# IBM SPSS Statistics 23 Step by Step

## A Simple Guide and Reference

FOURTEENTH EDITION

**Darren George**
*Burman University*
*(Formerly Canadian University College)*

**Paul Mallery**
*La Sierra University*

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

Fourteenth edition published 2016
by Routledge
711 Third Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

*Routledge is an imprint of the Taylor & Francis Group, an informa business*

Thirteenth edition published by Pearson 2014.

*Library of Congress Cataloging in Publication Data*
A catalog record for this book has been requested.

ISBN: 978-1-138-68134-7 (hbk)
ISBN: 978-0-13-432025-0 (pbk)
ISBN: 978-1-315-54589-9 (ebk)

Typeset in Palatino LT Pro 9.5/13
by Lumina Datamatics, Inc.

# IBM SPSS Statistics 23 Step by Step

*IBM SPSS Statistics 23 Step by Step: A Simple Guide and Reference, 14e,* takes a straightforward, step-by-step approach that makes SPSS software clear to beginners and experienced researchers alike. Extensive use of vivid, four-color screen shots, clear writing, and step-by-step boxes guide readers through the program. Exercises at the end of each chapter support students by providing additional opportunities to practice using SPSS.

All datasets used in the book are available for download at:
www.routledge.com/9780134320250

# Contents

# Preface

SPSS is a powerful tool that is capable of conducting just about any type of data analysis used in the social sciences, the natural sciences, or in the business world. While mathematics is generally thought to be the language of science, data analysis is the language of research. Research in many fields is critical for human progress, and as long as there is research, there will be the need to analyze data. The present book is designed to make data analysis more comprehendible and less toxic.

In our teaching, we have frequently encountered students so traumatized by the professor who cheerily says "Analyze these data on SPSS; get the manuals if you don't know how" that they dropped the course rather than continue the struggle. It is in response to this anguish that the present book was conceived. Darren George's background has been teaching high school mathematics, and Paul Mallery worked his way through college training people to use computers and programming computers. Both of us find great pleasure in the challenge of making a process that is intrinsically complex as clear as possible. The ultimate goal in all our efforts with the present book has been to make SPSS procedures, above all else, clear.

As the book started to take shape, a second goal began to emerge. In addition to making SPSS procedures clear to the beginner, we wanted to create a tool that was an effective reference for anyone conducting data analysis. This involved the expansion of the original concept to include most of the major statistical procedures that SPSS covers in the base module and many of the procedures in the advanced and regression modules as well. The result of years of effort you now hold in your hands.

The 14th edition is improved in both content and style. Most visibly, the new design and the switch from black-and-white to color make the text easier to use, and make the SPSS screens and output in the book more user-friendly. But more importantly than the improvements in style, changes in this edition have incorporated new content, including a much more extensive discussion of effect size throughout the entire book and a complete rewrite of the Nonparametrics chapter (Chapter 17).

Effect size: While effect size was mentioned in several chapters in prior editions, the discussion of this topic now allows the professor to assist students in a comprehensive understanding of this important issue across a number of different procedures. It includes a discussion of Carmer's *V* for chi-square ($\chi^2$) analysis, *r* and $r^2$ for correlations (in which the test statistic and the measure of effect size are the same), Cohen's *d* for *t* tests, eta ($\eta$) and eta-squared ($\eta^2$) for ANOVA, and *R* and $R^2$ for regression.

The rewrite of the Nonparametrics chapter reflects both a dramatic shift in the way SPSS handles nonparametric procedures, an introduction that provides a discussion of when nonparametric tests should be used (and when they might be avoided), and a more complete coverage of the material. Rather than simply covering nine different procedures (as in previous editions), five broad categories of nonparametric procedures are addressed: Do observed values differ from a hypothesized distribution? Is the order of observed values non-random? Is a continuous variable dissimilar in different groups? Do the medians differ for different groups? And, do my within-subjects measurements differ?

In the 22nd and 23rd releases, SPSS has cleaned up some bugs in the graphing programs, and there are a few access procedures that are more intuitive than in prior editions. As usual, all screens have been updated, all step-by-step sequences executed, and all outputs scrutinized to make certain everything in the current edition is accurate.

While the first 16 chapters of the book cover basic topics and would be understandable to many with very limited statistical background, the final 12 chapters involve procedures that progressively require a more secure statistical grounding. Those 12 chapters have provided our greatest challenge. At the beginning of each chapter we spend several pages describing the procedure that follows. But, how can one adequately describe, for instance, factor analysis or discriminant analysis in five or six pages? The answer is simple: We can't, but we can describe the procedures at a common sense, conceptual level that avoids excessive detail and excessive emphasis on computation that is useful as an introduction for beginners or as a useful adjunct to more advanced reading or mentoring for more advanced data analysts. Writing these introductions has not at all been simple. The chapter introductions are the most painstakingly worked sections of the entire book. Although we acknowledge the absence of much detail in our explanation of most procedures, we feel that we have done an adequate job at a project that few would even attempt. How successful have we been at achieving clarity in very limited space? The fact that this book is now in its 14th edition, has been an academic best seller for most of those editions, and is distributed in 85 countries of the world suggests that our efforts have not been in vain.

## Authors' Biographical Sketches and Present Addresses

Darren George is currently a professor of Psychology at (formerly Canadian University College)

Burman University
5415 College Avenue
Lacombe, AB, T4L 2E5
403-782-3381, Ext. 4082
dgeorge@burmanu.ca

where he teaches personality psychology, social psychology, and research methods. He completed his MA in Experimental Psychology (1982) at California State University, Fullerton; taught high school mathematics for nine years (1980–1989) at Mark Keppel High School (Alhambra, CA) and Mountain View High School (El Monte, CA), and then completed a Psychology PhD at UCLA (1992) with emphases in personality psychology, social psychology, and measurement and psychometrics. Darren has now been a professor at Burman University for 22 years.

Paul Mallery is currently a professor of Psychology at

La Sierra University
4500 Riverwalk Parkway
Riverside, CA, 92515
951-785-2528
pmallery@lasierra.edu

where he teaches social psychology and related courses and experimental methodology (including the application of SPSS). He received his PhD in Social Psychology from UCLA (1994), with emphases in statistics and political psychology. Paul formerly worked as a computer specialist, both programming and teaching computer usage.

## Acknowledgments

As we look over the creative efforts of the past years, we wish to acknowledge several people who have reviewed our work and offered invaluable insight and suggestions for improvement. Our gratitude is extended to Richard Froman of John Brown University, Michael A. Britt of Marist College, Marc L. Carter of the University of South Florida, Randolph A. Smith of Ouachita Baptist University, Roberto R. Heredia of Texas A&M International University, and several anonymous reviewers. And then there's the standard (but no less appreciated) acknowledgment of our families and friends who endured while we wrote this. Particular notice goes to our wives Elizabeth George and Suzanne Mallery as well as our families for their support and encouragement.

Chapter 1

# An Overview of IBM® SPSS® Statistics

## Introduction: An Overview of IBM SPSS Statistics 23

THIS BOOK gives you the step-by-step instructions necessary to do most major types of data analysis using SPSS. The software was originally created by three Stanford graduate students in the late 1960s. The acronym "SPSS" initially stood for "**S**tatistical **P**ackage for the **S**ocial **S**ciences." As SPSS expanded their package to address the hard sciences and business markets, the name changed to "Statistical Product and Service Solutions." In 2009 IBM purchased SPSS and the name morphed to **"IBM SPSS Statistics**." SPSS is now such a standard in the industry that IBM has retained the name due to its recognizability. No one particularly cares what the letters "SPSS" stand for any longer. IBM SPSS Statistics is simply one of the world's largest and most successful statistical software companies. In this book we refer to the program as **SPSS**.

## 1.1 Necessary Skills

For this book to be effective when you conduct data analysis with SPSS, you should have certain limited knowledge of statistics and have access to a computer that has the necessary resources to run SPSS. Each issue is addressed in the next two paragraphs.

**STATISTICS** You should have had at least a basic course in statistics or be in the process of taking such a course. While it is true that this book devotes the first two or three pages of each chapter to a description of the statistical procedure that follows, these descriptions are designed to refresh the reader's memory, *not* to instruct the novice. While it is certainly possible for the novice to follow the steps in each chapter and get SPSS to produce pages of output, a fundamental grounding in statistics is important for an understanding of which procedures to use and what all the output means. In addition, while the first 16 chapters should be understandable by individuals with limited statistical background, the final 12 chapters deal with much more complex and involved types of analyses. These chapters require substantial grounding in the statistical techniques involved.

**COMPUTER REQUIREMENTS** You must:

- Have access to a personal computer that has
  - Microsoft® Windows Vista® or Windows® 7 or 8.1 or 10; MAC OS® 10.8 (Mountain Lion) or higher installed
  - IBM SPSS Statistics 23.0 installed
- Know how to turn the computer on

- Have a working knowledge of the keys on the keyboard and how to use a mouse—or other selection device such as key board strokes or touch screen monitors.

This book will take you the rest of the way. If you are using SPSS on a network of computers (rather than your own PC or MAC) the steps necessary to *access* IBM SPSS Statistics may vary slightly from the single step shown in the pages that follow.

## 1.2 Scope of Coverage

IBM SPSS Statistics is a complex and powerful statistical program by any standards. The software occupies about 800 MB of your hard drive and requires at least 1 GB of RAM to operate adequately. Despite its size and complexity, SPSS has created a program that is not only powerful but is user friendly (you're the user; the program tries to be friendly). By improvements over the years, SPSS has done for data analysis what Henry Ford did for the automobile: made it available to the masses. SPSS is able to perform essentially any type of statistical analysis ever used in the social sciences, in the business world, and in other scientific disciplines.

This book was written for Version 23 of IBM SPSS Statistics. More specifically, the screen shots and output are based on Version 23.0. With some exceptions, what you see here will be similar to SPSS Version 7.0 and higher. Because only a few parts of SPSS are changed with each version, most of this book will apply to previous versions. It's 100% up-to-date with Version 23.0, but it will lead you astray only about 2% of the time if you're using Version 21.0 or 22 and is perhaps 60% accurate for Version 7.0 (if you can find a computer and software that old).

Our book covers the statistical procedures present in three of the *modules* created by SPSS that are most frequently used by researchers. A module (within the SPSS context) is simply a set of different statistical operations. We include the **Base Module** (technically called **IBM SPSS Statistics Base**), the module covering advanced statistics (**IBM SPSS Advanced Statistics**), and the module that addresses regression models (**IBM SPSS Regression**)—all described in greater detail later in this chapter. To support their program, SPSS has created a set of comprehensive manuals that cover all procedures these three modules are designed to perform. To a person fluent in statistics and data analysis, the manuals are well written and intelligently organized. To anyone less fluent, however, the organization is often undetectable, and the comprehensiveness (the equivalent of almost 2,000 pages of fine-print text) is overwhelming. To the best of our knowledge, hard-copy manuals are no longer available but most of this information may now be accessed from SPSS as PDF downloads. The same information is also available in the exhaustive online Help menu. Despite changes in the method of accessing this information, for sake of simplicity we still refer to this body of information as "SPSS manuals" or simply "manuals." Our book is about 400 pages long. Clearly we cannot cover in 400 pages as much material as the manuals do in 2,000, but herein lies our advantage.

The purpose of this book is to make the fundamentals of most types of data analysis clear. To create this clarity requires the omission of much (often unnecessary) detail. Despite brevity, we have been keenly selective in what we have included and believe that the material presented here is sufficient to provide simple instructions that cover 95% of analyses ever conducted by researchers. Although we cannot substantiate that exact number, our time in the manuals suggests that at least 1,600 of the 2,000 pages involve detail that few researchers ever consider. How often do you really need 7 different methods of extracting and 6 methods of rotating factors in factor analysis, or 18 different methods for post-hoc comparisons after a one-way ANOVA? (By the way, that last sentence should be understood by statistical geeks only.)

We are in no way critical of the manuals; they do well what they are designed to do and we regard them as important adjuncts to the present book. When our space limitations prevent explanation of certain details, we often refer our readers to the SPSS manuals. Within the context of presenting a statistical procedure, we often show a window that includes several options but describe only one or two of them. This is done without apology except for the occasional "description of these options extends beyond the scope of this book" and cheerfully refer you to the appropriate SPSS manual. The ultimate goal of this format is to create clarity without sacrificing necessary detail.

## 1.3 Overview

This chapter introduces the major concepts discussed in this book and gives a brief overview of the book's organization and the basic tools that are needed in order to use it.

If you want to run a particular statistical procedure, have used IBM SPSS Statistics before, and already know which analysis you wish to conduct, you should read the Typographical and Formatting Conventions section in this chapter (pages 5–7) and then go to the appropriate chapter in the last portion of the book (Chapters 6 through 28). Those chapters will tell you exactly what steps you need to perform to produce the output you desire.

If, however, you are new to IBM SPSS Statistics, then this chapter will give you important background information that will be useful whenever you use this book.

## 1.4 This Book's Organization, Chapter by Chapter

This book was created to describe the crucial concepts of analyzing data. There are three basic tasks associated with data analysis:

A. You must type data into the computer, and organize and format the data so both SPSS and you can identify it easily,

B. You must tell SPSS what type of analysis you wish to conduct, and

C. You must be able to interpret what the SPSS output means.

After this introductory chapter, Chapter 2 deals with basic operations such as types of SPSS windows, the use of the toolbar and menus, saving, viewing, and editing the output, printing output, and so forth. While this chapter has been created with the beginner in mind, there is much SPSS-specific information that should be useful to anyone. Chapter 3 addresses the first step mentioned above—creating, editing, and formatting a data file. The SPSS data editor is an instrument that makes the building, organizing, and formatting of data files wonderfully clear and straightforward.

Chapters 4 and 5 deal with two important issues—modification and transformation of data (Chapter 4) and creation of graphs or charts (Chapter 5). Chapter 4 deals specifically with different types of data manipulation, such as creating new variables, reordering, restructuring, merging files, or selecting subsets of data for analysis. Chapter 5 introduces the basic procedures used when making a number of different graphs; some graphs, however, are described more fully in the later chapters.

Chapters 6 through 28 then address Steps B and C—analyzing your data and interpreting the output. It is important to note that each of the analysis chapters is self-contained. If the beginner, for example, were instructed to conduct *t* tests on certain data, Chapter 11 would give complete instructions for accomplishing that procedure. In the Step by Step section, Step 1 is always "start the SPSS program" and refers the reader to Chapter 2 if there are questions about how to do this. The second step is always "create a data file or edit (if necessary) an already existing file," and the reader is then referred to Chapter 3 for instructions if needed. Then the steps that follow explain exactly how to conduct a *t* test.

As mentioned previously, this book covers three modules produced by SPSS: **IBM SPSS Statistics Base, IBM SPSS Advanced Statistics**, and **IBM SPSS Regression**. Since some computers at colleges or universities may not have all of these modules (the **Base** module is always present), the book is organized according to the structure SPSS has imposed: We cover almost all procedures included in the **Base** module and then selected procedures from the more complex **Advanced** and **Regression** Modules. Chapters 6–22 deal with processes included in the Base module. Chapters 23–27 deal with procedures in the Advanced Statistics and Regression Modules, and Chapter 28, the analysis of residuals, draws from all three.

**IBM SPSS STATISTICS BASE,** Chapters 6 through 10 describe the most fundamental data analysis methods available, including frequencies, bar charts, histograms, and percentiles (Chapter 6); descriptive statistics such as means, medians, modes, skewness, and ranges (Chapter 7); crosstabulations and chi-square tests of independence (Chapter 8); subpopulation means (Chapter 9); and correlations between variables (Chapter 10).

The next group of chapters (Chapters 11 through 17) explains ways of testing for differences between subgroups within your data or showing the strength of relationships between a dependent variable and one or more independent variables through the use of t tests (Chapter 11); ANOVAs (Chapters 12, 13, and 14); linear, curvilinear, and multiple regression analysis (Chapters 15 and 16); and the most common forms of nonparametric tests are discussed in Chapter 17.

Reliability analysis (Chapter 18) is a standard measure used in research that involves multiple response measures; multidimensional scaling is designed to identify and model the structure and dimensions of a set of stimuli from dissimilarity data (Chapter 19); and then factor analysis (Chapter 20), cluster analysis (Chapter 21), and discriminant analysis (Chapter 22) all occupy stable and important niches in research conducted by scientists.

**IBM SPSS ADVANCED STATISTICS AND REGRESSION:** The next series of chapters deals with analyses that involve multiple dependent variables (SPSS calls these procedures General Linear Models; they are also commonly called MANOVAs or MANCOVAs). Included under the heading General Linear Model are simple and general factorial models and multivariate models (Chapter 23), and models with repeated measures or within-subjects factors (Chapter 24).

The next three chapters deal with procedures that are only infrequently performed, but they are described here because when these procedures are needed they are indispensable. Chapter 25 describes logistic regression analysis and Chapters 26 and 27 describe hierarchical and nonhierarchical log-linear models, respectively. As mentioned previously, Chapter 28 on residuals closes out the book.

## 1.5 An Introduction to the Example

A single data file is used in 17 of the first 19 chapters of this book. For more complex procedures it has been necessary to select different data files to reflect the particular procedures that are presented. Example data files are useful because often, things that appear to be confusing in the SPSS documentation become quite clear when you see an example of how they are done. Although only the most frequently used sample data file is described here, there are a total of 12 data sets that are used to demonstrate procedures throughout the book, in addition to data sets utilized in the exercises. Data files are available for download at *www.spss-step-by-step.net*. These files can be of substantial benefit to you as you practice some of the processes presented here without the added burden of having to input the data. We suggest that you make generous use of these files by trying different procedures and then comparing your results with those included in the output sections of different chapters.

The example has been designed so it can be used to demonstrate most of the statistical procedures presented here. It consists of a single data file used by a teacher who teaches three sections of a class with approximately 35 students in each section. For each student, the following information is recorded:

- ID number
- Name
- Gender
- Ethnicity
- Year in school
- Upper- or lower-division class person
- Previous GPA
- Section
- Whether or not he or she attended review sessions or did the extra credit
- The scores on five 10-point quizzes and one 75-point final exam

In Chapter 4 we describe how to create four new variables. In all presentations that follow (and on the data file available on the website), these four variables are also included:

- The total number of points earned
- The final percent
- The final grade attained
- Whether the student passed or failed the course

The example data file (the entire data set is displayed at the end of Chapter 3) will also be used as the example in the introductory chapters (Chapters 2 through 5). If you enter the data yourself and follow the procedures described in these chapters, you will have a working example data file identical to that used through the first half of this book. Yes, the same material is recorded on the downloadable data files, but it may be useful for you to practice data entry, formatting, and certain data manipulations with this data set. If you have your own set of data to work with, all the better.

One final note: All of the data in the **grades** file are totally fictional, so any findings exist only because we created them when we made the file.

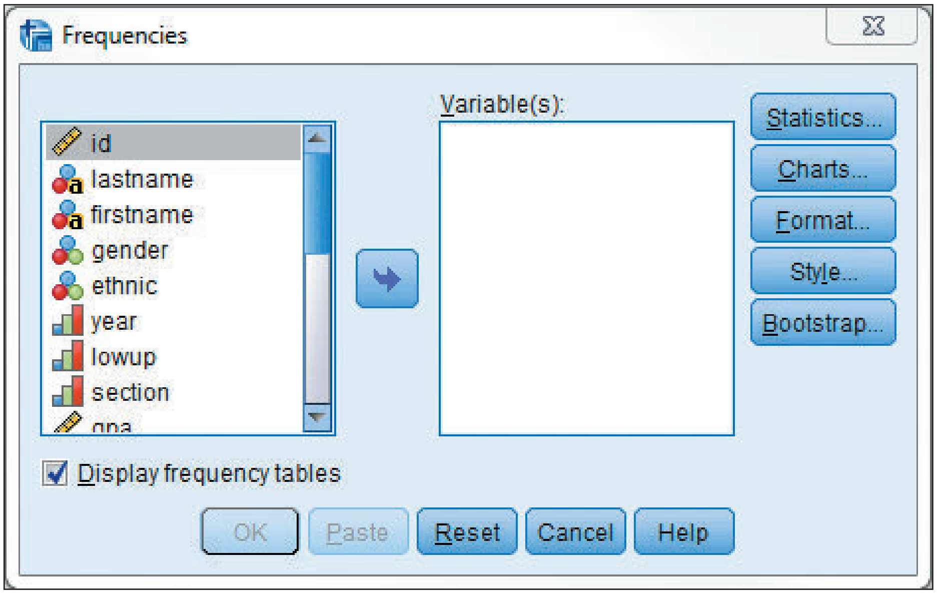
## 1.6 Typographical and Formatting Conventions

**CHAPTER ORGANIZATION** Chapters 2 through 5 describe IBM SPSS Statistics formatting and procedures, and the material covered dictates each chapter's organization. Chapters 6 through 28 (the analysis chapters) are, with only occasional exceptions, organized identically. This format includes:

1. The **Introduction** in which the procedure that follows is described briefly and concisely. These introductions vary in length from one to seven pages depending on the complexity of the analysis being described.
2. The **Step by Step** section in which the actual steps necessary to accomplish particular analyses are presented. Most of the typographical and formatting conventions described in the following pages refer to the Step by Step sections.
3. The **Output** section, in which the results from analyses described earlier are displayed—often abbreviated. Text clarifies the meaning of the output, and all of the critical output terms are defined.

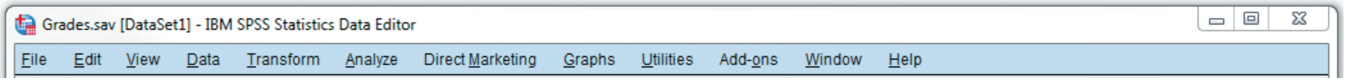
**THE SCREENS** Due to the very visual nature of SPSS, every chapter contains pictures of screens or windows that appear on the computer monitor as you work. The first picture from Chapter 6 (below) provides an example. These pictures are labeled "Screens" despite the fact that sometimes what is pictured is a screen (everything that appears on the monitor at a given time) and other times is a portion of a screen (a window, a dialog box, or something smaller). If the reader sees reference to Screen 13.3, she knows that this is simply the third picture in Chapter 13. The screens are typically positioned within breaks in the text (the screen icon and a title are included) and are used for sake of reference as procedures involving that screen are described. Sometimes the screens are separate from the text and labels identify certain characteristics of the screen (see the inside front cover for an example). Because screens take up a lot of space, frequently used screens are included on the inside front and back covers of this book. At other times, within a particular chapter, a screen from a different chapter may be cited to save space.

**Screen 1.1** The Frequencies Window



☑ Display frequency tables

Sometimes a portion of a screen or window is displayed (such as the menu bar included here) and is embedded within the text without a label.



The Step by Step boxes: Text that surrounds the screens may designate a procedure, but it is the Step by Step boxes that identify exactly what must be done to execute a procedure. The following box illustrates:

| In Screen | Do This | | Step 3 (sample) |
|---|---|---|---|
| Front1 | ✱ **File** → **Open** → ✱ **Data** | [*or* ✱ 📂] | |
| Front2 | **type** `grades.sav` → ✱ **Open** | [*or* ✱ ✱ **grades.sav**] | **Data** |

Sequence Step 3 means: "Beginning with Screen 1 (displayed on the inside front cover), click on the word **File,** move the cursor to **Open,** and then click the word **Data**. At this point a new window will open (Screen 2 on the inside front cover); type 'grades.sav' and then click the **Open** button, at which point a screen with your data file opens." Notice that within brackets shortcuts are sometimes suggested: Rather than the **File** → **Open** → **Data** sequence, it is quicker to click the 📂 icon. Instead of typing grades.sav and then clicking **Open**, it is quicker to double click on the **grades.sav** (with or without the **".sav"** suffix; this depends on your settings) file name. Items within Step by Step boxes include:

**Screens**: A small screen icon will be placed to the left of each group of instructions that are based on that screen. There are three different types of screen icons:

| Type of Screen Icon | Example Icon | Description of Example |
|---|---|---|
| Inside Cover Screens | **Front1** | Screen #1 on the inside front cover |
| General Screens | **Menu** | Any screen with the menu bar across the top |
| | **Graph** | Any screen that displays a graph or chart |
| Chapter Screens | **4.3** | The third screen in Chapter 4 |
| | **21.4** | The fourth screen in Chapter 21 |

Other images with special meaning inside of Step by Step boxes include:

| Image | What it Means |
|---|---|
| ✳ | A single click of the left mouse button (or select by touch screen or key strokes) |
| ✳ ✳ | A double-click of the left mouse button (or select by touch screen or key strokes) |
| **type** | A "type" icon appears before words that need to be typed |
| **press** | A "press" icon appears when a button such as the **TAB** key needs to be pressed |
| → | Proceed to the next step. |

Sometimes fonts can convey information, as well:

| Font | What it Means |
|---|---|
| `Monospaced font`<br>(Courier) | Any text within the boxes that is rendered in the Courier font represents text (numbers, letters, words) to be typed into the computer (rather than being clicked or selected). |
| *Italicized text* | *Italicized* text is used for information or clarifications within the Step by Step boxes. |
| **Bold font** | The **bold font** is used for words that appear on the computer screen. |

The groundwork is now laid. We wish you a pleasant journey through the exciting and challenging world of data analysis!

