheses. *See* Significance test  
 The Times Higher Education Supplement, 638  
 Three-way table, 148  
 Ties, 15-10–15-11  
 Time plot, 23–24, 25  
 Titanic, 25, 54, 149, 157, 16-12, 16-23  
 Transformation, 93  
     linear, 45–47, 48  
     logarithm, 93, 436, 582  
     rank, 15-4  
 Treatment, experimental, 172, 174, 175, 178, 188  
 Tree diagram, 290–291, 294  
 Tuskegee study, 222–223  
 Twitter, 25–26, 261, 525  
 Two-sample problems, 448  
 Two-way table, 139–140, 148, 530  
     data analysis for, 139–148  
     inference for, 530–550  
     models for, 545–548, 550  
     relationships in 143–144  
 Type I and II errors, 407–408  
  
 Unbiased estimator, 210–211, 215  
 Undercoverage, 198, 201  
 Unimodal distribution. *See* Distribution, unimodal  
 Union of events, 283, 294  
 Unit of measurement, 3, 45  
 Unit, experimental, 175  
 U.S. Agency for International Development, 15-27  
 U.S. Department of Education, 346  
  
 Value of a variable, 2, 8  
 Variability, 47, 211  
 Variable, 2, 8  
     categorical, 3, 8, 97, 487  
     dependent, 86  
     explanatory, 84, 86, 97  
     independent, 86

- lurking, 133, 136, 176
- quantitative, 3, 8
- response, 84, 86
- row and column, 140, 148
- Variance, 42, 48
  - of a difference between two sample means, 449
  - of a difference between two sample proportions, 509
  - of a random variable, 273–274, 279
  - a pooled estimator, 462, 467
  - rules for, 275–276, 279
  - of a sample mean, 306
- Variation
  - among groups, 658, 678
  - between groups, 647, 678
  - common cause, 17-7
  - special cause, 17-7
  - within group, 647, 658, 678,
- Venn diagram, 240
- Voluntary response, 194
  
- Wald statistic, 14-10, 14-20
- Whiskers, 38
- Wilcoxon rank sum test, 15-3–15-15
- Wilcoxon signed rank test, 15-18–15-25
- Wording questions, 200, 201
- World Bank, 31, 78, 100, 16-3
- World Database of Happiness, 638
  
- z -score, 61, 72
- z statistic
  - for one proportion, 495
  - for two proportions, 517
  - one-sample for mean, 419, 440, 440
  - two-sample for means, 448–450



[www.launchpadportal.com](http://www.launchpadportal.com)

At W. H. Freeman, we've taken what we've learned from thousands of statistics instructors and students to create a new generation of W. H. Freeman/Macmillan technology. The new online homework system, LaunchPad, offers Freeman's acclaimed content, curated and organized for easy assignability in a simple but powerful interface.

**Available in LaunchPad:**

**LEARNINGCurve**

[www.learningcurveworks.com](http://www.learningcurveworks.com)

**LearningCurve** provides students and instructors with powerful adaptive quizzing in a game-like format that features direct links to the e-Book and instant feedback. Questions are tailored specifically to the text and adapt to the student's responses, varying difficulty level and topic coverage based on student performance.

**CRUNCH IT!**

[www.whfreeman.com/crunchit](http://www.whfreeman.com/crunchit)

**Crunchit!** A web-based statistical program that allows users to perform all of the statistical operations and graphing needed for an introductory statistics course and more. It saves users time by automatically loading data from W. H. Freeman's statistics textbooks, and provides the flexibility to edit and import additional data.

**StatTutors** These multimedia tutorials explore important concepts and procedures in a presentation that combines video, audio, and interactive features.

**Statistical Applets** These activities give students hands-on opportunities to familiarize themselves with important statistical concepts and procedures, in an interactive setting that allows them to manipulate variables and see the results graphically. Icons in the textbooks indicate when an applet is available for the material being covered.

**Stepped Tutorials** These new exercise tutorials (2-3 per chapter) are easily assignable and assessable. The tutorials are centered on algorithmically-generated quizzing with step-by-step feedback to help students work their way toward the correct solution.

**About the Cover:**

Made from thousands of Italian countertop samples, the cover image presents the data of the artist's changing mood as tracked by [www.moodjam.com](http://www.moodjam.com). *IPS* explores mood by looking at how stress and lack of sleep are linked in a series of examples in Chapter 2.



W. H. Freeman and Company  
41 Madison Avenue, New York, NY 10010  
Houndmills, Basingstoke RG21 6XS, England  
Cover Design: Vicki Tomaselli  
Cover Image: Moodjam by Laurie Frick