## Chapter 2A

# IBM SPSS Statistics Processes for PC

WE MENTIONED in the introductory chapter that it is necessary for the user to understand how to turn the computer on and get as far as the Windows desktop. This chapter will give you the remaining skills required to use SPSS for Windows: how to use the mouse, how to navigate using the taskbar, what the various buttons (on the toolbar and elsewhere) do, and how to navigate the primary windows used in SPSS.

If you are fluent with computers, you may not need to read this chapter as carefully as someone less familiar. But everyone should read at least portions of this chapter carefully; it contains a great deal of information unique to SPSS for Windows.

## 2.1 The Mouse

Over the years SPSS has modified their product so that mouse operations parallel those of many major programs. The left-mouse button **point and click**, **double click**, and **dragging** operate in ways similar to major word processing programs, although sometimes SPSS has unique responses to these common applications. Important differences will be noted in the chapters where they apply. A **right-click** and running the cursor over a word or object also produces similar results. In early editions of the book we provided a thorough description of mouse operations. Now, the computer world is moving in a direction where one day the mouse may be obsolete—touch screens, key-stroke operations, and other selection devices may one day predominate. Because of this we have shifted our former "mouse-click" icon (🖱) to an icon designed to mean "select" (↖). If you are operating with a mouse this icon still means "left mouse click."

## 2.2 The Taskbar and Start Menu

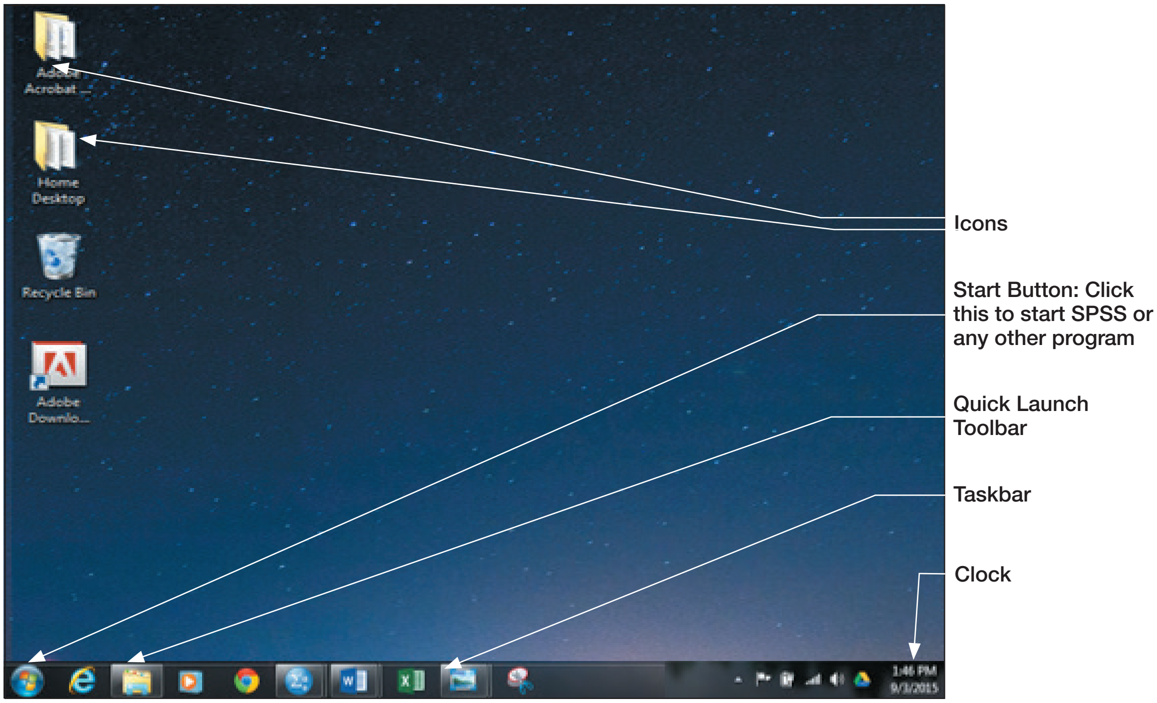
Once you have arrived at the Windows desktop, your screen should look something like that shown on the following page. It will certainly not look exactly like this, but it will be similar. There will typically be a number of icons along the left side of the screen, and a bar across the bottom (or top) of the screen (with the word "Start" on the left, and the time on the right).

There are two main types of icons on the Windows desktop: **Program** icons represent a particular program, while **folder** icons actually contain other icons (usually several programs that are related in some way).

The most important thing you need to know about the Windows desktop (at least as long as you are reading this book) is how to start the SPSS program. To do this on most computers, you need to click the ⊞ button, move the cursor over the ▸ All Programs menu folder, and then over the IBM SPSS Statistics program icon. Once IBM SPSS Statistics emerges, click on the icon, and SPSS will begin. On most computers, the Windows desktop will look similar to that shown in Screen 2.2 (following page) immediately before you click on IBM SPSS Statistics.

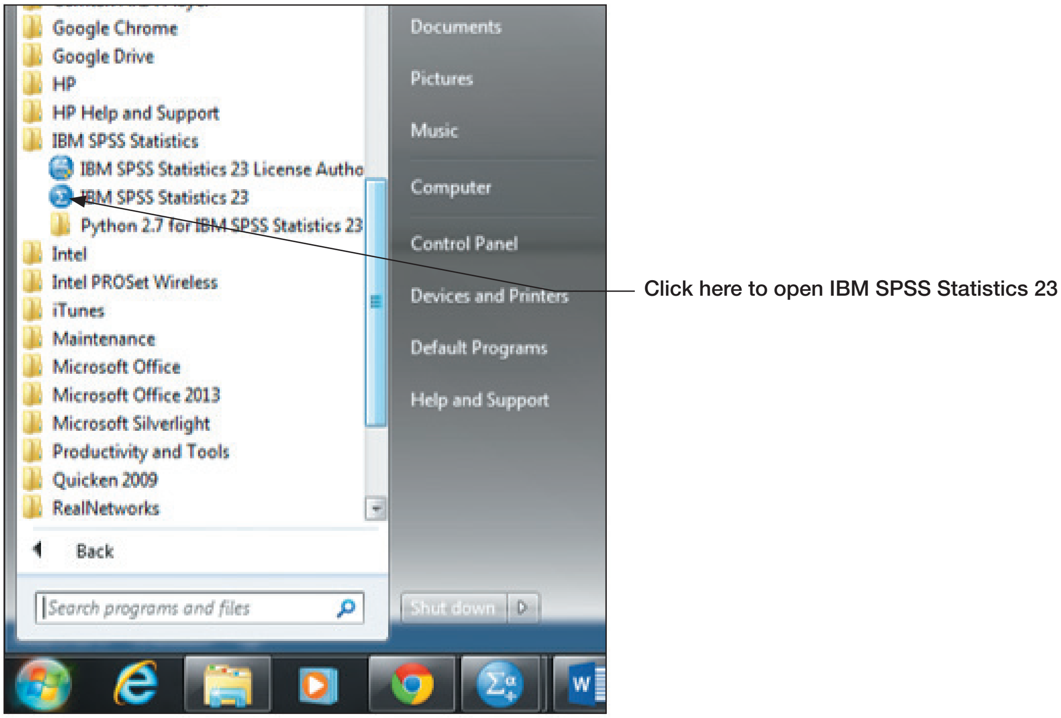
Reprinted with permission of Apple Inc. and Reprinted with permission of Microsoft, Inc.

**Screen 2.1** Windows desktop



One word of warning: On some computers, the SPSS program icon may be in a different location within the Start menu. You may have to move the cursor around the Start menu (look especially for any folders labeled "IBM" or "SPSS"). Occasionally, the icon is on the Windows desktop (along the left side), and you don't have to use the Start button at all.

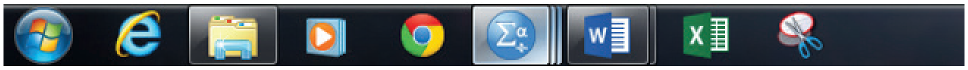
**Screen 2.2** View of the Windows desktop Start menu immediately before clicking to start the SPSS program



In addition to starting the SPSS program, the other important required skill when using the Taskbar is changing between programs. This is especially important because SPSS is actually a collection of several programs. When you first start the SPSS system, the Data Editor program is started, but as soon as you perform some statistics on your data, another program (in another window) is started: the Output program. You may have other SPSS windows open, but the Data Editor and Output windows are the main ones you will use.

Sometimes, SPSS will change between windows automatically; other times, you may need to change between the windows yourself. This is the other option that the taskbar provides.

When several programs are running at the same time, each program will have a button on the taskbar. Here's what the taskbar may look like when the SPSS data editor and SPSS output programs are both running:



If you want to change from one program window to the other, simply move the cursor down to the taskbar and click the appropriate button. Often, when many programs are open, several data sets or output screens may be concealed under a single SPSS icon. A click identifies which programs are available.

## 2.3 Common Buttons

A number of buttons occur frequently within screens and around the borders of windows and screens. The standard push buttons have the same function in any context. The most frequently used ones are identified and explained below:

| Icon | Description |
|---|---|
| □ | This is the maximize button and is located in the upper right corner of a window. Several of these buttons may be visible at one time if several windows are open. A mouse click on this button will maximize the window, that is, expand the designated window to fill the entire screen. |
| ▁ | This is the minimize button and is usually located to the immediate left of its brother. Its function is to minimize the designated window causing it to entirely disappear or to be represented by a small icon. |
| ❐ | This is called the restore button. A click on this button will reduce the size of the window with which it is associated. Sometimes this reduction will result in additional windows becoming visible. This allows you to see and then click on whichever window you wish to use. |
| ➔ | Within the SPSS context, this button will usually paste (move text from one location to another) a designated variable from a list in one location to an active box in another location. Such a move indicates that this variable is to be used in an analysis. |
| ⬅ | This button is the reverse. It will move a variable from the active box back to the original list. |
| ▲ ▼<br>▶ ◀ | These are up, down, right, and left scroll arrows that are positioned on each end of a scroll bar. To illustrate, see Screen 2.3 in which both vertical and horizontal scroll bars are identified. Most data or output files are longer than the open window and in many instances also wider. The scroll arrows and scroll bars allow you to view either above, below, to the right, or to the left of what is visible on the screen. |
| ▼ | A click on a down arrow will reveal a menu of options (called a drop-down menu) directly below. Screen 2.4 (page 13) shows two such buttons to the right of **Look in** and **Files of type**. |

## 2.4 The Data and Other Commonly Used Windows

What follows are pictures and descriptions of the most frequently used windows or screens in SPSS processes. These include:

1. The initial screen on entry into SPSS. This includes detail concerning the meaning of each icon and a brief description of the function of each command,

2. The **Open File** dialog, which identifies several different ways to access a previously created file, and
3. A main dialog box, which, although different for each procedure, has certain similarities that will be highlighted.

Following these three presentations the Output Window (the initial window that appears following completion of data analysis) and instructions (dealing with how to edit and manipulate output prior to printing results) are presented. This chapter concludes with a description of how to print or export output and a variety of options available with the **Options** option (page 24). We now begin with the screen that appears when you first enter SPSS.

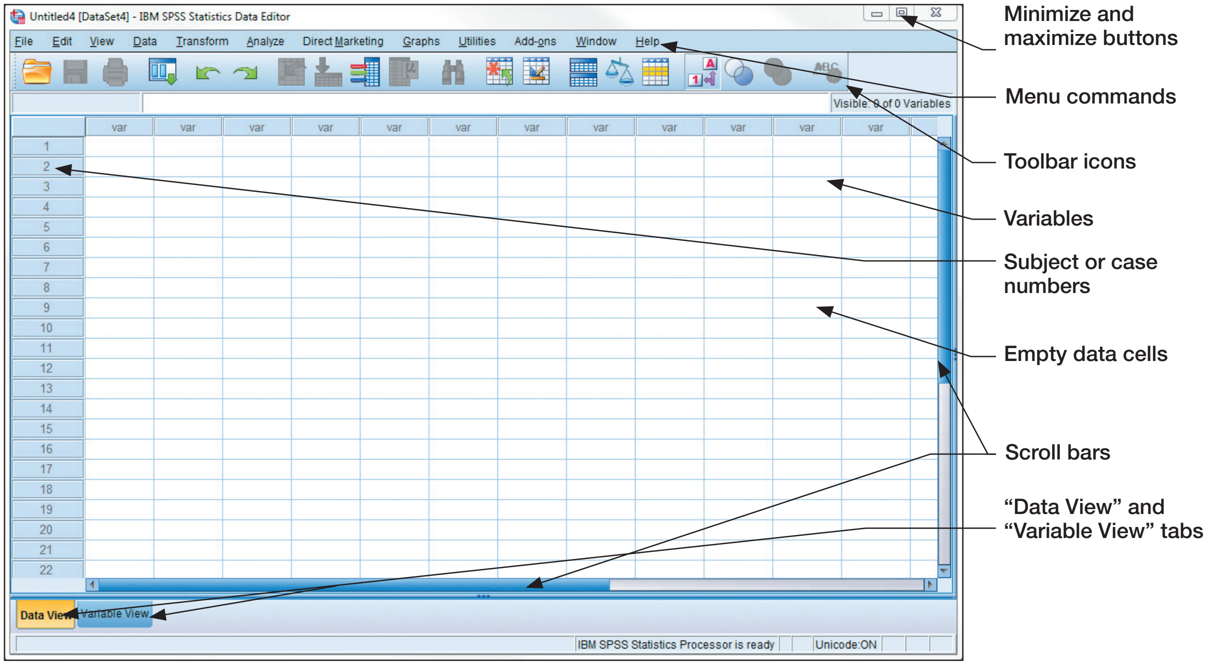
## 2.4.1 The Initial Screen, Icon Detail, Meaning of Commands

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | ✱ ⊞ → ✱ All Programs ▷ → ✱ IBM SPSS Statistics → ✱ IBM SPSS Statistics | Front1 |

Screen 2.3 (shown below) is a slightly modified version of Screen 1 on the inside front cover. When you start the SPSS program, Screen 2.3 is the first screen to appear:

**Screen 2.3** Initial data screen


Untitled4 [DataSet4] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 0 of 0 Variables
var
1
22
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON
Minimize and maximize buttons
Menu commands
Toolbar icons
Variables
Subject or case numbers
Empty data cells
Scroll bars
"Data View" and "Variable View" tabs

Screen 2.3 pictures a full-screen image of the data editor, with a detailed breakdown of the toolbar buttons. The menu bar (the commands) and the toolbar are located at the top of the screen and are described in the following pages. When you start SPSS, there are no data in the data editor. To fill the data editor window, you may type data into the empty cells (see Chapter 3) or access an already existing data file (described later in this chapter).

**TOOLBAR** The toolbar icons are located below the menu bar at the top of the screen. The icons were created specifically for ease of point-and-click mouse operations.

| Icon | Function | Icon | Function |
|---|---|---|---|
| 📂 | Click this to open a file | 🔭 | Find data |
| 💾 | Save current file | ✱ | Insert subject or case into the data file |
| 🖨 | Print file | ✎ | Insert new variable into the data file |
| ↓ | Recall a recently-used command | ▦ | Split file into subgroups |
| ↶ | Undo the last operation | ⚖ | Weight cases |
| ↷ | Redo something you just undid | ▦ | Select cases |
| + | (upper left corner) the "+" sign indicates that this is the active file | A 1 | Shifts between numbers and labels for variables with several levels |
| ↖ ↓ | Go to a particular variable or case number | ○ ● | Use subsets of variables/use all variables |
| ▤ | Access information about the current variable | ABC ✓ | Spell check |

It must be noted that even in SPSS applications the format of the icon bar may vary slightly. The toolbar shown above applies to the data editor window; a different toolbar is shown (pages 17–18) that applies to the output window. Also note that some of the icons are bright and clear and others are "grayed." Grayed icons are those that are not currently available. Note, for instance, that the Print File icon is grayed because there are no data to print. When data are entered into the data editor, then these icons become clear because they are now available. The best way to learn how the icons work is to click on them and see what happens.

**THE MENU BAR** The menu bar (just above the toolbar) displays the commands that perform most of the operations that SPSS provides. You will become well acquainted with these commands as you spend time in this book. Whenever you click on a particular command, a series of options appears below and you will select the one that fits your particular need. The commands are now listed and briefly described:

- **File**: Deals with different functions associated with files, including opening, reading, and saving, as well as exiting SPSS.
- **Edit**: A number of editing functions, including copying, pasting, finding, and replacing.
- **View**: Several options that affect the way the screen appears; the option most frequently used is **Value Labels.**
- **Data**: Operations related to defining, configuring, and entering data; also deals with sorting cases, merging or aggregating files, and selecting or weighting cases.
- **Transform**: Transformation of previously entered data, including recoding, computing new variables, reordering, and dealing with missing values.
- **Analyze**: All forms of data analysis begin with a click of the Analyze command.
- **Graphs**: Creation of graphs or charts can begin either with a click on the Graphs command or (often) as an option while other statistics are being performed.
- **Utilities**: Utilities deals largely with fairly sophisticated ways of making complex data operations easier. Most of these commands are for advanced users, and will not be described in this book.

- **Add-ons**: If you want to do advanced statistics that aren't already in SPSS, these menu options will direct you to other programs and services that SPSS can sell you.
- **Window**: Deals with the position, status, and format of open windows. This menu may be used instead of the taskbar to change between SPSS windows.
- **Help**: A truly useful aid with search capabilities, tutorials, and a statistics coach that can help you decide what type of SPSS procedure to use to analyze your data.

## 2.5 The Open Data File Dialog Window

The **Open Data** dialog window provides opportunity to access previously created data files. On the following page we show the **Open Data** window with key elements identified. To access this window, it is necessary to perform the following sequence of steps.

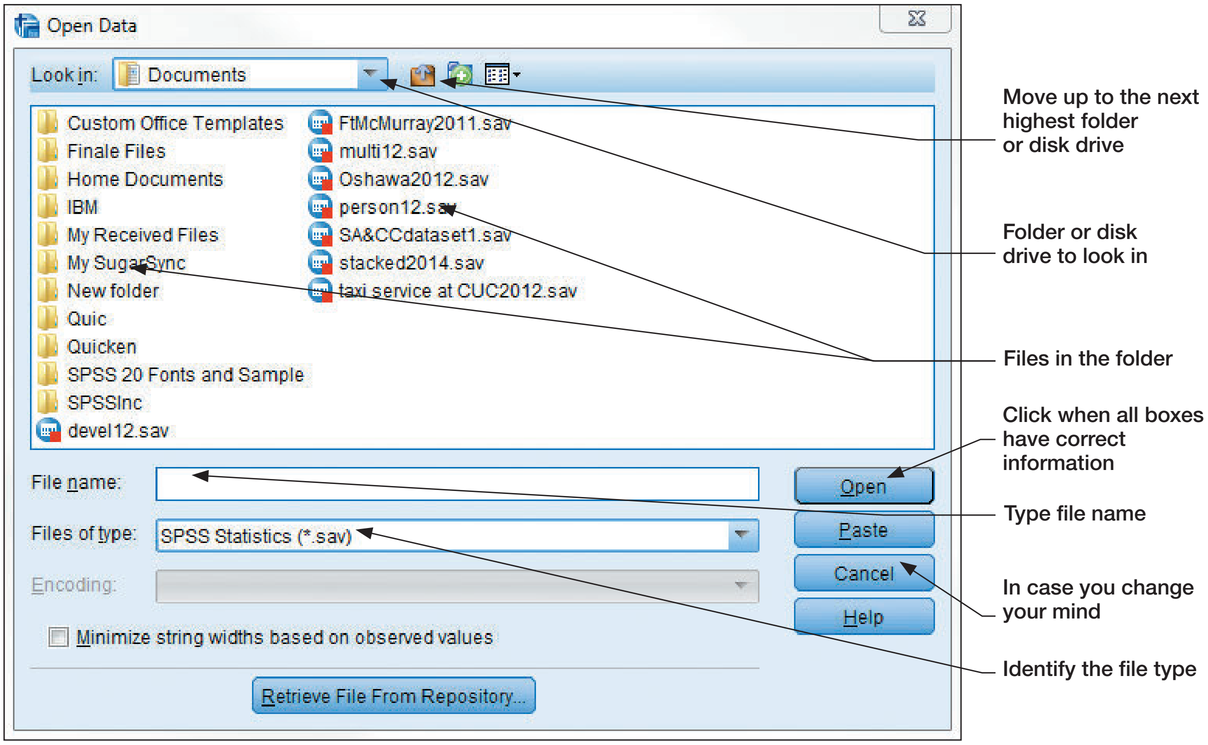
*From* Screen 2.3, *perform the following procedure to access the* **Open Data** *window.*

| In Screen | Do This | Step 2 |
|---|---|---|
| 2.3 | **File** → **Open** → **Data** [*or* 📂] | 2.4 |

To assist in description of the functions of this window, we will make use of the data file presented in the first chapter (and described in greater detail in Chapter 3).

The name of the data file is **grades.sav**. All data files created in the SPSS for Windows data editor are followed by a period (.) and the three-letter extension, **.sav**; depending on your Windows settings, the **.sav** may or may not appear in the Open Data dialog window.

**Screen 2.4** Open Data dialog window



Look in: Documents

Files of type: SPSS Statistics (*.sav)

□ Minimize string widths based on observed values

Move up to the next highest folder or disk drive

Folder or disk drive to look in

Files in the folder

Click when all boxes have correct information

Type file name

In case you change your mind

Identify the file type

If you wish to open a data file, there are several ways this may be accomplished. The two methods shown in the following page assume that you have already successfully navigated to the folder containing the **grades.sav** file.

*If the* **grades.sav** *file is visible in the list of files:*

| In Screen | Do This | Step 3 |
|---|---|---|
| 2.4 | ✳ ✳ **grades.sav** | Menu |

*A second option is to simply type the file name in the* **File name** *box, making sure the folder and disk drive (above) are correct, and then click on the* **Open** *button:*

| In Screen | Do This | Step 3a |
|---|---|---|
| 2.4 | type `grades.sav` → ✳ **Open** | Menu |

If you are reading a data file that was not created on the SPSS for Windows data editor, you need to identify the file type before SPSS can read it accurately. If, for instance, you are reading an Excel file (**grades.xls** in this case) your file will not appear unless you click the ▾ to the right of **Files of type** and select **Excel (*.xls, *.xlsx, *.xlsm)**. Then the file name should appear in the window above and you can double click to access that file.

The result of any of these operations is to produce a screen with the menu of commands across the top and the data from the **grades.sav** file in the data editor window. Notice (Steps 3 and 3a above) that the menu screen is the outcome (shown to the extreme right in each of the boxes) of each of these operations. The menu of commands across the top of the screen is what allows further analyses to take place.

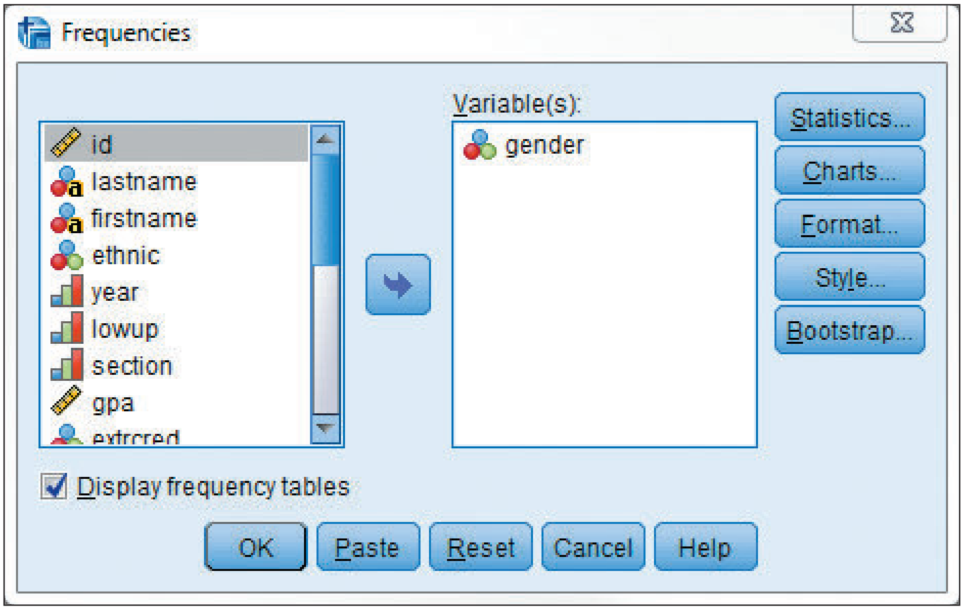
## 2.5.1 An Example of a Statistical-Procedure Dialog Window

There are as many statistical-procedure dialog windows as there are statistical procedures. Despite the fact that each is different from its fellow, there are fundamental similarities in each one. We illustrate by using the dialog window for Frequencies (Chapter 6). In all such dialog windows, the list of available variables resides in a box on the left. To the right will be another box or even two or three boxes. Although the title of this (or these) box(es) (to the right) will differ, they are designed to show which variables will be processed in the analysis. Between the boxes will be ➡ and/or ⬅ buttons to move variables from one box to another.

For all statistical-procedure dialog boxes five buttons are the same; the five buttons are always at the bottom of the dialog box:

- **OK**: When all variables and specifications are selected, click the **OK** button to compute the statistics.
- **Paste**: A click of the paste button will open a syntax window with the specifications identified in the dialog boxes pasted into the syntax box in a command-file format. Those acquainted with computer programming will enjoy this option. The reason this option is available is because the windows format (although simpler) has certain limitations that make it less flexible than the command-file format. If you wish to tailor your program to fit your specific needs then it is often useful to work with a command file. In SPSS 23 the command file is printed at the top of the Output section of each analysis.
- **Reset**: One characteristic of these main dialog boxes is called *persistence*. That means if one analysis is completed and the researcher wishes to conduct another, upon returning to the dialog box, all the same variables and specifications remain from the previous analysis. If that is what you wish, fine. If not, click the **Reset**

**Screen 2.5** Sample statistical procedure dialog window



☑ Display frequency tables

button and the variables will all return to the variable list and all specifications will return to the defaults.

- **Cancel**: For immediate exit from that procedure.
- **Help**: Similar to **Help** on the main command menu except that this help is context specific. If you are working with frequencies, then a click on the help button will yield information about frequencies.

Finally the three buttons to the right of the window (**Statistics, Charts, Format**) represent different procedural and formatting options. Most main-dialog boxes have one or more options similar to these. Their contents, of course, are dependent on the statistical procedure you happen to be performing.

### 2.5.2 Keyboard Processing, Check Boxes, and Radio Buttons

Some people do not like mice and SPSS statistics can be negotiated almost entirely from the keyboard. Let's clarify: Navigating within a data file (both moving from one data point to another and selecting blocks of data) can occur by use of the keyboard alone. When you open any data file, the upper left cell is selected by default; selection is designated by a yellow highlight of that cell. Some examples of movement within the data file follow:

**Tab**: yellow highlight moves to the right

**Enter**: highlight moves down

**Cursor keys**: highlight moves up, down, right, or left

**Home**: moves highlight to the first variable in the row

**End**: moves highlight to the last variable in the row

**Ctrl-Home** or **Ctrl-End**: moves highlight to the upper left or lower right corner of the data set

**Shift + cursor keys**: selects blocks of data starting with any highlighted cell

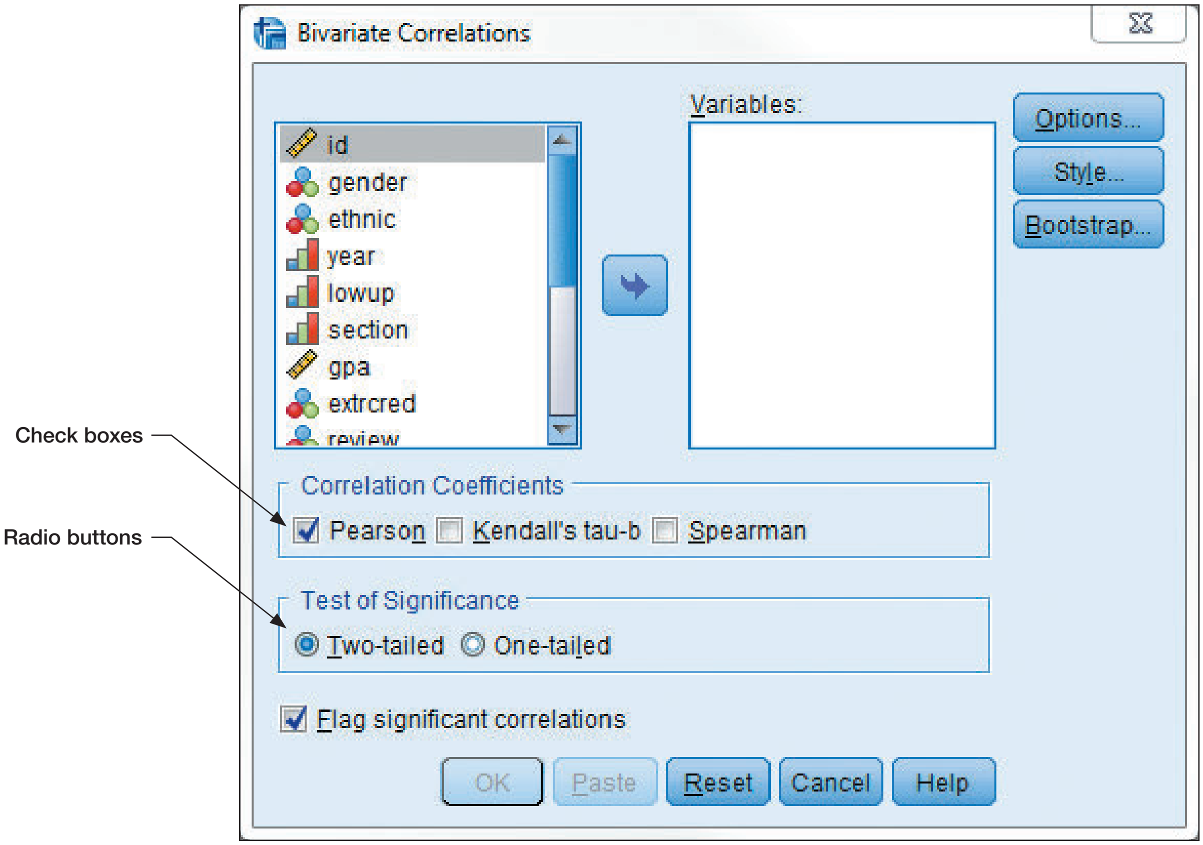
Another keyboard option deals with commands in which one letter of that option is underlined (e.g., **Edit, Options**). To access a particular procedure, press the **ALT** key, then press the underlined letter on the keyboard to select that option. For instance, to select **Edit**, press the **ALT** key and then the letter **E**. The menu under **Edit** will immediately appear. Even for mouse users, the keystrokes are often faster than mouse clicks once you have learned them. There are many more keyboard operations but this should give you a start.

Selecting icons from the tool bar and shifting from **Data View** to **Variable View** (lower left corner of the screen; see page 11) require the mouse or an equivalent selection device.

Finally, there are two different types of selections of options that may take place within dialog boxes. Note the window (in the following page) from Chapter 10. Notice the **Test of Significance** box near the bottom; it contains **radio buttons**. They are called radio buttons because if one is selected the others are automatically deselected (as in a car radio—if you select one channel you deselect the others).

**Check boxes** are shown under the **Correlation Coefficients** heading and at the very bottom of the dialog box. With check boxes you simply select those you desire. You may select all three, two, one, or none of them. A selection is indicated when a ☑ appears in the box to the left of the option. For radio buttons, click on a different option to deselect the present option. In check boxes, click on the box a second time to deselect it.

**Screen 2.6** Sample dialog window with radio buttons and check boxes



Correlation Coefficients

☑ Pearson □ Kendall's tau-b □ Spearman

Test of Significance

● Two-tailed ○ One-tailed

☑ Flag significant correlations

Check boxes

Radio buttons

## 2.6 The Output Window

"Output" is the term used to identify the results of the previous analyses. It is the objective of all data analysis. SPSS has a long history of efforts to create a format of output that is clear yet comprehensive. The current version uses a tables-with-borders format. When utilizing options described below, this is relatively clear, but output can still be awkward and occupy many pages. It is hoped that the information that follows maximizes your ability to identify, select, edit, and print out the most relevant output. This section is important since Output is almost always edited before it is printed.

The initial output screen is shown on the inside back cover (and below). Each chapter that involves output will give instructions concerning how to deal with the results displayed on this screen. Our focus here is to explain how to *edit* output so that, when printed, it will be reproduced in a format most useful to you. Of course, you do not have to edit or reorganize output before you print, but there are often advantages to doing so:

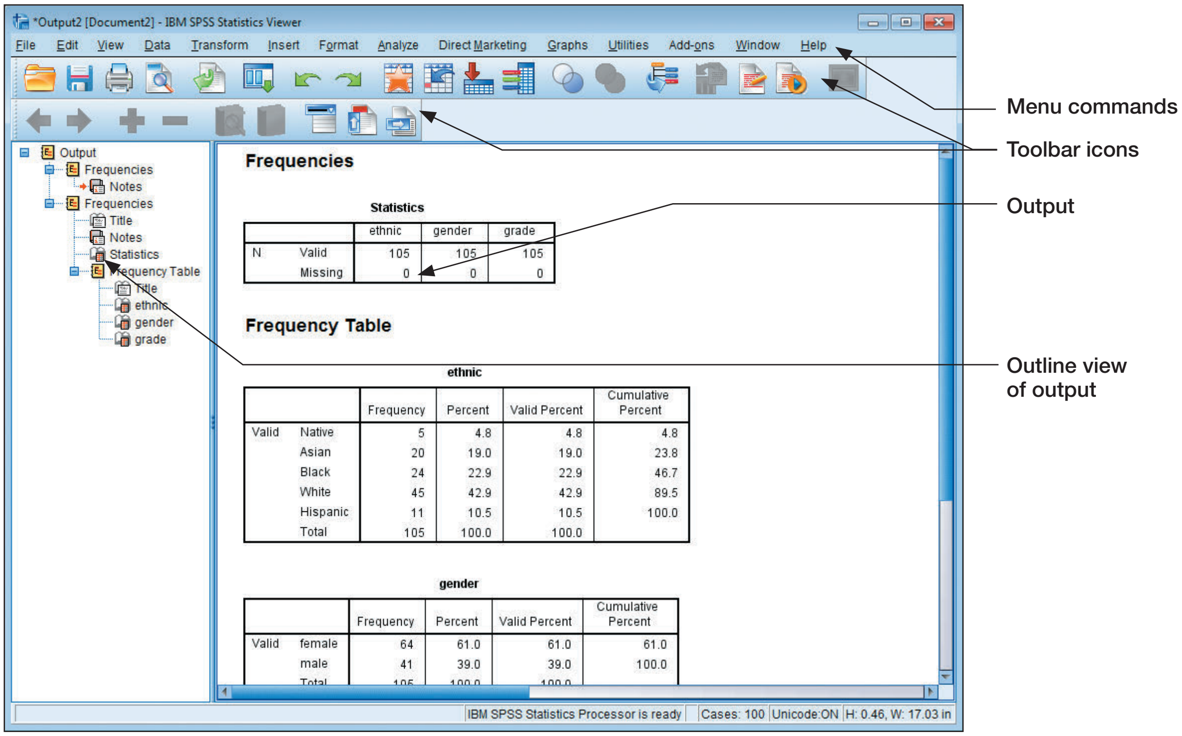
- Extensive outputs will often use/waste many pages of paper.
- Most outputs will include some information that is unnecessary.
- At times a large table will be clearer if it is reorganized.
- You may wish to type in comments or titles for ease or clarity of interpretation.
- SPSS systematically prints the syntax file at the top of every output and the location of your file. You may simply delete this information if you wish.

This chapter will explain the SPSS Output window (and toolbar), how to delete output that you no longer need, how to add comments to your file, how to re-arrange the order of the output, and how to save the output.

Screen 2.7 (below) summarizes the key elements of the output window, as well as providing a detailed description of toolbar items on the output window. In this sample, several procedures have been performed on the **grades.sav** file; don't worry about the content of the analysis, just pay attention to what SPSS will allow you to do with the output.

You will notice that several of the toolbar icons are identical to those in the SPSS Data Editor window; these buttons do the same thing that they do in the Data Editor, but with the output instead of the data. For example, clicking the print icon prints the output instead of the data.

**Screen 2.7** The SPSS Output navigator window



Frequencies

Statistics

| | | ethnic | gender | grade |
|---|---|---|---|---|
| N | Valid | 105 | 105 | 105 |
| | Missing | 0 | 0 | 0 |

Frequency Table

ethnic

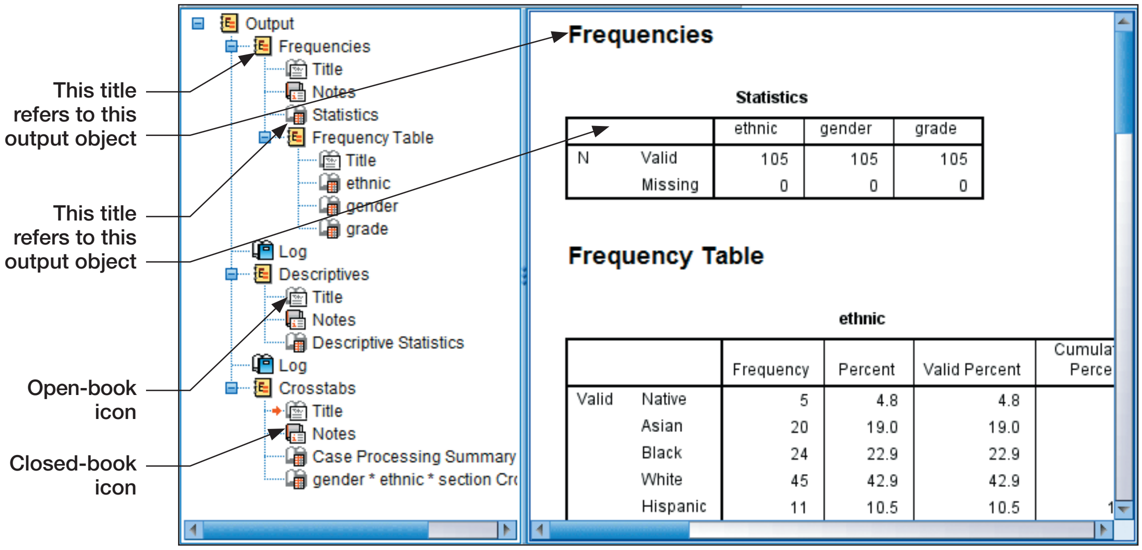
| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Native | 5 | 4.8 | 4.8 | 4.8 |
| | Asian | 20 | 19.0 | 19.0 | 23.8 |
| | Black | 24 | 22.9 | 22.9 | 46.7 |
| | White | 45 | 42.9 | 42.9 | 89.5 |
| | Hispanic | 11 | 10.5 | 10.5 | 100.0 |
| | Total | 105 | 100.0 | 100.0 | |

gender

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | female | 64 | 61.0 | 61.0 | 61.0 |
| | male | 41 | 39.0 | 39.0 | 100.0 |
| | Total | 105 | 100.0 | 100.0 | |

| Icon | Function | Icon | Function |
|---|---|---|---|
| | Click to open a file | | Show all variables |
| | Click to save a file | | Select most recent output |
| | Print output | ⬅ | Promote: Make the currently selected output into a heading |
| | Print preview (see what the printout will look like) | ➡ | Demote: Make the currently selected output into a subheading |
| | Export output to text or web (HTML) file | ✚ | Display everything under the currently selected heading |
| | Recall a recently used command | ▬ | Hide everything under the currently selected heading |
| | Undo or redo the last operation | | Display the currently selected object (if it is hidden) |
| | Go to SPSS Data Editor (frequently used!) | | Hide the currently selected object (visible in outline view only) |
| | Go to a particular variable or case number | | Insert heading (above titles, tables, and charts) |
| | Get information about variables | | Insert title above output |
| | User-defined subset of the data | | Insert text at strategic locations within the output |

One of the most important things to learn about the SPSS Output window is the use of the outline view on the left of the screen. On the right side of the window is the output from the SPSS procedures that were run, and on the left is the outline (like a table of contents without page numbers) of that output. The SPSS output is actually composed of a series of *output objects*; these objects may be titles (e.g., “Frequencies”), tables of numbers, or charts, among other things. Each of these objects is listed in the outline view:



Frequencies

**Statistics**

| | | ethnic | gender | grade |
|---|---|---|---|---|
| N | Valid | 105 | 105 | 105 |
| | Missing | 0 | 0 | 0 |

Frequency Table

**ethnic**

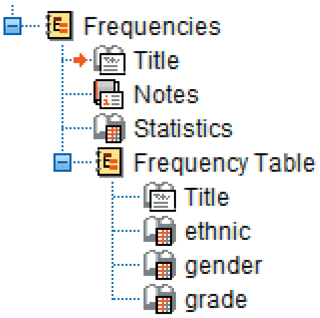
| | | Frequency | Percent | Valid Percent | Cumula Perce |
|---|---|---|---|---|---|
| Valid | Native | 5 | 4.8 | 4.8 | |
| | Asian | 20 | 19.0 | 19.0 | |
| | Black | 24 | 22.9 | 22.9 | |
| | White | 45 | 42.9 | 42.9 | |
| | Hispanic | 11 | 10.5 | 10.5 | 1 |

You will notice that there is no “Notes” section in the output window to correspond with the “Notes” title in the outline view. That’s because the notes are (by default) hidden. If you want to see the notes, first click on the word “Notes” and then click on the open-book icon (📖) at the top of the page. The closed-book icon will then become an open-book icon and the notes will materialize in the window to the right. Click the close-book icon (📙) and the notes will disappear and the book will close.

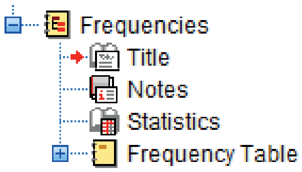
The outline view makes navigating the output easier. If you want to move to the Crosstabs output, for example, you merely need to click on the word "Crosstabs" in the outline view, and the crosstabs will appear in the output window. If you want to delete the Descriptives section (perhaps because you selected an incorrect variable), simply click on the word "Descriptives" (to select that menu item) and then click **Delete**. If you want to move some output from one section to another (to re-arrange the order), you can select an output object (or a group of output objects), and select **Edit** and then click **Cut**. After that, select another output object below which you want to place the output object(s) you have cut. Then select **Edit** and click **Paste After**.

If you have been working with the same data file for a while, you may produce a lot of output. So much output may be produced, in fact, that it becomes difficult to navigate through the output even with the outline view. To help with this problem, you can "collapse" a group of output objects underneath a heading. To do this, click on the minus sign to the left of the heading. For example, if you want to collapse the Frequencies output (to get it out of the way and come back to it later) simply click on the minus sign to the left of the title **Frequency Table**. If you want to expand the frequencies heading later, all you have to do is click on the plus sign to the left of the title **Frequency Table**. This operation is illustrated below.

Clicking on the lower minus sign…



…will collapse the Frequency Table objects to this:

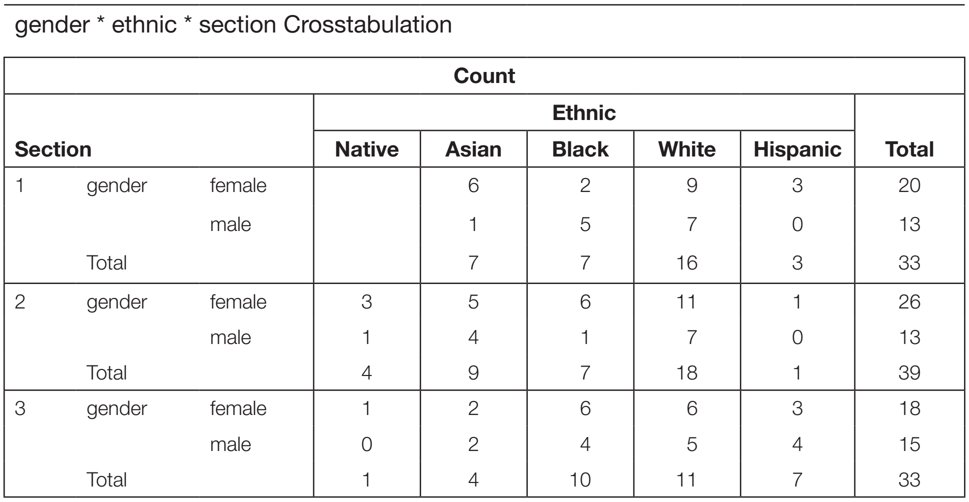


One particularly useful command when you are working with output is the insert text command (📝) When you click this button, an SPSS Text object is inserted. In this box, you can type comments to remind yourself what is interesting about the SPSS output (for example, "The next chart would look good in the results section" or "Oh, no, my hypothesis was not supported!"). Once you have typed your comments, click on another SPSS object to deselect the SPSS Text object.

## 2.7 Modifying or Rearranging Tables

Now that you have learned the basics of getting around within SPSS output, there is one more major concept to learn. When SPSS produces output, whether it be charts or tables, you can make changes in the format of that output. Although this book will not describe all of the procedures that you can perform to modify the appearance of charts and tables, some basic operations will be described that will allow you to make your charts and tables more attractive or easier to read. The basics of editing charts and graphs are described in Chapter 5; the basic processes of modifying tables are described here.

First, we will start with some background and theory of how tables work. It is important to realize that, although SPSS automatically arranges tables for you—it decides what to put in the columns and rows—the arrangement of a table is somewhat arbitrary. For example, in the crosstabulation of **gender x ethnicity x section** (described in Chapter 8), by default SPSS places the ethnicity in the columns, and section and gender in rows:



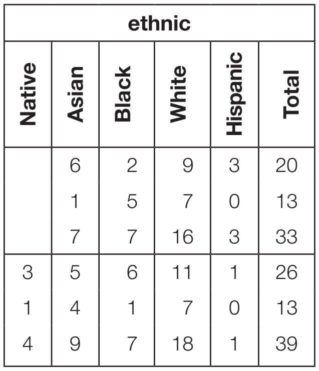
gender * ethnic * section Crosstabulation

| Count | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Ethnic | | | | | |
| Section | | | Native | Asian | Black | White | Hispanic | Total |
| 1 | gender | female | | 6 | 2 | 9 | 3 | 20 |
| | | male | | 1 | 5 | 7 | 0 | 13 |
| | Total | | | 7 | 7 | 16 | 3 | 33 |
| 2 | gender | female | 3 | 5 | 6 | 11 | 1 | 26 |
| | | male | 1 | 4 | 1 | 7 | 0 | 13 |
| | Total | | 4 | 9 | 7 | 18 | 1 | 39 |
| 3 | gender | female | 1 | 2 | 6 | 6 | 3 | 18 |
| | | male | 0 | 2 | 4 | 5 | 4 | 15 |
| | Total | | 1 | 4 | 10 | 11 | 7 | 33 |

Although there is nothing wrong with this arrangement, you might want to rearrange the table, depending on what exactly you were trying to demonstrate. It is important to remember that you are *not* changing the content of the statistics present within the table; you are only rearranging the *order* or the format in which the data and statistics are displayed.

You will notice in the table (above), there are both *rows* and *columns*. The columns contain the various categories of **ethnic**. The rows contain two different levels of headings: In the first level is the course **section**, and in the second level is the **gender** of the students. You will notice that each level of the second heading (each gender) is displayed for each level of the first heading (each section). This table has one *layer*: The word "**Count**" (the number of subjects in each category) is visible in the upper left hand corner of the table. In most tables, there is only one layer, and we will focus our discussion here on arranging rows and columns.

In order to edit and change the format of an output table, you must first double click on the output object in the output view (the right portion of Screen 2.7). When you do that, the toolbars will disappear, and in the menu bar will appear a new item: **Pivot**. In addition to this, some of the options available on the various menus change. We will describe some of the options available on the **Pivot** and **Format** menus, including **Transpose Rows and Columns**, **Pivoting Trays**, **Rotate Inner Column Labels**, and **Rotate Outer Row Labels**.



| ethnic | | | | | |
|---|---|---|---|---|---|
| Native | Asian | Black | White | Hispanic | Total |
| | 6 | 2 | 9 | 3 | 20 |
| | 1 | 5 | 7 | 0 | 13 |
| | 7 | 7 | 16 | 3 | 33 |
| 3 | 5 | 6 | 11 | 1 | 26 |
| 1 | 4 | 1 | 7 | 0 | 13 |
| 4 | 9 | 7 | 18 | 1 | 39 |

You will notice in the sample table shown above that the columns are much wider than they need to be to display the numbers. This is so the columns will be wide enough to display the labels. Clicking on **Format** followed by **Rotate Inner Column Labels** will fix this problem by making the column labels much taller and narrower, as shown to the left. If you decide you prefer the previous format, simply click on **Format** followed by **Rotate Inner Column Labels** once again, and the old format will be restored. It is also possible to change the orientation of the row titles. This makes the section numbers tall and narrow. But this also makes the heading "Section" tall and narrow, and renders the table itself narrower.

When SPSS produced the sample **gender x ethnic x section** table, it placed ethnicity in the columns, and course section and gender in the rows. A researcher may want these variables displayed differently, perhaps with ethnicity in the rows, and both gender and class section in the columns. In order to make this change, you may use either **Transpose Rows and Columns** or **Pivoting Trays**. Clicking **Pivot** followed by **Transpose Rows and Columns** is the simplest way to rearrange rows and columns. If we were to choose this command on our sample output table, then the rows would become columns and the columns rows, as shown below:

gender * ethnic * section Crosstabulation

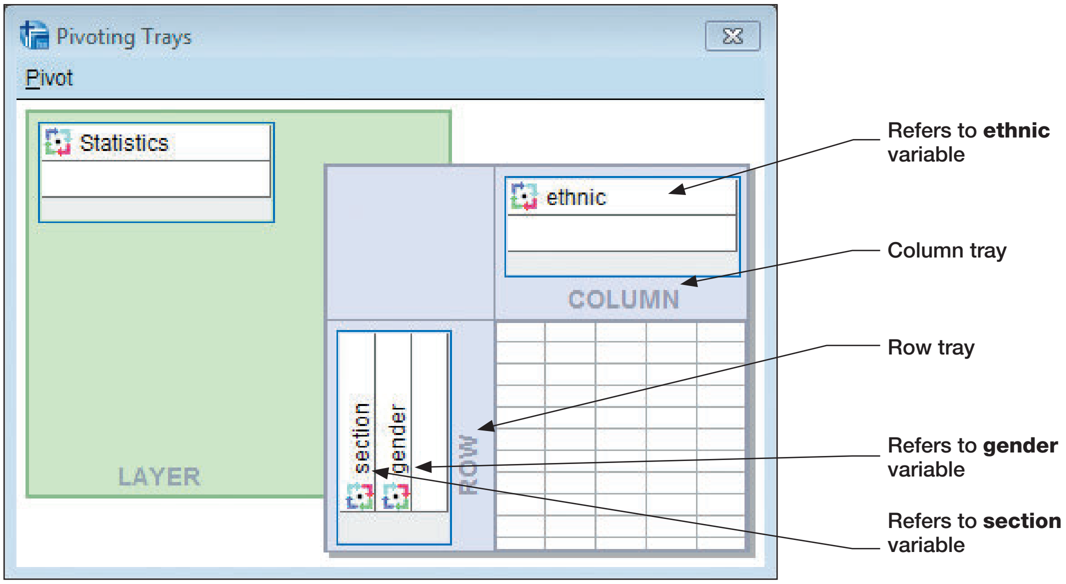
**Count**

| | | **Section** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **1** | | | **2** | | | **3** | | |
| | | **gender** | | **Total** | **gender** | | **Total** | **gender** | | **Total** |
| | | **female** | **male** | | **female** | **male** | | **female** | **male** | |
| Ethnic | Native | | | | 3 | 1 | 4 | 1 | 0 | 1 |
| | Asian | 6 | 1 | 7 | 5 | 4 | 9 | 2 | 2 | 4 |
| | Black | 2 | 5 | 7 | 6 | 1 | 7 | 6 | 4 | 10 |
| | White | 9 | 7 | 16 | 11 | 7 | 18 | 6 | 5 | 11 |
| | Hispanic | 3 | 0 | 3 | 1 | 0 | 1 | 3 | 4 | 7 |
| Total | | 20 | 13 | 33 | 26 | 13 | 39 | 18 | 15 | 33 |

A somewhat more complex and sophisticated way of rearranging columns and rows is to use **Pivoting trays**. Selecting this option (a subcommand on the **Pivot** menu) will produce a new window that allows you to rearrange rows and columns (Screen 2.8). This window will also allow you to rearrange layers, but because the arrangement of layers usually doesn't need to be changed, we won't worry about layers in this discussion.

The pivoting tray window consists of different trays: a tray for **ROW** variables, and a tray for **COLUMN** variables. "Statistics" is a separate **LAYER** and refers to the

**Screen 2.8** Pivoting Trays Window



Refers to **ethnic** variable

Column tray

Row tray

Refers to **gender** variable

Refers to **section** variable

data that fill the cells. In our example **ethnic**ity is in the column tray, with **section** and **gender** both in the row tray.

You may click and drag any of the variables from an original location to a different location. For example, if you want to rearrange the order of the **section** and **gender** variables, you could click and drag **section** to the blank box to the right of **gender**. This would rearrange the rows so that the gender was the main heading level, with three levels of each section for each of the two levels of gender.

You can also rearrange variables between rows and columns, by dragging a variable icon from one tray to another. For example, if you wanted to make **section** a major column heading with levels of ethnicity displayed for each section, you would drag the **section** variable icon to the same row as the **ethnic** variable and **section** will become the major heading.

The best way to understand pivoting tables is simply to experiment with them. Because of the persistence function, it's not at all difficult to run the analysis again if you get too confused.

Once you have done all of this work making your output look the way you want it to look (not to mention doing the statistical analyses to begin with), you may want to save your work. This, by the way, is quite different from saving the data file (always followed by a "**.sav**" suffix). Output is generally saved with a "**.spv**" suffix. The sequence step required to do that follows:

| In Screen | Do This | Step 4 |
|---|---|---|
| 2.7 | ✳ File → ✳ Save → type `yourname.spv` → ✳ Save | Menu |

## 2.8 Printing or Exporting Output

Toward the end of the Step by Step section of each chapter, skeletal information is provided concerning printing of output and/or graphs. Every time this is done the reader is referred to this section in this chapter for greater detail. If you desire more complete information than provided here, please refer to the *SPSS* manuals.

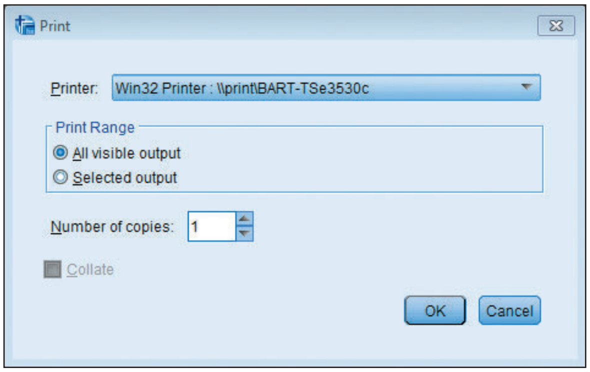
The first time you print, it is a good idea to select the **File** and then click the **Page setup** option. This very intuitive dialog box lets you specify whether the orientation of printed material will be portrait (longer dimension is vertical) or landscape (longer dimension is horizontal), the margins, and the size and source of your paper.

In order to print your output, the material that you wish to print must be *designated*—that is, you must select the portions of your output that you want printed. If you know that you want all of your output printed, then you don't have to designate the material prior to printing. We don't, however, recommend making this a habit: Save the trees. (You don't have to hug them if you don't want, but at least think about what you are going to print before you start.)

To print a section of the output you must highlight the portion of the file you wish to reproduce *before* beginning the print sequence. This may be accomplished by

1. Clicking on the folder or output object you want to print on the left side of the SPSS output window;
2. Clicking on a folder or output object on the left side of the SPSS output window, then holding down the **Shift key** while selecting another folder or output object (this will select both objects you clicked and all the objects in between); or
3. Clicking on a folder or object on the left side of the SPSS output window, then holding down the **Control** key while clicking on one or more additional folders or objects (this will select only the folders or objects that you click on).

## Screen 2.9 Print Dialog Box



Printer: Win32 Printer : \\print\BART-TSe3530c

Print Range

● All visible output

○ Selected output

Number of copies: 1

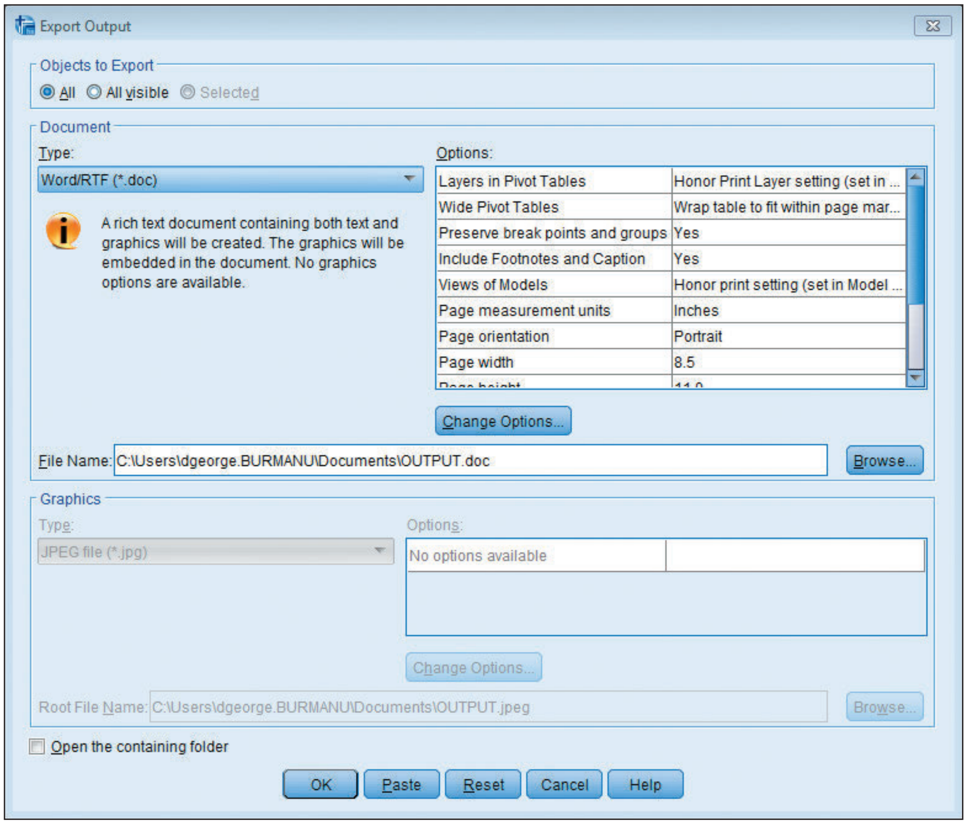
□ Collate

The process always starts with a click on the **File** command followed by a click on the **Print** option. This causes a small dialog box to open (Screen 2.9), allowing you to choose which printer you want to use, how many copies to print, and whether to print just the designated (selected) part of the output or the entire output. No formatting options are allowed in this dialog box. You must use the **Page setup** procedure (described in the previous page) prior to printing.

Once all of the options in the **Print** dialog have been specified as desired, click the **OK** button and watch your output pour from the printer.

Of course, you may not find it necessary to print the output on paper. You can always open the output on SPSS and view it on the screen . . . but what if you want to view the output on a different computer, one that doesn't have SPSS? In that case, instead of printing your output, you can export it in PDF format which can be read on any computer with Adobe Reader or a variety of other free programs. To do that, select **File**, then click the **Export** option. Screen 2.10 will appear.

## Screen 2.10 Export Output



Objects to Export

● All ○ All visible ○ Selected

Document

Type: Word/RTF (*.doc)

A rich text document containing both text and graphics will be created. The graphics will be embedded in the document. No graphics options are available.

Options:

| | |
|---|---|
| Layers in Pivot Tables | Honor Print Layer setting (set in ... |
| Wide Pivot Tables | Wrap table to fit within page mar... |
| Preserve break points and groups | Yes |
| Include Footnotes and Caption | Yes |
| Views of Models | Honor print setting (set in Model ... |
| Page measurement units | Inches |
| Page orientation | Portrait |
| Page width | 8.5 |
| Page height | 11.0 |

File Name: C:\Users\dgeorge.BURMANU\Documents\OUTPUT.doc

Graphics

Type: JPEG file (*.jpg)

Options: No options available

Root File Name: C:\Users\dgeorge.BURMANU\Documents\OUTPUT.jpeg

□ Open the containing folder

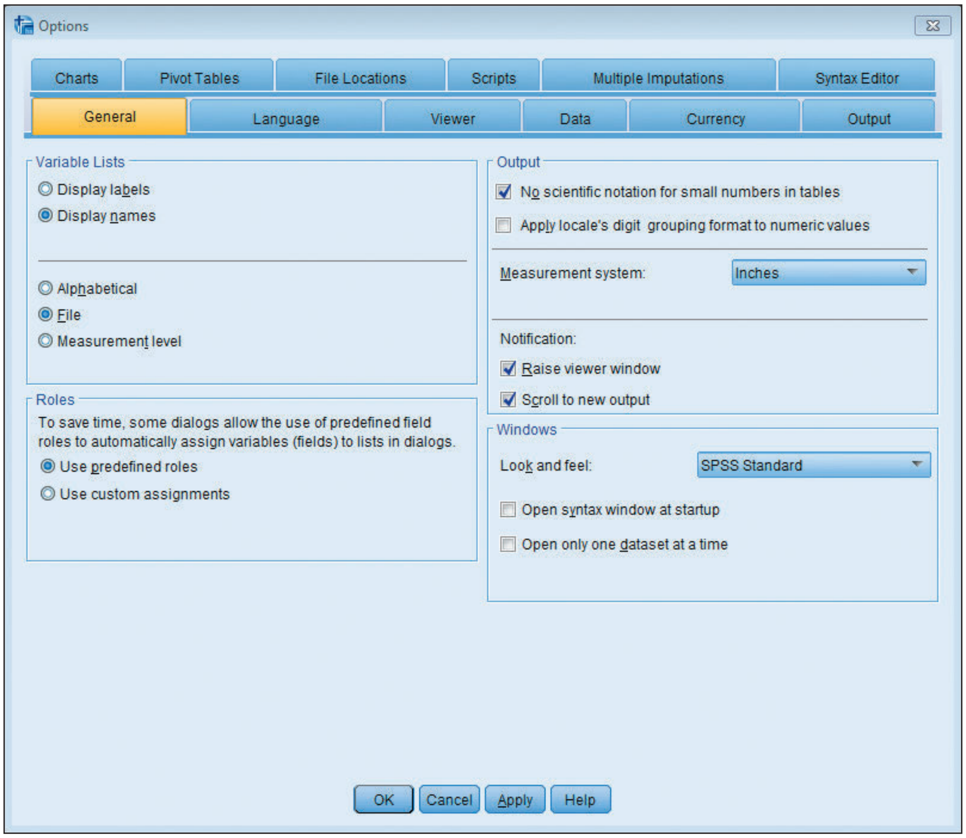
You can export your output to many formats—Microsoft Word, HTML, Excel, PowerPoint—but it is usually best to select **Type → Portable Document Format (*.pdf)**. Other formats are often difficult to read. Then you merely have to tell SPSS where to save your file and what to name it (by clicking the **Browse** button), and you will have an output that you can view anywhere.

## 2.9 The "Options . . . " Option: Changing the Formats

Under the **Edit → Options. . .** sequence lies a wide variety of alternatives to SPSS defaults. An awareness of the available options allows you to format many features associated with the appearance of the SPSS screen, dialog windows, and output. We reproduce the dialog window here and then identify the selections that we have found to be the most useful.

Notice that there are 11 different tabs across the top (see Screen 2.11, below); each deals with a different set of functions. Here are the options we use most frequently:

**Screen 2.11** Options box



Variable Lists
○ Display labels
● Display names

○ Alphabetical
● File
○ Measurement level

Roles
To save time, some dialogs allow the use of predefined field roles to automatically assign variables (fields) to lists in dialogs.
● Use predefined roles
○ Use custom assignments

Output
☑ No scientific notation for small numbers in tables
□ Apply locale's digit grouping format to numeric values

Measurement system: Inches

Notification:
☑ Raise viewer window
☑ Scroll to new output

Windows
Look and feel: SPSS Standard
□ Open syntax window at startup
□ Open only one dataset at a time

**General** tab (currently visible): Under **Variable Lists**, SPSS 21.0 has introduced some useful defaults that you will usually want to keep: **Display names** (rather than labels) and **No scientific notation for small numbers in tables**. If you know the data file well, stay with the default **File** so that variables will be listed in the original order. If you are dealing with a large data file (or one that you don't know very well), then **Alphabetical** may be a better solution.

The **Viewer** tab allows you to specify the format of your output by selecting font, font size, or special effects such as bold, underline, italics, and color.

**Output Labels** tab: In the four boxes to the left be sure to change to a sequence (from top to bottom) of **Names, Labels, Names, Labels** rather than the default of all **Labels**. Then click **Apply** to secure the new setting. The default option (all four are **Labels**) often creates a mess by filling your output with cumbersome descriptions rather than precise names.

**Pivot Tables** tab: There are 18 different options that deal with the format of the tables used in your outputs. You may selectively choose those output formats that best suit your needs. For instance, in a large table, one of the "narrow" formats will allow you to get more information on a single page.

There are others, many others, but this gives you a start.

# Chapter 2B

# IBM SPSS Statistics Processes for Mac

WE DEAL in this chapter with the grim reality that SPSS was for many years designed primarily for the PC. We as Mac users “know” that we have a superior operating system, have more intuitive operations and sequences, and in general believe we can thrash the heck out of any PC. To add insult to injury the first eleven editions of this book included the PC bias right in its title: *SPSS for Windows Step by Step*.

SPSS created a version of their program for Mac eons ago, but for many years the Mac version worked differently (and not good differently). Only with SPSS 16.0 did they create fully parallel programs, that is, once you get into the program, whether you are a PC or a Mac user, the screens and the sequences are essentially identical.

What we as authors have chosen to do is to write two complete versions of Chapter 2 introducing SPSS sequences and processes: one Chapter 2 for PC users and the other for Mac users. In *this* Chapter 2 everything you see employs Mac screen shots and terminology. This chapter will also alert you to any significant differences between the two operating systems and provide you with the grounding to make maximum use of the rest of the book.

Throughout the remainder of the book we retain the PC screens, but, other than minor variations, the screens and sequences for PC and Mac are identical. Many of our readers have used prior editions of our book even though they have the Mac OS version of SPSS. Even though they have not had difficulty, this chapter should now make it even easier.

## 2.1 Selecting

For most purposes you can use your mouse, touchpad, or touchscreen the same as you would most other programs, such as Microsoft Word. The point and click, double click, and dragging operate in ways similar to major word processing programs although sometimes SPSS has unique responses. Important differences will be noted in the chapters where they apply. The right click (on a mouse) or two-finger click (on a touchpad) is usually used to provide options of what you can do to whatever you clicked. In this book a left-mouse click (on a mouse), a one-fingered click (on a touchpad), or a one-fingered touch (on a touchscreen) is designated by a small picture of clicking cursor (↖). A right- or two-fingered mouse click will be identified in words.

## 2.2 The Desktop, Dock, and Application Folder

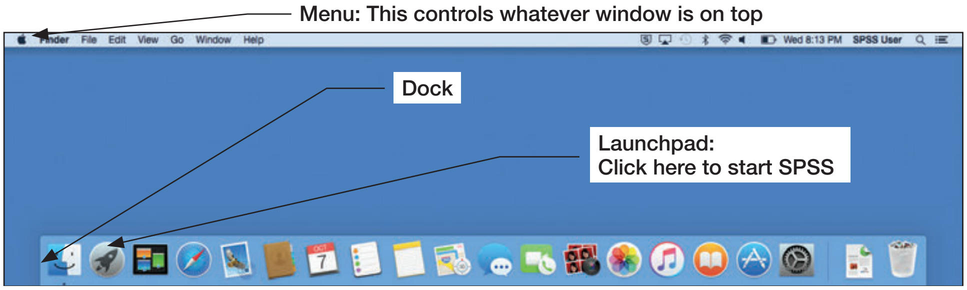
Once you have arrived at the desktop, your screen should look something like that shown in the following page. It may not look exactly like this, but it will be similar. The menu of commands across the top control the currently selected program, day

Reprinted with permission of Apple Inc. and Reprinted with permission of Microsoft, Inc.

and time appear to the right, and the Dock, once activated, reveals icons that access the different programs on your machine.

There are two main types of icons on the Dock: Application icons represent a particular program, while folder icons actually contain other icons (usually several programs that are related).

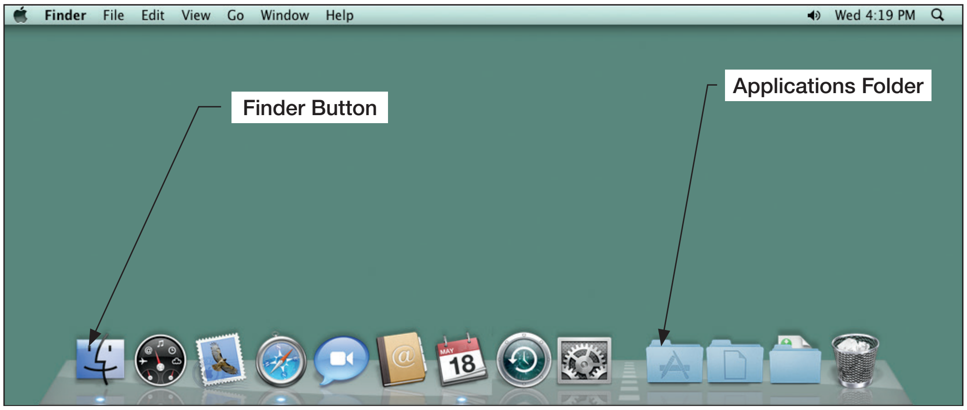
## Screen 2.1 The Mac Desktop and Dock


Menu: This controls whatever window is on top
Finder File Edit View Go Window Help
Wed 8:13 PM SPSS User
Dock
Launchpad:
Click here to start SPSS

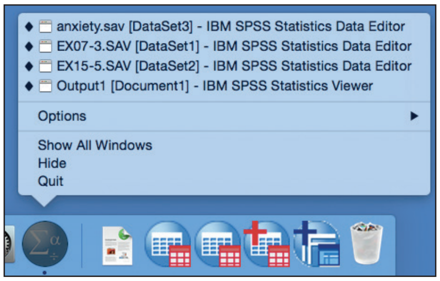
On most newer Macs, you will first click the Launchpad icon (🚀), then click on the IBM folder (▦), and then click on SPSS Statistics icon (Σ).

If you don't have a Launchpad icon, then this procedure won't work and you'll have to instead click on the Applications button on the Dock (Screen 2.2), or (if there isn't one) the Finder button, then the Applications button. Look for a folder labeled IBM. You may have to click through several other folders before you get to the SPSS Statistics icon.

## Screen 2.2 The Mac Dock on Older Machines



In addition to starting the SPSS program, the other important skill when using the Dock is changing between programs. This is especially important because SPSS is actually a collection of several programs. When you first start the SPSS system, the Data Editor program is started, but as soon as you perform some statistics on your data, another program (in another window) is started: the Output program. In addition, it is possible to have several data files open at the same time.


anxiety.sav [DataSet3] - IBM SPSS Statistics Data Editor
EX07-3.SAV [DataSet1] - IBM SPSS Statistics Data Editor
EX15-5.SAV [DataSet2] - IBM SPSS Statistics Data Editor
Output1 [Document1] - IBM SPSS Statistics Viewer
Options
Show All Windows
Hide
Quit

The SPSS icon that appears at the bottom of the screen can include within it several different programs (data files and output). To access any of the programs that aren't easily seen on the screen, move you cursor to the SPSS icon and right-click or two-finger click. In the picture (to the right) there are three data files open (the first three files listed three icons to the left) and one output file open (the fourth file listed). To select a particular file, just click on its name.

## 2.3 Common Buttons

A number of buttons occur frequently within screens and around the borders of windows and screens. The standard push buttons have the same function in any context. The most frequently used ones are identified and explained below:

| Icon | Description |
|---|---|
| ⊕ | This is the green zoom button. It is similar to the maximize button on a PC but does more. The zoom button, when clicked, will not only expand a program to fill the screen but will return it to its initial size if you click the button again. |
| ⊖ | This is the yellow minimize button. When you click it the program disappears from the screen and moves down to the right side of the Dock as a small icon. A click on the icon will restore the program again. |
| ⊗ | This is the red close button. Its function is to quit the open window. However, it does not necessarily quit the program. As a general rule, applications that can have multiple windows open at any given time remain open when the window's close button is clicked. Single window applications on the other hand will quit. |
| ➡ | Within the SPSS context, this button will usually paste (move text from one location to another) a designated variable from a list in one location to an active box in another location. Such a move indicates that this variable is to be used in an analysis. |
| ⬅ | This button is the reverse. It will move a variable from the active box back to the original list. |
| ▲ ▼<br>▶ ◀ | These are up, down, right, and left scroll arrows that are positioned on each end of a scroll bar. Most data or output files are longer than the open window and in many instances also wider. The scroll arrows and scroll bars allow you to view either above, below, to the right, or to the left of what is visible on the screen. |
| ▼ | A click on a down arrow will reveal a menu of options (called a drop-down menu) directly below. Screen 2.4 shows two such buttons to the right of Look in, and Files of type. |

## 2.4 The Data and Other Commonly used Windows

What follows are pictures and descriptions of the most frequently used windows or screens in SPSS processes. These include:

1. The initial screen on entry into SPSS. This includes detail concerning the meaning of each icon and a brief description of the function of each command,
2. The **Open File** dialog, which identifies several different ways to access a previously created file, and
3. A main dialog box, which, although different for each procedure, has certain similarities that will be highlighted.

Following these three presentations the output window (the initial window that appears following completion of data analysis) and instructions (dealing with how to edit and manipulate output prior to printing results) are presented. This chapter concludes with a description of how to print or export output and a variety of options available with the **Options** option.

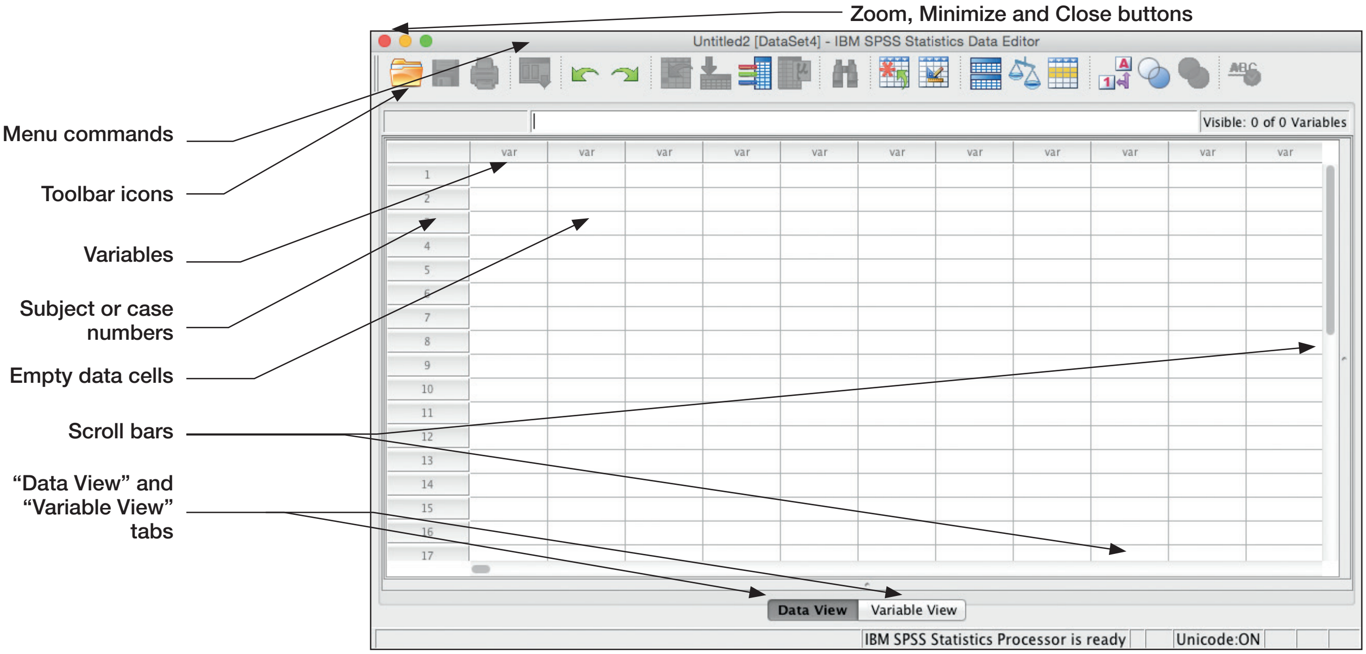
### 2.4.1 The Initial Screen, Icon Detail, and Meaning of Commands

*To access the initial SPSS screen successively click the following icons:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | ✳ 🚀 → ✳ ⬛ → ✳ Σα+ | Front1 |

Screen 2.3 (shown below) is a slightly modified version of Screen 1 on the inside front cover. When you start the SPSS program, Screen 2.3 is the first screen to appear:

**Screen 2.3** Initial Data Screen


Zoom, Minimize and Close buttons
Menu commands
Toolbar icons
Variables
Subject or case numbers
Empty data cells
Scroll bars
"Data View" and "Variable View" tabs
Untitled2 [DataSet4] - IBM SPSS Statistics Data Editor
Visible: 0 of 0 Variables
var
1
2
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

| Icon | Function | Icon | Function | Icon | Function |
|---|---|---|---|---|---|
| | Click this to open a file | | Find data | | (upper left corner) the "+" sign indicates that this is the active file |
| | Save current file | | Insert subject or case into the data file | A 1 | Shifts between numbers and labels for variables with several levels |
| | Print file | | Insert new variable into the data file | | Go to a particular variable or case number |
| | Recall a recently-used command | | Split file into subgroups | | Use subsets of variables/use all variables |
| | Undo the last operation | | Weight cases | | Access information about the current variable |
| | Redo something you just undid | | Select cases | ABC | Spell check |

Screen 2.3 pictures a full-screen image of the data editor, with a detailed breakdown of the toolbar buttons. The menu bar (the commands) and the toolbar are located at the top of the screen and are described below. When you start SPSS, there are no data in the data editor. To fill the data editor window, you may type data into the empty cells (see Chapter 3) or access an already existing data file (described later in this chapter).

**TOOLBAR** The toolbar icons are located below the menu bar at the top of the screen. The icons were created specifically for ease of point-and-click mouse operations. It must be noted that even in SPSS applications the format of the icon bar may vary slightly. The toolbar shown above applies to the Data Editor window; a different toolbar is shown (pages 34–35) that applies to the Output window. Also note that some of the icons are bright and clear and others are "grayed." Grayed icons are those that are not currently available. Note for instance that the Print File icon is grayed because there are no data to print. When data are entered into the data editor, this icon becomes clear because it is now available. The best way to learn how the icons work is to click on them and see what happens.

**THE MENU BAR** The menu bar (just above the toolbar) displays the commands that perform most of the operations that SPSS provides. You will become well acquainted with these commands as you spend time in this book. Whenever you click on a particular command, a series of options appears below and you will select the one that fits your particular need. The commands are now listed and briefly described:

- **SPSS Statistics**: Mostly used to access **Preferences**.
- **File**: Deals with different functions associated with files, including opening, reading, and saving, as well as exiting SPSS.
- **Edit**: A number of editing functions, including copying, pasting, finding, and replacing.
- **View**: Several options that affect the way the screen appears; the option most frequently used is Value Labels.
- **Data**: Operations related to defining, configuring, and entering data; also deals with sorting cases, merging or aggregating files, and selecting or weighting cases.
- **Transform**: Transformation of previously entered data, including recoding, computing new variables, reordering, and dealing with missing values.
- **Analyze**: All forms of data analysis begin with a click of the Analyze command.
- **Graphs**: Creation of graphs or charts can begin either with a click on the Graphs command or (often) as an additional option while other statistics are being performed.
- **Utilities**: Utilities deals largely with fairly sophisticated ways of making complex data operations easier. Most of these commands are for advanced users, and will not be described in this book.
- **Add-ons**: If you want to perform advanced statistics that aren't already in SPSS, these menu options will direct you to other programs and services that SPSS can sell you.
- **Window**: Deals with the position, status, and format of open windows. This menu may be used instead of the taskbar to change between SPSS windows.
- **Help**: A truly useful aid with search capabilities, tutorials, and a statistics coach that can help you decide what type of SPSS procedure to use to analyze your data.

## 2.5 The Open Data File Dialog Window

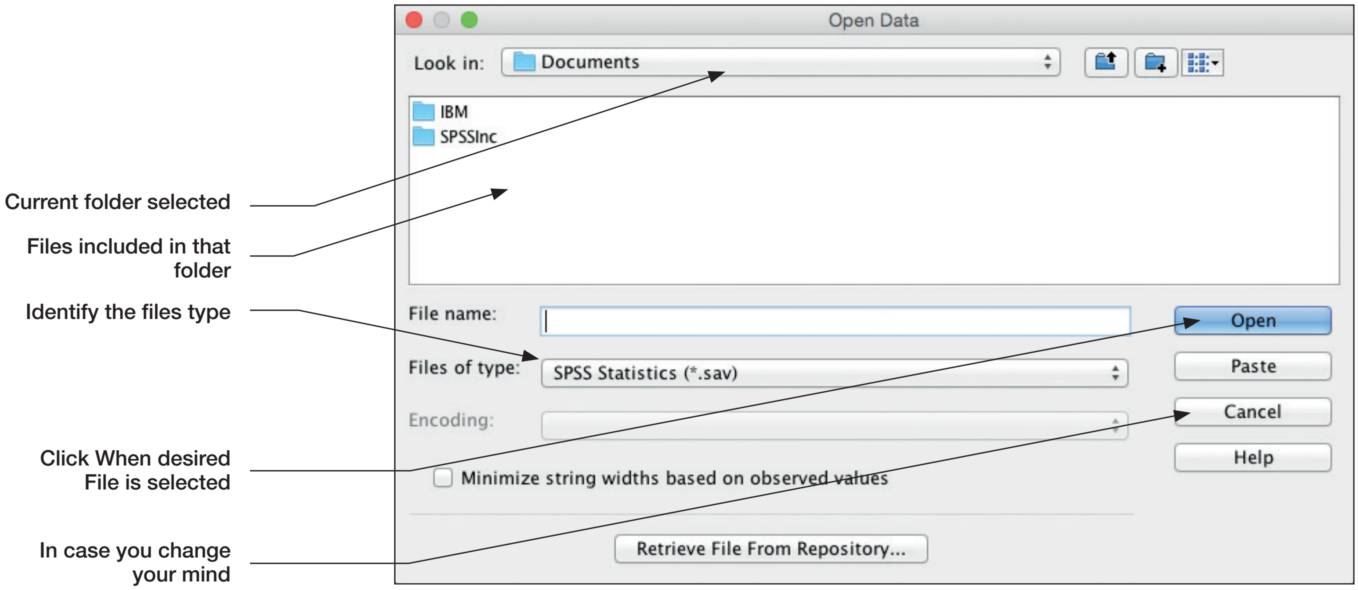
The Open File dialog window provides opportunity to access previously created data files. On the following page we show the Open File window with key elements identified. To access this window, it is necessary to perform the following sequence of steps.

*From Screen 2.3, perform the following procedure to access the Open File window.*

| In Screen | Do This | Step 2 |
|---|---|---|
| 2.3 | 🖱 **File** → **Open** → 🖱 **Data** [*or* 🖱 📂] | 2.4 |

To assist in description of the functions of this window, we will make use of the data file presented in the first chapter (and described in greater detail in Chapter 3). The name of the file is **grades.sav**. All data files created in the IBM SPSS Statistics Data Editor are followed by a period (.) and the three-letter extension, **.sav**; depending on your settings, the **.sav** may or may not appear in the Open Data dialog window.

**Screen 2.4** Open Data Dialog Window


Open Data
IBM
SPSSInc
File name:
Encoding:
□ Minimize string widths based on observed values
Open
Paste
Cancel
Help
Retrieve File From Repository...

Look in: Documents

Files of type: SPSS Statistics (*.sav)

Current folder selected

Files included in that folder

Identify the files type

Click When desired File is selected

In case you change your mind

If you wish to open a data file, there are two simple ways this may be accomplished.

*If the* **grades.sav** *file is visible in the list of files:*

| In Screen | Do This | Step 3 |
|---|---|---|
| 2.4 | ✳ ✳ **grades.sav** | Menu |

*A second option is to simply type the file name in the* **File name** *box, making sure the folder and disk drive (above) are correct, and then click on the* **Open** *button:*

| In Screen | Do This | Step 3a |
|---|---|---|
| 2.4 | type `grades.sav` → ✳ **Open** | Menu |

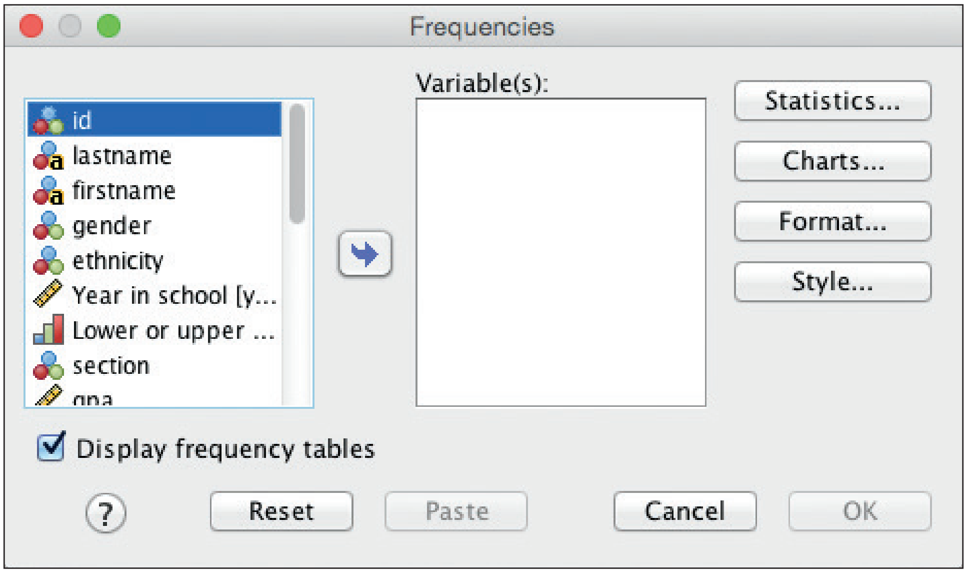
If you are reading a data file that was not created on the SPSS for Windows data editor, you need to identify the file type before SPSS can read it accurately. If, for instance, you are reading an Excel file (**grades.xls** in this case) your file will not appear unless you click the ▼ to the right of **Files of type** and select **Excel (*.xls, *.xlsx, *.xlsm)**. Then the file name should appear in the window above and you can double click to access that file.

The result of either of these operations is to produce a screen with the menu of commands across the top and the data from the **grades.sav** file in the data editor window. Notice (Steps 3 or 3a above) that the menu screen is the outcome (shown to the extreme right in each of the boxes) of each of these operations. The menu of commands across the top of the screen is what allows further analysis to take place.

## 2.5.1 An Example of a Statistical-Procedure Dialog Window

There are as many statistical-procedure dialog windows as there are statistical procedures. Despite the fact that each is different from its fellow, there are fundamental similarities in each one. We illustrate by use of the dialog window for Frequencies (Chapter 6). In all such boxes, to the left will be a list of available variables from the data file. To the right will be another box or even two or three boxes. Although the title of this (or these) box(es) (to the right) will differ, they are designed to show which

Screen 2.5 Sample Statistical Procedure Dialog Window



☑ Display frequency tables

variables will be processed in the analysis. Between the boxes will be → and/or ← buttons to move variables from one box to another.

For all statistical-procedure dialog boxes five buttons are the same; the five buttons are always at the bottom of the dialog box:

- **OK**: When all variables and specifications are selected, click the OK button to compute the statistics.
- **Paste**: A click of the paste button will open a syntax window with the specifications identified in the dialog boxes pasted into the syntax box in a command-file format. Those acquainted with computer programming will enjoy this option. The reason this option is available is because the current format (although simpler) has certain limitations that make it less flexible than the command file format. If you wish to tailor your program to fit your specific needs then it is often useful to work with a command file. Clearly this requires reasonable fluency in SPSS syntax. In SPSS 23 the command file is also printed at the top of the Output section of each analysis.
- **Reset**: One characteristic of these main dialog boxes is called persistence. That means if one analysis is completed and the researcher wishes to conduct another, upon returning to the dialog box, all the same variables and specifications remain from the previous analysis. If that is what you wish, fine. If not, click the Reset button and the variables will all return to the variable list and all specifications will return to the defaults. There is one other way of returning all statistical procedure dialog windows to the default settings: open a different data file (or close and re-open SPSS and open the same data file again).
- **Cancel**: For immediate exit from that procedure.
- ?: Similar to Help on the main command menu except that this help is context specific. If you are working with frequencies, then a click on the help button will yield information about frequencies.

Finally the four buttons to the right of the window (**Statistics, Charts, Format, Style**) represent different procedural and formatting options. Most main-dialog boxes have one or more options similar to these. Their contents, of course, depend on the statistical procedure you are performing.

## 2.5.2 Keyboard Processing, Check Boxes, and Radio Buttons

Keyboard operations can greatly speed up certain SPSS operations. As a starting point the keyboard can be used to negotiate the data screen, in most instances, much more quickly. When you open an SPSS data screen and click on the upper left cell (or any cell for that matter) that cell is highlighted with blue. Here is how different keystrokes allow you to negotiate the screen.

**Tab**: blue highlight moves down

**Return**: highlight moves down

**Cursor keys**: highlight moves up, down, right, or left

**Home**: moves highlight to the first variable in the row

**End**: moves highlight to the last variable in the row

**⌘-Home** or **⌘-End**: moves highlight to the upper left or lower right corner of the data set

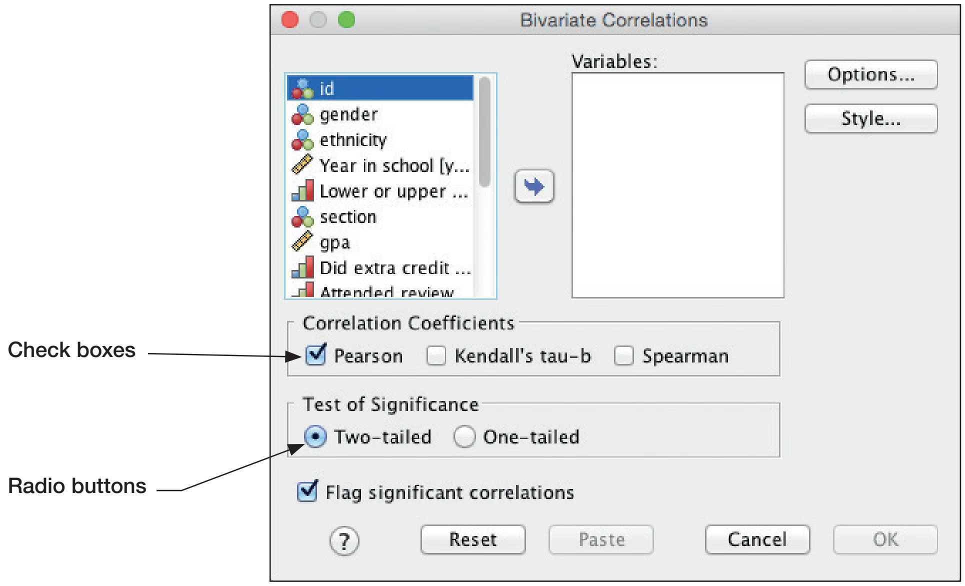
**Shift + cursor keys**: selects blocks of data starting with any highlighted cell

Mac also makes use of the command (⌘) key to allow certain frequently used operations to take place with the use of key strokes. These key strokes are identified by including the command-key stroke combination to the right of the operation. For instance, for the operation “Quit SPSS Statistics” the screen shows Quit SPSS Statistics ⌘Q. This means that you can exit the program either by clicking this button or by pressing the keys **⌘-Q**.

Finally, there are two different types of selections of options that may take place within dialog boxes. Note the window from Chapter 10. Notice the **Test of Significance** box near the bottom; it contains radio buttons. They are called radio buttons because if one is selected the others are automatically deselected (as in a car radio—if you select one channel you deselect the others).

Check boxes are shown under the **Correlation Coefficients** heading and at the very bottom of the dialog box. With check boxes you simply select those you desire. You may select all three, two, one, or none of them. A selection is indicated when a ☑ appears in the box to the left of the option. For radio buttons, click on a different option to deselect the present option. In check boxes, click on the box a second time to deselect it.

**Screen 2.6** Sample Dialog Window with Radio Buttons and Check Boxes



Correlation Coefficients

☑ Pearson □ Kendall's tau-b □ Spearman

Test of Significance

● Two-tailed ○ One-tailed

☑ Flag significant correlations

## 2.6 The Output Window

The "Output" is the term used to identify the results of previously conducted analyses. It is the objective of all data analysis. SPSS has a long and storied history of efforts to create a format of output that is clear yet comprehensive. The current version uses a tables-with-borders format. When utilizing options described below, this is relatively clear, but output can still be very awkward and occupy many pages. It is hoped that the information that follows will maximize your ability to identify, select, edit, and print out the most relevant output. This is an important section since Output is almost always edited before it is printed.

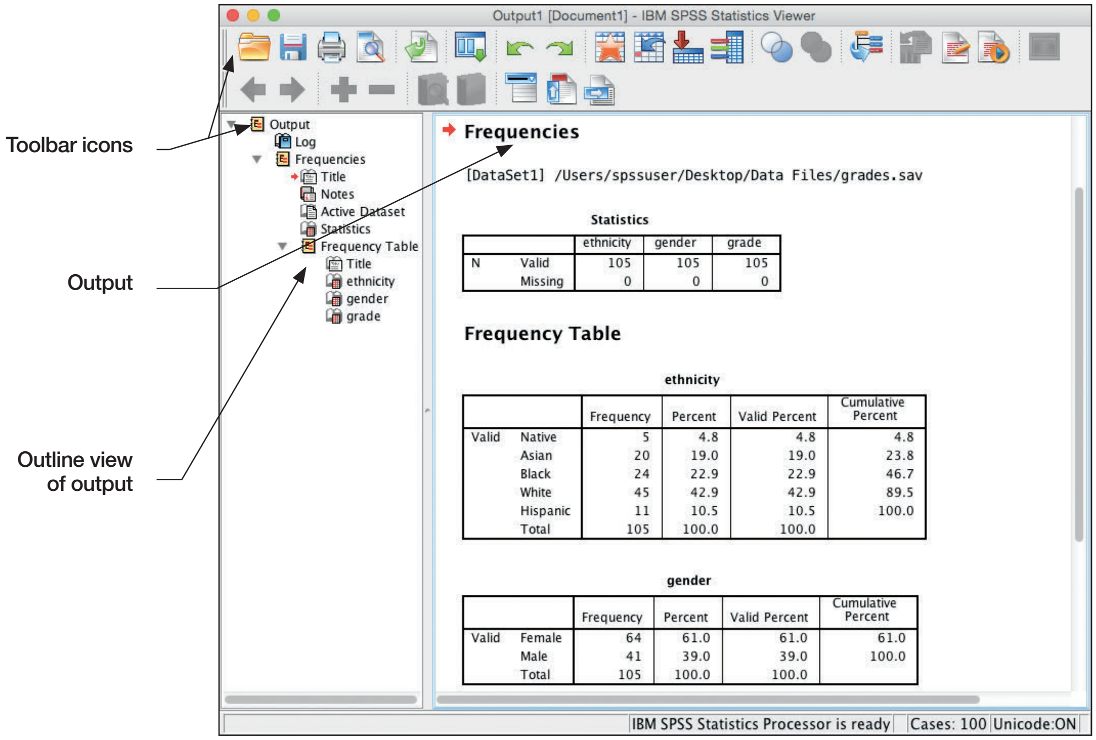
The initial output screen is shown on the inside back cover (and below). Each chapter that involves output will give instructions concerning how to deal with the results displayed on this screen. Our focus here is to explain how to edit output so that, when printed, it will be reproduced in a format most useful to you. Of course, you do not have to edit or reorganize output before you print, but there are often advantages to doing so:

- Extensive outputs will often use/waste many pages of paper.
- Most outputs will include some information that is unnecessary.
- At times a large table will be clearer if it is reorganized.
- You may wish to type in comments or titles for ease or clarity of interpretation.
- SPSS systematically prints the syntax file at the top of every output and the location of your file. You may simply delete this information if you wish.

This chapter will explain the SPSS Output window (and toolbar), how to delete output that you no longer need, how to add comments to your file, how to re-arrange the order of the output, and how to save the output.

Screen 2.7 summarizes the key elements of the output window, as well as providing a detailed description of toolbar items on the output window. In this sample, several procedures have been performed on the **grades.sav** file; don't worry about the content of the analysis, just pay attention to what SPSS will allow you to do with the output.

**Screen 2.7** The SPSS Output Navigator Window



Frequencies

```
[DataSet1] /Users/spssuser/Desktop/Data Files/grades.sav
```

Statistics

| | | ethnicity | gender | grade |
|---|---|---|---|---|
| N | Valid | 105 | 105 | 105 |
| | Missing | 0 | 0 | 0 |

Frequency Table

ethnicity

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Native | 5 | 4.8 | 4.8 | 4.8 |
| | Asian | 20 | 19.0 | 19.0 | 23.8 |
| | Black | 24 | 22.9 | 22.9 | 46.7 |
| | White | 45 | 42.9 | 42.9 | 89.5 |
| | Hispanic | 11 | 10.5 | 10.5 | 100.0 |
| | Total | 105 | 100.0 | 100.0 | |

gender

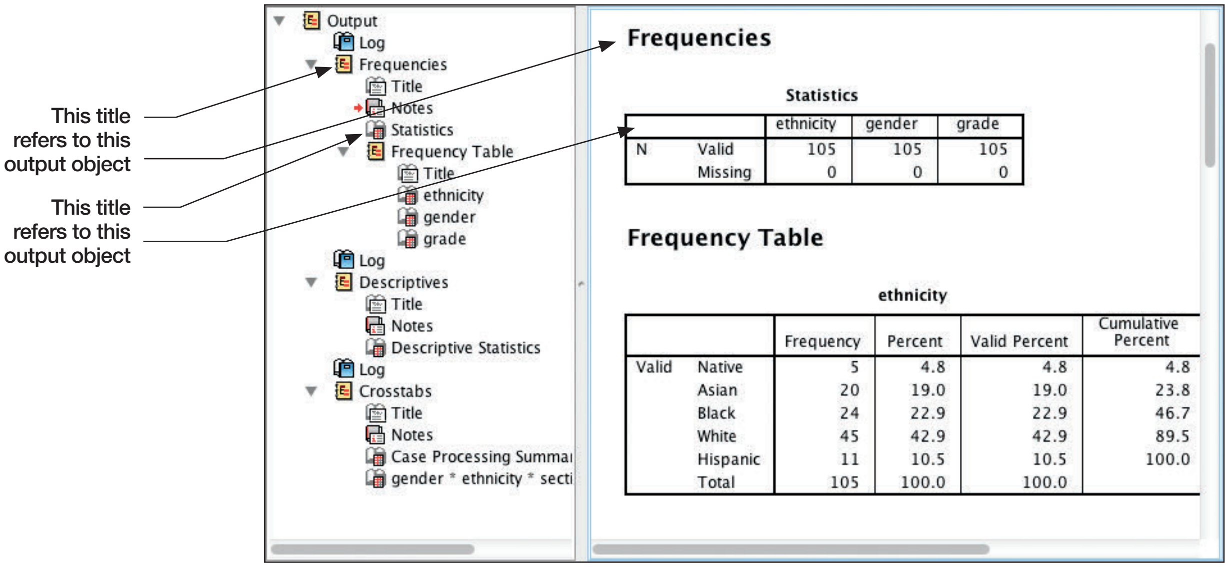
| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Female | 64 | 61.0 | 61.0 | 61.0 |
| | Male | 41 | 39.0 | 39.0 | 100.0 |
| | Total | 105 | 100.0 | 100.0 | |

IBM SPSS Statistics Processor is ready | Cases: 100 | Unicode:ON

| Icon | Function | Icon | Function |
|---|---|---|---|
| | Click to open a file | | Show all variables |
| | Click to save a file | | Select most recent output |
| | Print output | ← | Promote: Make the currently selected output into a heading |
| | Print preview (see what the printout will look like) | → | Demote: Make the currently selected output into a subheading |
| | Export output to text or web (HTML) file | + | Display everything under the currently selected heading |
| | Recall a recently used command | − | Hide everything under the currently selected heading |
| ↶ ↷ | Undo or redo the last operation | | Display the currently selected object (if it is hidden) |
| | Go to SPSS Data Editor (frequently used!) | | Hide the currently selected object (visible in outline view only) |
| | Go to a particular variable or case number | | Insert heading (above titles, tables, and charts) |
| | Get information about variables | | Insert title above output |
| | User-defined sub set of the data | | Insert text at strategic locations within the output |

You will notice that several of the toolbar icons are identical to those in the SPSS Data Editor window; these buttons do the same thing that they do in the Data Editor, but with the output instead of the data. For example, clicking the print icon prints the output instead of the data.

One of the most important things to learn about the SPSS Output window is the use of the outline view on the left of the screen. On the right side of the window is the output from the SPSS procedures that were run, and on the left is the outline (like a table of contents without page numbers) of that output. The SPSS output is actually composed of a series of output objects; these objects may be titles (e.g., “Frequencies”), tables of numbers, or charts among other things. Each of these objects is listed in the outline view:



This title refers to this output object

This title refers to this output object

Frequencies

Statistics

| | | ethnicity | gender | grade |
|---|---|---|---|---|
| N | Valid | 105 | 105 | 105 |
| | Missing | 0 | 0 | 0 |

Frequency Table

ethnicity

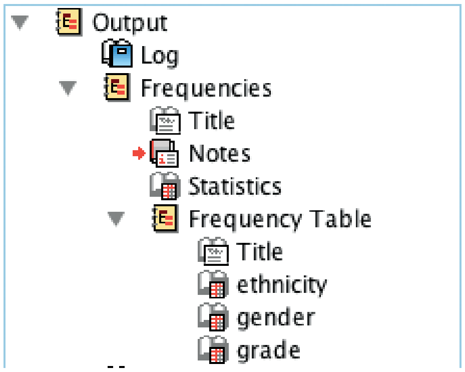
| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Native | 5 | 4.8 | 4.8 | 4.8 |
| | Asian | 20 | 19.0 | 19.0 | 23.8 |
| | Black | 24 | 22.9 | 22.9 | 46.7 |
| | White | 45 | 42.9 | 42.9 | 89.5 |
| | Hispanic | 11 | 10.5 | 10.5 | 100.0 |
| | Total | 105 | 100.0 | 100.0 | |

You will notice that there is no "Notes" section in the output window to correspond with the "Notes" title in the outline view. That's because the notes are (by default) hidden. If you want to see the notes, first click on the word "Notes" and then click on the open-book icon (📖) at the top of the page. The closed-book icon will then become an open-book icon and the notes will materialize in the window to the right. Click the close-book icon (📕) and the notes will disappear and the book will close.

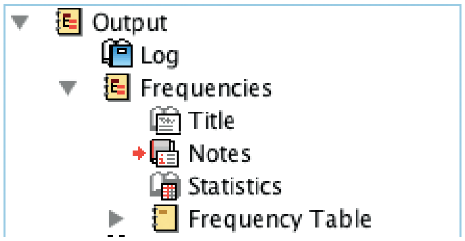
The outline view makes navigating the output easier. If you want to move to the Crosstabs output, for example, you merely need to click on the word "Crosstabs" in the outline view, and the crosstabs will appear in the output window. If you want to delete the Descriptives section (perhaps because you selected an incorrect variable), simply click on the word "Descriptives" (to select that menu item) and then click Delete. If you want to move some output from one section to another (to re-arrange the order), you can select an output object (or a group of output objects), and select Edit and then click Cut. After that, select another output object below which you want to place the output object(s) you have cut. Then select Edit and click Paste After.

If you have been working with the same data file for a while, you may produce a lot of output. So much output may be produced, in fact, that it becomes difficult to navigate through the output even with the outline view. To help with this problem, you can "collapse" a group of output objects underneath a heading. To do this, click on the ▼ sign to the left of the heading. For example, if you want to collapse the Frequencies output (to get it out of the way and come back to it later) simply click on the ▼ to the left of the title Frequency Table. If you want to expand the frequencies heading later, all you have to do is click on the ▼ to the left of the title Frequency Table. This operation is illustrated below.

Clicking on the lower ▼ . . .



. . . Will collapse the Frequency Table Objects to this



One very useful command when you are working with output is the insert-text command (📝). When you click this button, an SPSS Text object is inserted. In this box, you can type comments to remind yourself what is interesting about the SPSS output (for example, "The next chart would look good in the results section" or "Oh, no, my hypothesis was not supported!"). Once you have typed your comments, click on another SPSS object to deselect the SPSS Text object.

Now that you have learned the basics of getting around within SPSS output, there is one more major concept to learn. When SPSS produces output, whether it be charts or tables, you can make changes in the format of that output. Although this book will not describe all of the procedures that you can perform to modify the appearance of charts and tables, some basic operations will be described that will allow you to make your charts and tables more attractive or easier to read. The basics of editing charts and graphs are described in Chapter 5; the basic processes of modifying tables are described here.

## 2.7 Modifying or Rearranging Tables

First, we will start with some background and theory of how tables work. It is important to realize that, although SPSS automatically arranges tables for you—it decides what to put in the columns and rows—the arrangement of a table is somewhat arbitrary. For example, in the crosstabulation of **gender × ethnicity × section** (described in Chapter 8), by default SPSS places the ethnicity in the columns, and section and gender in rows:

### gender * ethnic * section Crosstabulation

**Count**

| | | | ethnic | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Section** | | | **Native** | **Asian** | **Black** | **White** | **Hispanic** | **Total** |
| 1 | gender | female | | 6 | 2 | 9 | 3 | 20 |
| | | male | | 1 | 5 | 7 | 0 | 13 |
| | Total | | | 7 | 7 | 16 | 3 | 33 |
| 2 | gender | female | 3 | 5 | 6 | 11 | 1 | 26 |
| | | male | 1 | 4 | 1 | 7 | 0 | 13 |
| | Total | | 4 | 9 | 7 | 18 | 1 | 39 |
| 3 | gender | female | 1 | 2 | 6 | 6 | 3 | 18 |
| | | male | 0 | 2 | 4 | 5 | 4 | 15 |
| | Total | | 1 | 4 | 10 | 11 | 7 | 33 |

Although there is nothing wrong with this arrangement, you might want to rearrange the table, depending on what exactly you are trying to demonstrate. It is important to remember that you are not changing the content of the statistics present within the table, you are only rearranging the order or the format in which the data and statistics are displayed.

You will notice in the table (above that), there are both rows and columns. The columns contain the various categories of **ethnic**. The rows contain two different levels of headings: In the first level is the course **section**, and in the second level is the **gender** of the students. You will notice that each level of the second heading (each gender) is displayed for each level of the first heading (each section). This table has one layer: The word "**Count**" (the number of subjects in each category) is visible in the upper left hand corner of the table. In most tables, there is only one layer, and we will focus our discussion here on arranging rows and columns.

In order to edit and change the format of an output table, you must first double click on the output object in the output view (the right portion of Screen 2.7). When you do that, the toolbars will disappear, and in the menu bar will appear a new item: Pivot. In addition to this, some of the options available on the various menus change. We will describe some of the options available on the Pivot and **Format** menus, including **Transpose Rows and Columns, Pivoting Trays, Rotate Inner Column Labels**, and **Rotate Outer Row Labels**.

You will notice in the sample table shown above that the columns are much wider than they need to be to display the numbers. This is so the columns will be wide enough to display the labels. Clicking on **Format** followed by **Rotate Inner Column Labels** will fix this problem by making the column labels much taller and narrower, as shown to the right. If you decide you prefer the previous format, simply click on **Format** followed by **Rotate Inner Column Labels** once again, and the old format will be restored. It is also possible to change the orientation of the row titles. This makes the section numbers tall and narrow. But this also makes the heading "Section" tall and narrow, and renders the table itself narrower.

When SPSS produced the sample **gender × ethnic × section** table, it placed ethnicity in the columns, and course section and gender in the rows. A researcher may want these variables displayed differently, perhaps with ethnicity in the rows, and both gender and class section in the columns. In order to make this change, you may use either **Transpose Rows and Columns** or **Pivoting Trays**. Clicking **Pivot** followed by **Transpose Rows and Columns** is the simplest way to rearrange rows and columns.

| ethnic | | | | | |
|---|---|---|---|---|---|
| **Native** | **Asian** | **Black** | **White** | **Hispanic** | **Total** |
| | 6 | 2 | 9 | 3 | 20 |
| | 1 | 5 | 7 | 0 | 13 |
| | 7 | 7 | 16 | 3 | 33 |
| 3 | 5 | 6 | 11 | 1 | 26 |
| 1 | 4 | 1 | 7 | 0 | 13 |
| 4 | 9 | 7 | 18 | 1 | 39 |

If we were to choose this command on our sample output table, then the rows would become columns and the columns rows, as shown below:

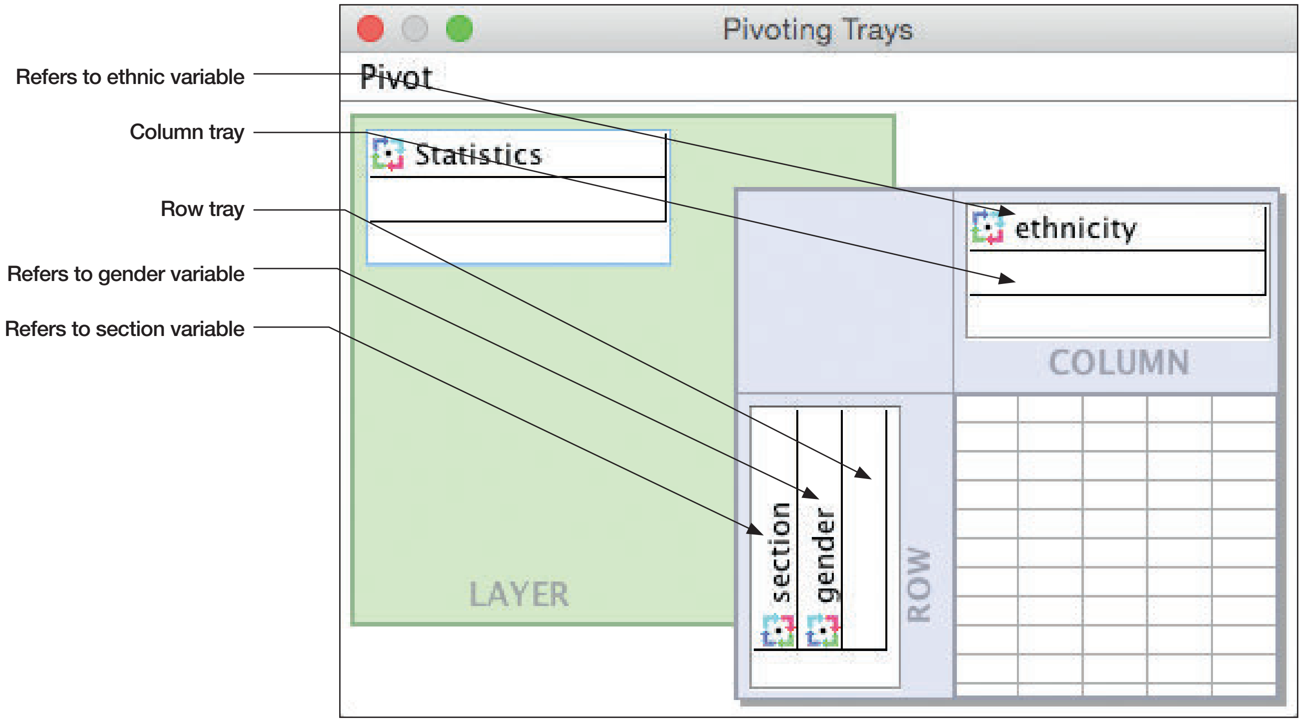
### gender * ethnic * section Crosstabulation

**Count**

| | | section | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | | | 2 | | | 3 | | |
| | | gender | | Total | gender | | Total | gender | | Total |
| | | female | male | | female | male | | female | male | |
| ethnic | Native | | | | 3 | 1 | 4 | 1 | 0 | 1 |
| | Asian | 6 | 1 | 7 | 5 | 4 | 9 | 2 | 2 | 4 |
| | Black | 2 | 5 | 7 | 6 | 1 | 7 | 6 | 4 | 10 |
| | White | 9 | 7 | 16 | 11 | 7 | 18 | 6 | 5 | 11 |
| | Hispanic | 3 | 0 | 3 | 1 | 0 | 1 | 3 | 4 | 7 |
| Total | | 20 | 13 | 33 | 26 | 13 | 39 | 18 | 15 | 33 |

A somewhat more complex and sophisticated way of rearranging columns and rows is to use **Pivoting Trays**. Selecting this option (a subcommand on the **Pivot** menu, or double-click on the table in the output window) will produce a new window that allows you to rearrange rows and columns (Screen 2.8). This window will also allow you to rearrange layers, but because the arrangement of layers usually doesn't need to be changed, we won't worry about layers in this discussion.

**Screen 2.8** Pivoting Trays Window



The pivoting tray window consists of different trays: a tray for **ROW** variables and a tray for **COLUMN** variables. "Statistics" is a separate **LAYER** and refers to the data that fill the cells. In our example **ethnic**ity is in the column tray, with **section** and **gender** both in the row tray.

You may click and drag any of the variables from an original location to a different location. For example, if you want to rearrange the order of the **section** and **gender** variables, you could click and drag **section** to the blank box to the right of **gender**. This would rearrange the rows so that the gender was the main heading level, with three levels of each section for each of the two levels of gender.

You can also rearrange variables between rows and columns, by dragging a variable icon from one tray to another. For example, if you wanted to make **section** a major column heading with levels of ethnicity displayed for each section, you would drag the **section** variable icon to the same row as the **ethnic** variable and **section** will become the major heading.

The best way to understand pivoting tables is simply to experiment with them. Because of the persistence function, it's not at all difficult to run the analysis again if you get too confused.

Once you have done all of this work making your output look the way you want it to look (not to mention doing the statistical analyses to begin with), you may want to save your work. This, by the way, is quite different from saving the data file (always followed by a "**.sav**" suffix). Output is generally saved with a "**.spv**" suffix. The required sequence step follows:

| In Screen | Do This | Step 4 |
|---|---|---|
| 2.7 | ✳ **File** → ✳ **Save** → **type** `yourname.spv` → ✳ **Save** | Menu |

## 2.8 Printing or Exporting Output

Toward the end of the Step by Step section of each chapter, skeletal information is provided concerning printing of output and/or graphs. Every time this is done the reader is referred to this section in this chapter for greater detail. If you desire more complete information than provided here, please refer to the *SPSS* manuals.

The first time you print, it is a good idea to select **File** and then click the **Page setup** option. This very intuitive dialog box lets you specify whether the orientation of printed material will be portrait (longer dimension is vertical) or landscape (longer dimension is horizontal), the scale (100%, larger, or smaller), and the size and source of your paper.

In order to print your output, the material that you wish to print must be designated—that is, you must select the portions of your output that you want printed. If you know that you want all of your output printed, then you don't have to designate the material prior to printing. We don't, however, recommend making this a habit: Save the trees. (You don't have to hug them if you don't want, but at least think about what you are going to print before you start.)

To print a section of the output you must highlight the portion of the file you wish to reproduce before beginning the print sequence. This may be accomplished by

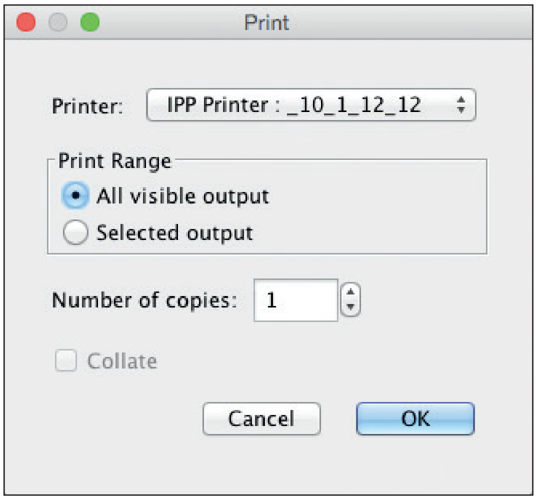
1. Clicking on the folder or output object you want to print on the left side of the SPSS output window;
2. Clicking on a folder or output object on the left side of the SPSS output window, then holding down the **Shift** key while selecting another folder or output object (this will select both objects you clicked and all the objects in between); or
3. Clicking on a folder or object on the left side of the SPSS output window, then holding down the ⌘ key while clicking on one or more additional folders or objects (this will select only the folders or objects that you click on).

Printing starts with a click on the **File** command followed by a click on the **Print** option. This causes a small dialog box to open (Screen 2.9) allowing you to choose which printer you want to use, how many copies to print, and whether to print just the designated (selected) part of the output or the entire output. No formatting options are allowed in this dialog box. You must use the **Page setup** procedure (described in the previous page) prior to printing.

Once all of the options in the **Print** dialog have been specified as desired, click the **OK** button and watch your output pour from the printer.

Of course, you may not find it necessary to print the output on paper. You can always open the output on SPSS and view it on the screen . . . but what if you want to view the output on a different computer, one that doesn't have SPSS? In that case, instead of printing your output, you can export it in PDF format, which can be read on any computer with Adobe Reader or a variety of other free programs. To do that, select **File** and then click the **Export** option. Screen 2.10 will appear.

## Screen 2.9 Print Dialog Box



Printer: IPP Printer : _10_1_12_12

Print Range

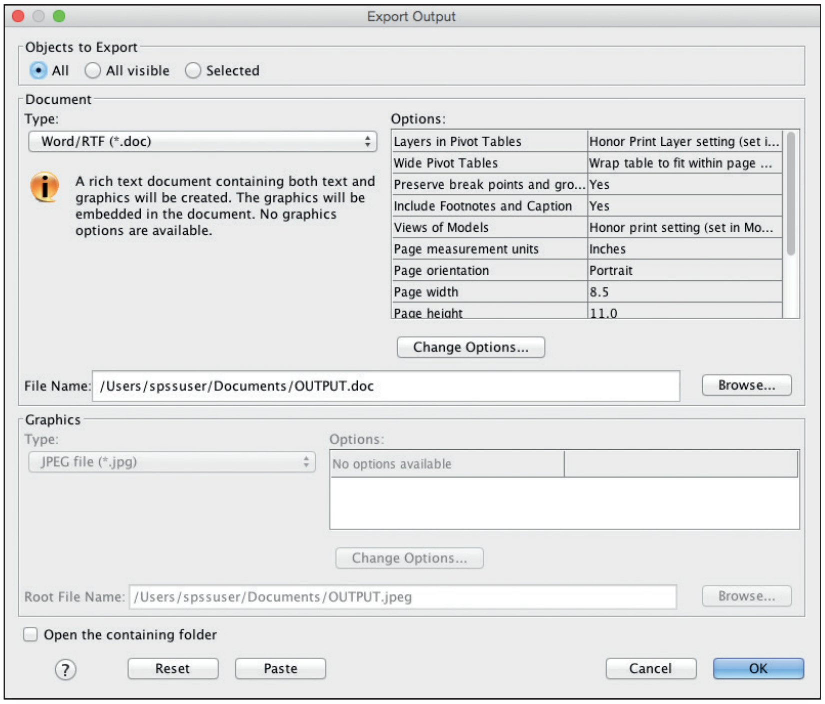
● All visible output

○ Selected output

Number of copies: 1

□ Collate

## Screen 2.10 Export Output



Objects to Export

● All ○ All visible ○ Selected

Document

Type: Word/RTF (*.doc)

A rich text document containing both text and graphics will be created. The graphics will be embedded in the document. No graphics options are available.

Options:

| | |
|---|---|
| Layers in Pivot Tables | Honor Print Layer setting (set i... |
| Wide Pivot Tables | Wrap table to fit within page ... |
| Preserve break points and gro... | Yes |
| Include Footnotes and Caption | Yes |
| Views of Models | Honor print setting (set in Mo... |
| Page measurement units | Inches |
| Page orientation | Portrait |
| Page width | 8.5 |
| Page height | 11.0 |

File Name: /Users/spssuser/Documents/OUTPUT.doc

Graphics

Type: JPEG file (*.jpg)

Options: No options available

Root File Name: /Users/spssuser/Documents/OUTPUT.jpeg

□ Open the containing folder

You can export your output to many formats—Microsoft Word, HTML, Excel, PowerPoint—but it is usually best to select **Type → Portable Document Format (*.pdf)**. Other formats are often difficult to read. Then you merely have to tell SPSS where to save your file and what to name it (by clicking the **Browse** button), and you will have an output that you can view anywhere.

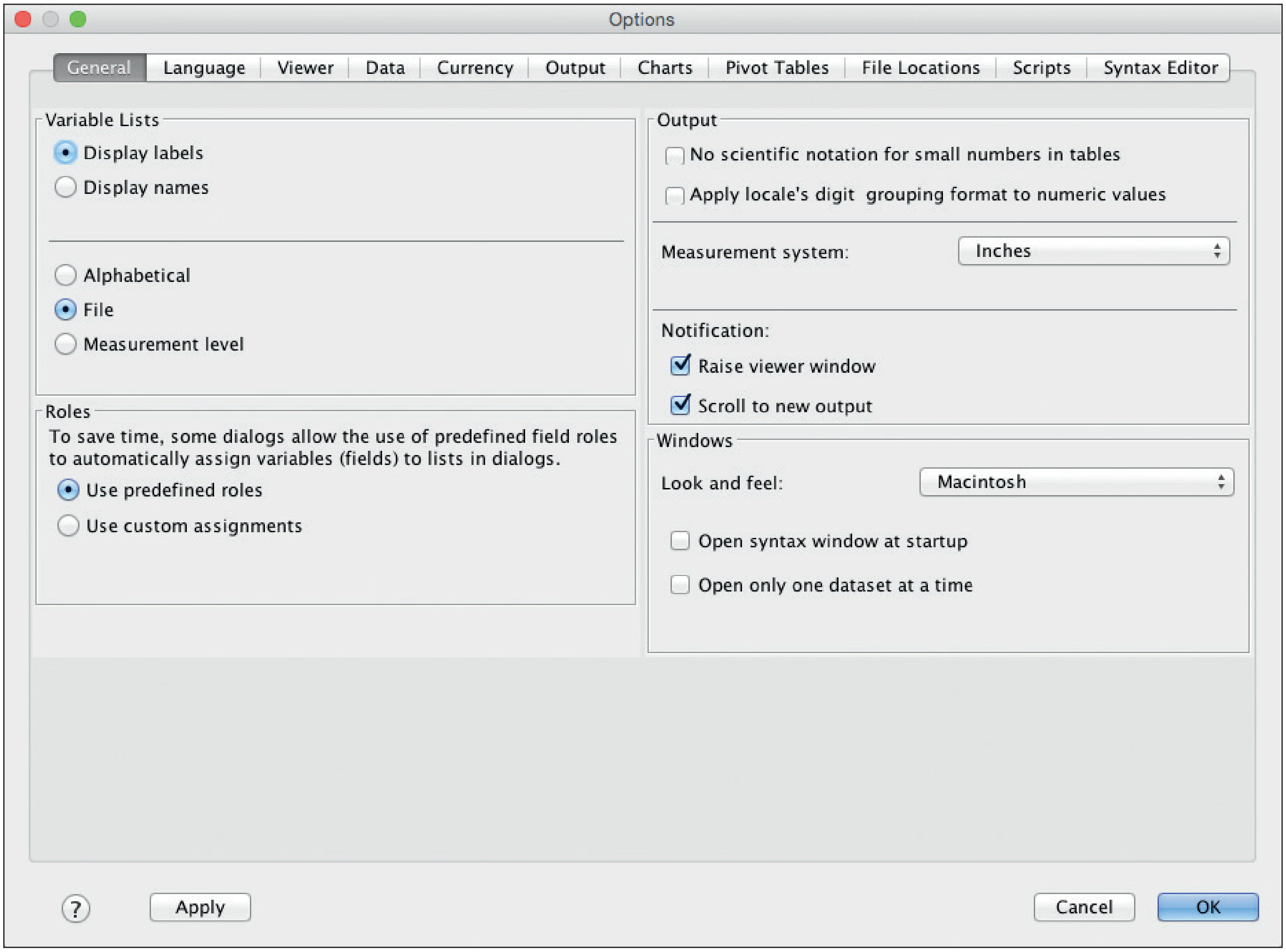
## 2.9 The "Options . . ." Option: Changing the Formats

Under the **Edit → Options . . .** (or **SPSS Statistics → Preferences . . .**) sequence lies a wide variety of alternatives to SPSS defaults. An awareness of the available options allows you to format many features associated with the appearance of the SPSS screen, dialog windows, and output. We reproduce the dialog window here and then identify the selections that we have found to be the most useful.

Notice (Screen 2.11) that there are 11 more tabs across the top; each deals with a different set of functions. Here are the options we use most frequently:

**General** tab (currently visible below): Under **Variable Lists**, you may find it useful to **Display names** (rather than labels); it makes your output more informative, particularly if you have a lot of variables with names that aren't obvious. It is sometimes

**Screen 2.11** Options Window



Variable Lists

● Display labels
○ Display names

○ Alphabetical
● File
○ Measurement level

Roles

To save time, some dialogs allow the use of predefined field roles to automatically assign variables (fields) to lists in dialogs.

● Use predefined roles
○ Use custom assignments

Output

□ No scientific notation for small numbers in tables
□ Apply locale's digit grouping format to numeric values

Measurement system: Inches

Notification:

☑ Raise viewer window
☑ Scroll to new output

Windows

Look and feel: Macintosh

□ Open syntax window at startup
□ Open only one dataset at a time

useful to select **No scientific notation for small numbers in tables**. If you know the data file well, stay with the default **File** order so that variables will be listed in the original order. If you are dealing with a large data file (or one that you don't know very well), then **Alphabetical** may be a better solution.

The **Viewer** tab allows you to specify the format of your output by selecting font, font size, or special effects such as bold, underline, italics, and color.

**Output** tab: If your variable labels are long, you will probably want to change the four boxes to the left to (from top to bottom) of **Names, Labels, Names, Labels** rather than the default of all **Labels**. Then click **Apply** to secure the new setting. The default option (all four are **Labels**) often creates a mess by filling your output with cumbersome descriptions rather than precise names.

**Pivot Tables** tab: There are 28 different options that deal with the format of the tables used in your outputs. You may selectively choose those output formats that best suit your needs. For instance in a large table, one of the "compact" formats will allow you to get more information on a single page.

There are others, many others, but this gives you a start.

# Chapter 3

# Creating and Editing a Data File

THIS CHAPTER describes the first step necessary to analyze data: Typing the data into the computer. SPSS uses a spreadsheet for entering and editing data. Variable names are listed across the top columns, case or subject numbers are listed along the left rows, and data may be entered in cells in the middle. Vertical and horizontal scroll bars allow you to maneuver rapidly through even a large file, and there are several automatic *go-to* functions that allow you to identify or find a particular case number or variable quickly.

This is the first chapter in which we use a step-by-step sequence to guide you through the data entry procedure. A data file was introduced in the first chapter that will be used to illustrate the data-entry process. The name of the file is **grades.sav**, and it involves a course with 105 students taught in three different sections. For each student a variety of information has been recorded, such as quiz and final exam scores. The content of this entire file is listed on the last pages of this chapter and provides an opportunity to enter the file into your own computer. This file is available for download (along with the other data files used in exercises) at ***www.spss-step-by-step.net***. Because of this, you don't have to enter these data yourself if you don't want to, but you may wish to do so (or at least enter the first few students) to gain practice in the data entry procedure.

## 3.1 Research Concerns and Structure of the Data File

Before creating the data file, you should think carefully to ensure that the structure of your file allows all the analyses you desire. Actually, this careful consideration needs to occur much earlier than the data entry phase. We have both had individuals come to us for help with data analysis. They are often working on some large project (such as a dissertation) and have collected data from hundreds of subjects that assess hundreds of variables. Clearly the thought of what to actually *do* with these data has never occurred to them. If such individuals are wealthy and not our students, we're in for a nice little financial benefit as we attempt to tidy up the mess they have created. On other occasions, their lack of planning makes it impossible to make important comparisons or to find significant relationships. Examples of common errors include failure to include key variables (such as gender or age) when such variables are central to the study; requesting yes-no answers to complex personal questions (irritating the subjects and eliminating variability); including many variables without a clear *dependent* variable to identify the objective of the study; having a clear dependent variable but no independent variables that are designed to influence it; and a variety of other outrages.

If the original research has been carefully constructed, then creating a good data file is much easier. For instance, many studies use **ordinal data** (that is, it has an intrinsic order such as seven levels of income from low to high, or levels of education from junior high to Ph.D.). Ordinal data should be coded with numbers rather than letters to allow SPSS to conduct analyses with such data. Another example: If you have missing values (as almost all studies do), it is far better to deal with these at the data-entry stage rather than resorting to SPSS defaults for automatic handling of such. A simple example illustrates one way to handle missing data: If you have an ethnicity variable, several subjects may not answer such a question or may be of an ethnic background that doesn't fit the categories you have created. If you have provided four categories such as White, Black, Asian, and Hispanic (coded as 1, 2, 3, and 4), you might create an additional category (5 = Other or Decline to State) when you enter the data. Ethnicity then has no missing values and you can make any type of ethnic comparisons you wish. The issue of missing values is dealt with in greater detail in Chapter 4.

The order in which you enter variables also deserves some thought before you enter the data. One of us had an individual come with a data file 150 variables wide and the demographics (such as gender, age, and marital status) were positioned at the end of the file. Since demographics were important to the interpretation of the data, they had to be moved back to the beginning of the file where they were more accessible. In determining the order for data entry, IDs and demographics are usually placed at the beginning of the file, and then entry of other variables should follow some sort of logical order.

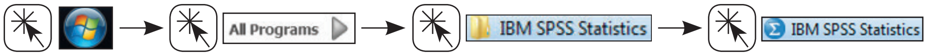
It is also important to think about the way your data were collected, and how that influences the way your file should be structured. An important rule to remember is that each **row** in your data file represents one **case**. Usually, one case is one person or subject, but sometimes a case may be more than one subject (for example, if a husband and wife rate each other, that may be one case). So, if you collect data from one person for several different variables (for a correlational analysis, paired-samples *t* test, or within-subjects ANOVA) then all of the data from that person should be entered in one row. Specifically, each **column** represents a separate **variable** and the variable labels at the top of the columns should make it clear what you are measuring.

The **grades.sav** sample data file illustrates: There are 105 cases in 105 rows representing 105 students. Case #1 in row #1 is the first student (Maria Ross). Case #2 in row #2 is the second student (Mihaela Shima) and so forth. The variables in the **grades.sav** are listed across the top in a useful order. For instance the five quizzes are listed in order (**quiz1, quiz2, quiz3, quiz4, quiz5**) for ease of access when we are performing analyses.

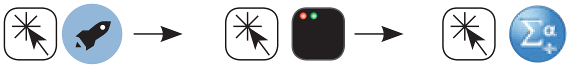
If, however, you collect data from different subjects in different conditions (for an independent-samples *t* test or between-subjects ANOVA, for example) each subject, as above, should be listed on a different row, but a variable in a column should make it clear which condition is involved for each subject. For instance, in the sample data file (as in pretty much all data files) the **gender** variable categorizes subjects into the "female" category or the "male" category. A similar principle applies to the section variable (**section**) which separates subjects into three different categories, "section 1 ," "section 2 ," and "section 3 ."

## 3.2 Step by Step

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 |  | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

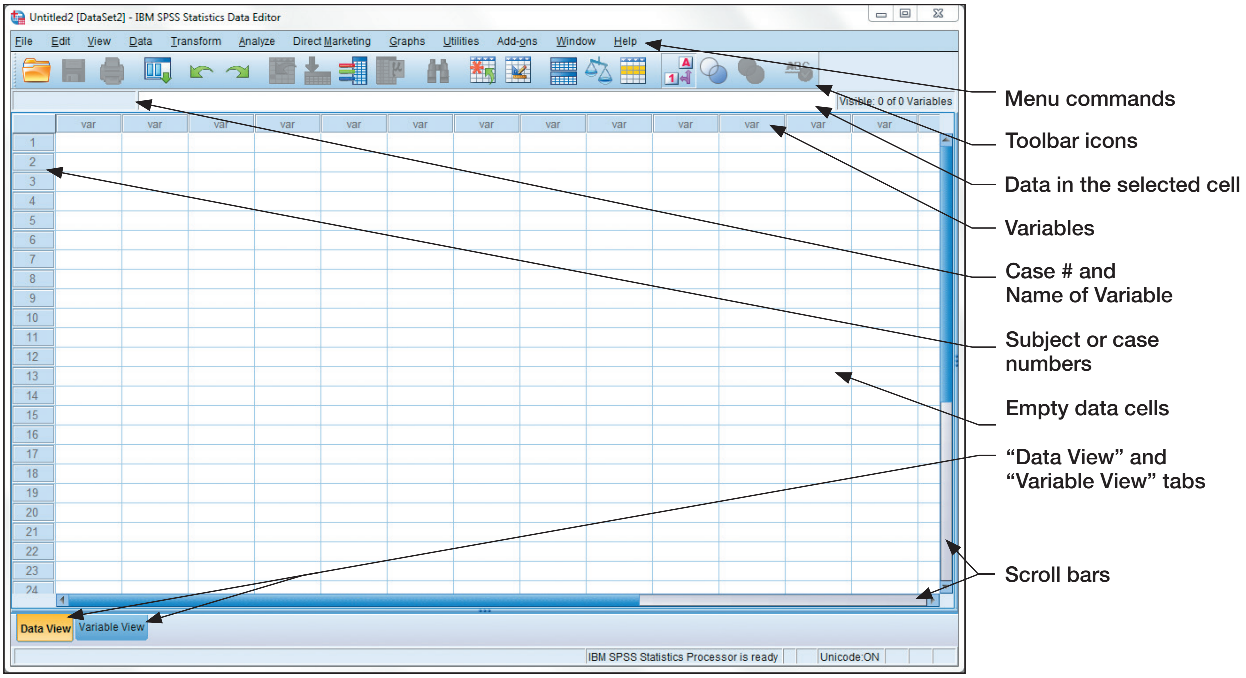
| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 |  | Front1 |

Your screen should now look something like Screen 3.1 (following page). Critical elements important for data entry are labeled.

*With a simple click on the Variable View tab (at the bottom of the screen) the second screen (Screen 3.2, below) emerges that allows you to create your data file. The Step by Step box (requiring only one click) follows.*

| In Screen | Do This | Step 2 |
|---|---|---|
| 3.1 | **Variable View** *tab* | 3.2 |

## Screen 3.1 Data view screen


Untitled2 [DataSet2] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 0 of 0 Variables
var var var var var var var var var var var var var
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

Menu commands

Toolbar icons

Data in the selected cell

Variables

Case # and
Name of Variable

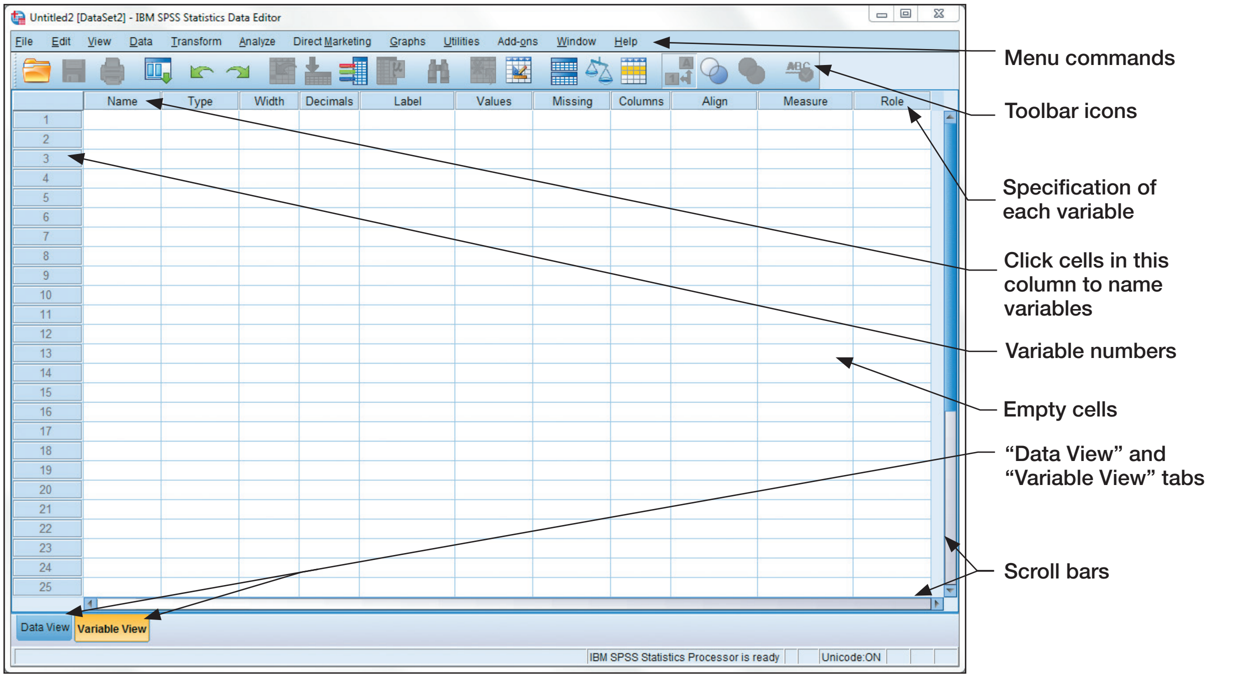
Subject or case
numbers

Empty data cells

“Data View” and
“Variable View” tabs

Scroll bars

## Screen 3.2 *Variable view screen*


Untitled2 [DataSet2] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Name Type Width Decimals Label Values Missing Columns Align Measure Role
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

Menu commands

Toolbar icons

Specification of
each variable

Click cells in this
column to name
variables

Variable numbers

Empty cells

“Data View” and
“Variable View” tabs

Scroll bars

The two screens serve distinct and complementary purposes: **Screen 3.1** is designed to enter data after the data file has been created. **Screen 3.2** is designed to name, label, and determine specifications for each variable. Notice the 11 words in the row near the top of Screen 3.2 labeled “Specifications of each variable.” In case you left your spectacles at school (and thus can't read the fine print) they are **Name, Type, Width, Decimals, Label, Values, Missing, Columns, Align, Measure,** and **Role**. This is the order (with one minor exception) in which we will present the various processes associated with setting up a data file.

We now take you step by step through the procedures necessary to enter the data included in the **grades.sav** file. The entire file is listed beginning on page 55 for sake of reference. While there are a number of different *orders* in which you might execute the steps to create a data file, for the sake of simplicity we will have you first enter all variable names, then format the variables, and then enter all of the data. If you want to conduct data analysis before all of the data are entered, however, there is nothing wrong with entering the names and formats for several variables, entering the data for those variables, performing data analysis, and then continuing with the next set of variables. A file of class records provides an excellent example of this kind of data entry: You would enter data by variables (quizzes, homework assignments, etc.) as they are completed.

## 3.2.1 Name

Beginning with the variable view screen (Screen 3.2), simply type the names of your variables one at a time in the first column. After a name is typed, you may use the cursor keys or the tab to move to the next cell to type the next variable name. There are 17 variable names in the original **grades.sav** data file. You may begin the process by typing all 17 names in the first 17 rows of the Variable View screen. Here's step by step on the first five variables.

*We continue on from sequence Step 2 (on page 45) beginning with Screen 3.2.*

| In Screen | Do This | Step 3 |
|---|---|---|
| 3.2 | 🖱 *first cell under* **Name** *(upper left)* → | |
| | **type** `id` → 🖱 *next cell in the* **Name** *column* | |
| | **type** `lastname` → 🖱 *next cell* | |
| | **type** `firstname` → 🖱 *next cell* | |
| | **type** `gender` → 🖱 *next cell* | |
| | **type** `ethnic` → 🖱 *next cell* | |
| | *. . . and so on for the next 12 variables* | |

You will notice that when you initially press **Home** to the right of the variable name just typed, 9 of the other 10 columns fill in with default information. The default settings will typically need to be altered, but that can be done easily after you have typed the names of each of the (in this case) 17 variables.

Each variable name you use must adhere to the following rules. There was a time when these rules made sense. Now, you just have to follow them.

- Each variable may be any length but shorter than 10 characters is usually desirable.
- It must begin with a letter, but after that any letters, numbers, a period, or the symbols @, #, _, or $ may be used. However, the name may not end with a period.
- All variable names must be unique; duplicates are not allowed.
- Variable names are not sensitive to upper or lower case. **ID**, **Id**, and **id** are all identical to SPSS.

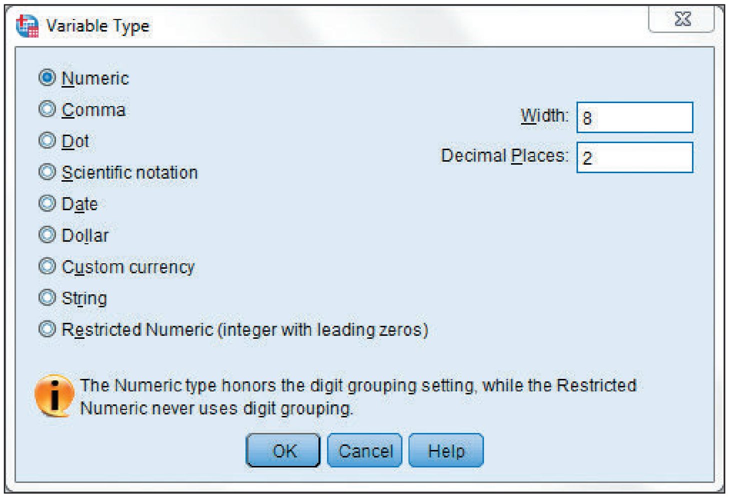
- Because they have a unique meaning in SPSS, certain variable names may not be used, including: **all, ne, eq, to, le, lt, by, or, gt, and, not, ge,** and **with**.

You now have all 17 variable names in the left column. We will now specify how to configure each variable to facilitate the most effective use of your data.

### 3.2.2 Type

When you click on a cell in any row under **Type** you notice a small grayed box (…) to the right of the word **Numeric**. A click on the box opens up Screen 3.3.

**Screen 3.3** Variable Type window



Variable Type

● Numeric
○ Comma
○ Dot
○ Scientific notation
○ Date
○ Dollar
○ Custom currency
○ String
○ Restricted Numeric (integer with leading zeros)

Width: 8
Decimal Places: 2

The Numeric type honors the digit grouping setting, while the Restricted Numeric never uses digit grouping.

Notice that **Numeric** is selected. Most of your variables *will* be numeric and you will retain the default setting. There are two nonnumeric variables in the **grades.sav** file, **lastname** and **firstname**. For those two variables you will click on the grayed box to select **String.**

A variable that contains letters (rather than only numbers) is called a *string* variable. String variables may contain numbers (e.g., **type2**, **JonesIII**) or even consist of numbers, but SPSS treats strings as nonnumeric, and only a very limited number of operations may be conducted with strings as variables. You may notice (see Screen 3.3) that eight different variable types are available. **Numeric** and **String** are by far the most frequently used options and the others will not be considered here. What follows is a step by step of how to change the **lastname** and **firstname** to string variables.

*We continue on from sequence Step 3 (on page 46) beginning with Screen 3.2.*

| In Screen | Do This | Step 4 |
|---|---|---|
| 3.2 | ✳ *cell to the right of the* **lastname** *variable* → ✳ … | |
| 3.3 | ✳ **String** → ✳ **OK** | |
| 3.2 | ✳ *cell to the right of the* **firstname** *variable* → 🖱 … | |
| 3.3 | ✳ **String** → ✳ **OK** | 3.2 |

### 3.2.3 Decimals

We consider **Decimals** before **Width** because if you select a width of 2 or 1, the default value for decimals (2) will prevent your selection of those values, and your computer will squawk most unhappily. The function of the **Decimal** column is to identify the number of decimal places for each variable. In the **grades.sav** file only the **GPA** variable requires decimals (the default 2 places), so all other variables will be assigned a "0" in the decimals column. For string variables the number of decimals is 0 by default. Since the decimal defaults are already correct for **lastname**, **firstname**, and **gpa**, we show in the following Step by Step box how to adjust four of the variables to 0 digits to the right of the decimal point.

*We continue on from sequence Step 4 (on page 45) beginning with Screen 3.2.*

| In Screen | Do This | Step 5 |
|---|---|---|
| 3.2 | *highlight "2" in* **Decimals** *cell to the right of* **id** → **type** 0 → | |
| | **press** ↓ *cursor key 3 X (to bypass first and last name)* → **type** 0 **(gender)** → | |
| | **press** ↓ *cursor key* → **type** 0 **(ethnic)** → | |
| | **press** ↓ *cursor key* → **type** 0 **(year)** → *continue for variables with "0" decimal* | |

## 3.2.4 Width

In the **Width** column you determine the largest number or longest string that will occur for each variable. For **id** the width will be 6 because all ID numbers in the data set have six digits. For **lastname** we might select 10. While there may be a student with a last name longer than 10 letters, that is typically long enough for identification. For **gpa** the width will be 4: one digit left of the decimal, two digits to the right of the decimal, and one more space for the decimal point.

When you click on the cell under **Width** to the right of any variable name the cell is highlighted and a small ⇕ appears in the right margin of the cell. Use this to increase or decrease width from the default of eight characters, or simply highlight the original number and type in the number you desire (this is the method shown in the Step by Step box below).

*We continue on from sequence Step 5 (above) beginning with Screen 3.2.*

| In Screen | Do This | Step 6 |
|---|---|---|
| 3.2 | *highlight "8" in* **Width** *cell to the right of* **id** → **type** 6 → **press** ↓ *cursor key* → | |
| | **(lastname)** **type** 10 → **press** ↓ *cursor key* → | |
| | **(firstname)** **type** 10 → **press** ↓ *cursor key* → | |
| | **(gender)** **type** 1 → **press** ↓ *cursor key* → | |

## 3.2.5 Label

The **Label** column allows you to label any variable whose meaning is not clear from the variable name. Many times the meaning is clear from the variable name itself (e.g., **id, gender, quiz1, quiz2**) and no label is required. Other times the meaning is NOT clear and a label is very useful. After an analysis, in the Output section, an **Options** selection will allow the label to be listed instead of the variable name to assist in clarity and interpretation.

The cells in the **Label** column are simply text boxes and you type the label you desire. The maximum length is 256 characters, but something that long is very cumbersome. Twenty or thirty characters is usually plenty to get your point across.

As you type a label, the width of the **Label** column will remain the same and the first portion of the label will be hidden. If you want to be able to read the entire label, simply position the cursor at the line to the right of **Label** in the top row. The cursor will become a ↔ and you can adjust the column back to a longer (or shorter) width. Use this procedure freely to custom format the Variable View screen.

*We continue on from sequence Step 6 (on page 48) and create labels for two variables.*

| In Screen | Do This | Step 7 |
|---|---|---|
| 3.2 | 🖱 **Label** *cell to the right of* **year** → **type** `year in school` | |
| | 🖱 **Label** *cell to the right of* **lowup** → **type** `lower or upper division` | |

## 3.2.6 Values

Value labels allow you to identify *levels* of a variable (e.g., **gender**: 1 = female, 2 = male; **marital**: 1 = married, 2 = single, 3 = divorced, 4 = widowed). Entering value labels for variables that have several distinct groups is absolutely critical for clarity in interpretation of output. SPSS can do the arithmetic whether or not you include value labels, but you'll never remember whether a 3 means single or divorced (or maybe it was widowed? I'll bet you had to look!). Another advantage of value labels is that SPSS can display these labels in your data file and in Output following analyses. The pleasure of having labels included in the data file is something that quickly becomes addictive. SPSS allows up to 60 characters for each value label, but clearly something shorter is more practical.

Just like the **Type** option, a click on any cell in the **Values** column will produce the small grayed box (...). A click on this box will produce a dialog box (Screen 3.4, following page) that will allow you to create value labels.

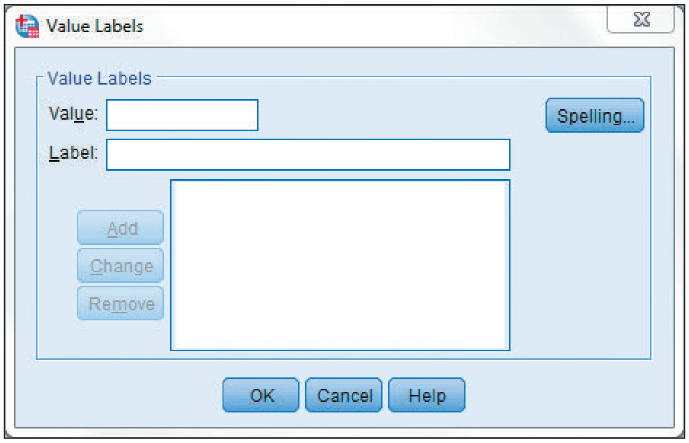
*In the Step by Step box below we show how to use this dialog box to create labels for* **gender** *and* **year***. Again we begin at Screen 3.2.*

| In Screen | Do This | Step 8 |
|---|---|---|
| 3.2 | 🖱 *cell in* **Values** *column in row with the* **gender** *variable* → 🖱 ... | |
| 3.4 | **type** 1 → **press** **Tab** → **type** `female` → 🖱 **Add** | |
| | **type** 2 → **press** **Tab** → **type** `male` → 🖱 **Add** → 🖱 **OK** | |
| 3.2 | 🖱 *cell in* **Values** *column in row with the* **year** *variable* → 🖱 ... | |
| 3.4 | **type** 1 → **press** **Tab** → **type** `frosh` → 🖱 **Add** | |
| | **type** 2 → **press** **Tab** → **type** `soph` → 🖱 **Add** | |
| | **type** 3 → **press** **Tab** → **type** `junior` → 🖱 **Add** | |
| | **type** 4 → **press** **Tab** → **type** `senior` → 🖱 **Add** 🖱 **OK** | 3.2 |

## 3.2.7 Missing

The **Missing** column is rarely used. Its purpose is to designate different types of missing values in your data. For instance, subjects who refused to answer the ethnicity question might be coded 8 and those who were of different ethnicity than those listed might be coded 9. If you have entered these values in the **Missing** value column, then

**Screen 3.4** The Value Labels window



you may designate that 8s and 9s do not enter into any of the analyses that follow. A small dialog box (not reproduced here) opens when you click the … that allows you to specify which values are user-missing.

## 3.2.8 Columns

The **Columns** column, by contrast, is used with most variables. This allows you to identify how much room to allow for your data and labels. The pros and cons on wide versus narrow columns are clear: If you have wide columns you can see the entire variable name and thus seem less crowded. If you have narrow columns, you have the advantage of getting many variables visible on the screen at one time but you may have to truncate variable names.

For instance if you have columns that are 3 characters wide you are able to fit 33 variables into the visible portion of the data screen (depending on your monitor); if the columns are all 8 characters wide (the default), you can fit only 12. Anyone who has spent time working with data files knows how convenient it is to be able to view a large portion of the data screen without scrolling. But many variable names would be truncated and thus unintelligible with only a three-character width. The same rationale applies to the value labels. Note that for the **ethnic** variable it might be more precise to create labels of "American Native or Inuit, Asian or Asian American, African or African American/Canadian, Caucasian, Hispanic," but the brief labels of "Native, Asian, Black, White, and Hispanic" get the point across without clutter.

Once again the ⬍ becomes visible in the right margin when you click a cell in the **Columns** column. You may either click the number higher or lower, or type in the desired number after highlighting the default number, 8.

*We enter the column width for the first five variables in the Step by Step box below. Notice how the width reflects the name of the variable. The starting point, again, is Screen 3.2.*

| In Screen | Do This | Step 9 |
|---|---|---|
| 3.2 | *highlight "8" in* **Columns** *cell to the right of* **id** → **type** 6 → **press** ↓ *cursor key* | |
| | **(lastname)** **type** 10 → **press** ↓ *cursor key* → | |
| | **(firstname)** **type** 10 → **press** ↓ *cursor key* → | |
| | **(gender)** **type** 6 → **press** ↓ *cursor key* → | |
| | **(ethnic)** **type** 5 → **press** ↓ *cursor key, and so forth for all variables* | 3.2 |

### 3.2.9 Align

The **Align** column provides a drop-down menu that allows you to align the data in each cell right, left, or center. By default, numeric variables align to the right, string variables align to the left. You may select otherwise if you wish.

### 3.2.10 Measure

The **Measure** column also provides a drop-down menu that allows you to select three options based on the nature of your data: **Scale**, **Ordinal**, and **Nominal**.

- **Scale** measures (designated with a 📏 icon) have intrinsic numeric meaning that allow typical mathematical manipulations. For instance, age is a scale variable: 16 is twice as much as 8, 4 is half as much as 8, the sum of a 4 and 8 is 12, and so forth. **Scale** is the default for all numeric variables.
- **Ordinal** measures (designated with a 📊 icon) have intrinsic order but mathematical manipulations are typically meaningless. On an aggression scale of 1 to 10, someone higher on the scale is more aggressive than someone lower on the scale, but someone who rates 4 is not twice as aggressive as someone who rates 2.
- **Nominal** measures (designated with a ● icon) are used for identification but have no intrinsic order (lesser to greater) such as gender, ethnicity, marital status, and most string variables. Nominal data may be used for categorization and for a number of other statistical procedures.

Sometimes it can be difficult to choose between scale and ordinal. If so, don't worry too much. In all analysis, SPSS handles both ordinal and scale variables identically.

### 3.2.11 Role

The **Role** column made its appearance for the first time in SPSS 18. This function is designed for large data sets in which the researcher wishes to keep track of which variables are undesignated (**Input**, the default), which are dependent variables (**Target**), and other functions unique to certain experimental designs. For most studies this column may be ignored.

## 3.3 Entering Data

After naming and formatting the variables, entering data is a simple process. We first give you the sequence of steps for saving the data file—an event that should occur frequently during the creation of the file and entry of data. Then we describe two different ways of entering the data; depending on the particular data file you are creating, either may be appropriate.

*When you save the file for the first time you need to enter the file name.*

| In Screen | Do This | Step 10 |
|---|---|---|
| 3.1 | **File** → **Save** → **type** `grades.sav` → **Save** | |

Note: These instructions assume that you want to save the file in the default folder (depending on your particular computer setup this could be a number of different folders). If you don't want to save your work in the default folder, you will need to click the up-one-level button (⇧) one or more times, and double click on the disk drives, network paths, or folders in which you want to save your file.

*After the first time you save the file, saving your changes requires just one or two clicks.*

| In Screen | Do This | Step 11 |
|---|---|---|
| 3.1 | **File** → **Save** [*or* 💾] | |

There are two different ways of entering data that we will describe: by variable and by case or subject. In different settings either method may be desirable, so we will describe both.

**ENTER DATA BY VARIABLE** Click on the first empty cell under the first variable, type the number (or word), press the **Down-arrow** key or **Enter** key**,** then type the next number/word, press the **Down-arrow** or **Enter** key, and so forth. When you finish one variable, scroll up to the top of the file and enter data for the next variable in the same manner.

**ENTER DATA BY CASE OR SUBJECT** Click on the first empty cell for the first subject under the first variable, and then type the first number/word, press the **Right-arrow** key or **Tab** key**,** type the next number, press the **Right-arrow** or **Tab** key, and so forth. When you finish one subject, scroll back to the first column and enter data for the next subject.

## 3.4 Editing Data

Just as data entry is a simple procedure, so also is editing previously entered data. The following options are available:

**CHANGING A CELL VALUE** Simply click on the cell of interest, type the new value, and then hit **Enter**, **Tab**, or any of the **Arrow keys**.

**INSERTING A NEW CASE** If you wish to insert data for a new subject or case, click on the case number above which you would like the new case to be. Then click on the insert-case toolbar icon () and a new line will open and push all other cases down by exactly one line. You may then enter data for the new subject or case.

**INSERTING A NEW VARIABLE** To insert a new variable, click on the variable to the right of where you would like the new variable to be located, click on the insert-variable icon (), and a new column will open (to the left) and push all other variables exactly one column to the right. You may then name and format the new variable and enter data in that column.

**TO COPY OR CUT CELLS** To copy rows or columns of data, first highlight the data you want to copy by a click on the variable name (to highlight all entries for one variable) or click on the case number (to highlight all the entries for one subject or case) and then use the shift and arrow keys to highlight larger blocks of data. Once the cells you want to copy are highlighted, right-click on the highlighted cells and (from the drop down menu) click **Copy** if you wish to *leave* the cells where they are but put them onto the clipboard for future pasting. Click the **Cut** option if you wish to *delete* the cells from the data editor and place them onto the clipboard.

**TO PASTE CELLS** At first glance, pasting cells may seem simply a matter of copying the cells you wish to paste, selecting the location where you wish to paste them, and then clicking **Edit** ➔ **Paste**. This process is actually somewhat tricky, however, for two reasons: (1) If you paste data into already existing cells, you will erase any data already in those cells and (2) SPSS lets you paste data copied from one variable into another variable (this can cause confusion if, for example, you copy data from the **gender** variable into the **marital** status variable). Here are some tips to avoid problems when pasting cells:

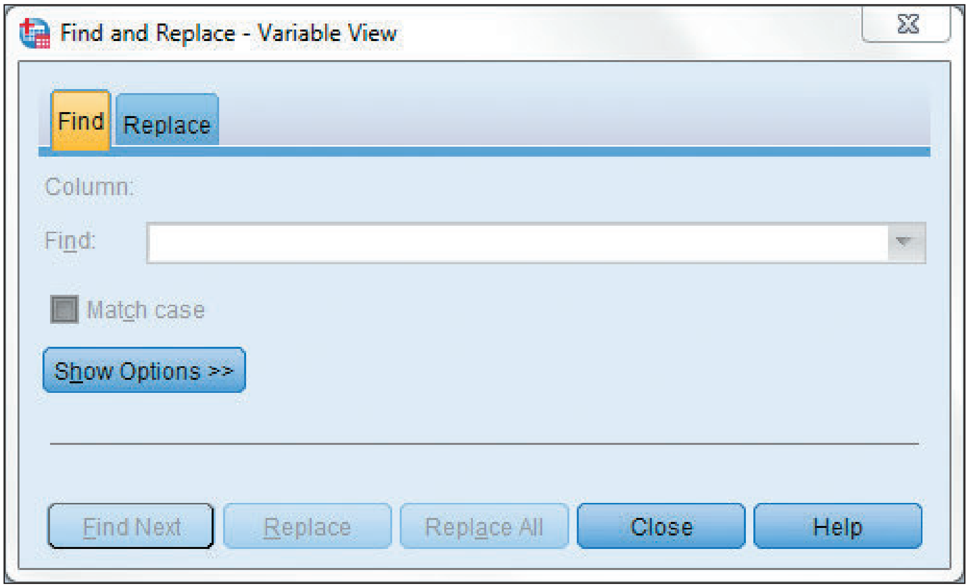
- Save your data file before doing major cutting and pasting, so that you can go back to an earlier version if you make a mistake.
- Create space for the new data by inserting new cases and variables (see instructions above) to make room for the cells you are going to paste.
- In most cases you will cut and paste entire rows (cases) or columns (variables) at the same time. It requires careful attention to paste *segments* of rows or columns.

- Be careful to align variables appropriately when pasting, so that you don't try pasting string variables into numeric variables or inflict other problematic variable mismatches.

**TO SEARCH FOR DATA** One of the handiest editing procedures is the **Find** function. A click on the **Edit** command followed by a click on the **Find** option (or a click on the 🔍 toolbar icon) opens up a screen (Screen 3.5) that allows you to search for a particular word or data value. This function is most frequently used for two different purposes:

- If you have a large file that includes names, you can quickly find a particular name that is embedded within the file.
- If you discover errors in your data file (e.g., a GPA of 6.72), the search function can quickly find those errors for correction.

Screen 3.5 The Find Data window



*If for instance we wish to find a student with the last name Osborne, execute the following sequence of steps:*

| In Screen | Do This | Step 12 |
|---|---|---|
| 3.1 | 🖱 **lastname** → 🖱 **Edit** → 🖱 **Find** [*or* 🖱 🔍] | |
| 3.5 | **type** `Osborne` → 🖱 **Find Next** → 🖱 **Close** | |

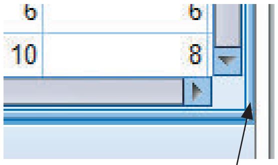
*If you are working with a data file with a large number of subjects, and notice (based on frequency print-outs) that you have three subjects whose scores are outside the normal range of the variable (e.g., for* **anxiety** *coded 1 to 7 you had three subjects coded 0), and you wish to locate these errors, perform the following sequence of steps.*

| In Screen | Do This | Step 13 |
|---|---|---|
| 3.1 | 🖱 **anxiety** → 🖱 **Edit** → 🖱 **Find** [*or* 🖱 🔍] | |
| 3.5 | **type** 0 → 🖱 **Find Next** → 🖱 **Close** | 3.1 |

### TO MAKE SUBJECT IDENTIFIERS ALWAYS VISIBLE

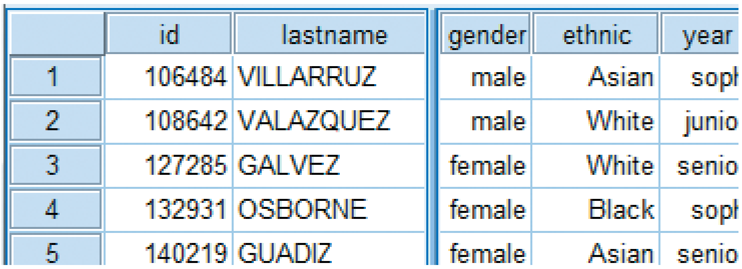
When a file has many variables, it is often useful to make some of the columns (those on the left, if you set the file up as we recommend) always visible, even when you are scrolling over to view or enter variables along the right.

This is easy to do, once you find where to put the cursor. Near the bottom right of the screen, look for a small area to the right of the vertical scroll bar.



*This is what you're looking for*

Drag that line to the left, until it is immediately to the right of any variables you want to always be visible. For example, in the **grades.sav** file, you might move it just to the right of the **lastname** variable. Then, your file would look like this:



| | id | lastname | gender | ethnic | year |
|---|---|---|---|---|---|
| 1 | 106484 | VILLARRUZ | male | Asian | sopl |
| 2 | 108642 | VALAZQUEZ | male | White | junio |
| 3 | 127285 | GALVEZ | female | White | senio |
| 4 | 132931 | OSBORNE | female | Black | sopl |
| 5 | 140219 | GUADIZ | female | Asian | senio |

This part is always visible        This part scrolls

This makes it *much* easier to work with a "wide" file that contains many variables.

## 3.5 Grades.sav: The Sample Data File

The data file is the raw data for calculating the grades in a particular class. The example consists of a single file, used by a teacher who teaches three sections of a class with approximately 35 students in each section. From left to right, the variables that are used in the data file are:

| Variable | Description |
|---|---|
| ID | Six-digit student ID number |
| LASTNAME | Last name of the student |
| FIRSTNAME | First name of the student |
| GENDER | Gender of the student: 1 = female, 2 = male |
| ETHNIC | Ethnicity of the student: 1 = Native (Native American or Inuit), 2 = Asian (or Asian American), 3 = Black, 4 = White (non-Hispanic), 5 = Hispanic |
| YEAR | Year in school; 1 = Frosh (1st year), 2 = Soph (2nd year), 3 = Junior (3rd year), 4 = Senior (4th year) |
| LOWUP | Lower or upper division student: 1 = Lower (1st or 2nd year), 2 = Upper (3rd or 4th year) |
| SECTION | Section of the class (1 through 3) |
| GPA | Cumulative GPA at the beginning of the course |
| EXTCR | Whether or not the student did the extra credit project: 1 = No, 2 = Yes |
| REVIEW | Whether or not the student attended the review sessions: 1 = No, 2 = Yes |
| QUIZ1 to QUIZ5 | Scores out of 10 points on five quizzes throughout the term |
| FINAL | Final exam worth 75 points |

| ID | Lastname | Firstname | Gender | Ethnic | Year | Lowup | Section | GPA | Extrcredit | Review | Quiz1 | Quiz2 | Quiz3 | Quiz4 | Quiz5 | Final |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 973427 | ROSS | MARIA | 1 | 4 | 4 | 2 | 1 | 3.19 | 1 | 2 | 9 | 7 | 10 | 9 | 7 | 65 |
| 390203 | SHIMA | MIHAELA | 1 | 2 | 3 | 2 | 2 | 2.28 | 1 | 2 | 6 | 7 | 9 | 6 | 8 | 61 |
| 703740 | SUNYA | DALE | 2 | 5 | 3 | 2 | 3 | 3.58 | 1 | 2 | 10 | 9 | 10 | 10 | 7 | 62 |
| 354601 | CARPIO | MARY | 1 | 2 | 2 | 1 | 1 | 2.03 | 1 | 2 | 10 | 10 | 10 | 10 | 9 | 71 |
| 979028 | NEUHARTH | JIM | 2 | 4 | 3 | 2 | 3 | 1.80 | 1 | 2 | 3 | 6 | 3 | 4 | 5 | 49 |
| 768995 | DUMITRESCU | STACY | 2 | 4 | 4 | 2 | 2 | 2.88 | 1 | 1 | 7 | 10 | 8 | 9 | 10 | 60 |
| 574170 | HURRIA | WAYNE | 2 | 1 | 2 | 1 | 2 | 3.84 | 1 | 1 | 4 | 5 | 6 | 6 | 6 | 48 |
| 380157 | LUTZ | WILLIAM | 2 | 4 | 3 | 2 | 2 | 2.25 | 2 | 2 | 10 | 9 | 10 | 10 | 8 | 61 |
| 167664 | SWARM | MARK | 2 | 4 | 3 | 2 | 3 | 2.35 | 1 | 2 | 8 | 10 | 10 | 10 | 9 | 71 |
| 245473 | DAYES | ROBERT | 2 | 4 | 3 | 2 | 1 | 2.74 | 1 | 1 | 8 | 9 | 6 | 7 | 10 | 48 |
| 436413 | PANG | SUZANNE | 1 | 2 | 3 | 2 | 1 | 2.66 | 1 | 2 | 8 | 6 | 7 | 8 | 7 | 60 |
| 515586 | FIALLOS | LAUREL | 1 | 4 | 2 | 1 | 2 | 3.90 | 1 | 1 | 7 | 8 | 8 | 6 | 6 | 63 |
| 106484 | VILLARUIZ | ALFRED | 2 | 2 | 2 | 1 | 2 | 1.18 | 1 | 2 | 6 | 5 | 7 | 6 | 3 | 53 |
| 725987 | BATILLER | FRED | 2 | 2 | 2 | 1 | 2 | 1.77 | 1 | 2 | 6 | 7 | 7 | 7 | 5 | 60 |
| 870810 | REYNO | NICHOLAS | 2 | 4 | 3 | 2 | 3 | 3.66 | 2 | 1 | 10 | 8 | 10 | 10 | 10 | 68 |
| 127285 | GALVEZ | JACKIE | 1 | 4 | 4 | 2 | 2 | 2.46 | 2 | 2 | 10 | 7 | 8 | 9 | 7 | 57 |
| 623857 | CORTEZ | VIKKI | 1 | 3 | 4 | 2 | 3 | 2.56 | 1 | 2 | 5 | 7 | 6 | 5 | 6 | 58 |
| 905109 | JENKINS | ERIC | 2 | 3 | 2 | 1 | 3 | 2.84 | 1 | 1 | 6 | 8 | 6 | 6 | 10 | 64 |
| 392464 | DOMINGO | MONIKA | 1 | 4 | 3 | 2 | 3 | 3.02 | 2 | 1 | 10 | 10 | 10 | 9 | 9 | 55 |
| 447659 | GLANVILLE | DANA | 1 | 5 | 4 | 2 | 3 | 2.77 | 1 | 1 | 6 | 8 | 9 | 5 | 8 | 63 |
| 958384 | RONCO | SHERRY | 1 | 4 | 2 | 1 | 1 | 2.30 | 1 | 2 | 10 | 9 | 10 | 10 | 7 | 60 |
| 414775 | RATANA | JASON | 2 | 2 | 3 | 2 | 1 | 2.38 | 1 | 2 | 8 | 9 | 10 | 10 | 9 | 50 |
| 237983 | LEE | JONATHAN | 2 | 2 | 4 | 2 | 2 | 1.66 | 2 | 2 | 5 | 7 | 4 | 7 | 6 | 63 |
| 818528 | CARRINGTON | JYLL | 1 | 4 | 3 | 2 | 1 | 1.95 | 1 | 2 | 9 | 10 | 10 | 8 | 8 | 53 |
| 938881 | YEO | DENISE | 1 | 1 | 3 | 2 | 3 | 3.53 | 1 | 2 | 7 | 10 | 9 | 8 | 9 | 72 |
| 944702 | LEDESMA | MARTINE | 1 | 4 | 3 | 2 | 2 | 3.90 | 1 | 2 | 6 | 7 | 7 | 5 | 9 | 67 |
| 154441 | LIAN | JENNY | 1 | 5 | 2 | 1 | 1 | 3.57 | 1 | 2 | 10 | 9 | 10 | 10 | 10 | 71 |
| 108642 | VALAZQUEZ | SCOTT | 2 | 4 | 3 | 2 | 2 | 2.19 | 2 | 1 | 10 | 10 | 7 | 6 | 9 | 54 |
| 985700 | CHA | LILY | 1 | 4 | 2 | 1 | 1 | 2.43 | 2 | 2 | 10 | 9 | 10 | 10 | 7 | 63 |
| 911355 | LESKO | LETITIA | 1 | 3 | 2 | 1 | 3 | 3.49 | 1 | 2 | 10 | 9 | 10 | 10 | 8 | 71 |
| 249586 | STOLL | GLENDON | 2 | 4 | 3 | 2 | 2 | 2.51 | 1 | 1 | 5 | 9 | 5 | 6 | 10 | 63 |
| 164605 | LANGFORD | DAWN | 1 | 3 | 3 | 2 | 2 | 3.49 | 2 | 1 | 10 | 10 | 9 | 10 | 10 | 75 |
| 419891 | DE CANIO | PAULA | 1 | 4 | 3 | 2 | 2 | 3.53 | 1 | 2 | 6 | 7 | 7 | 9 | 9 | 54 |
| 615115 | VASENIUS | RUSS | 2 | 3 | 3 | 2 | 3 | 1.77 | 1 | 2 | 6 | 7 | 6 | 8 | 6 | 59 |
| 192627 | MISCHKE | ELAINE | 1 | 4 | 1 | 1 | 2 | 2.90 | 1 | 1 | 3 | 8 | 4 | 6 | 8 | 55 |
| 595177 | WILLIAMS | OLIMPIA | 1 | 3 | 3 | 2 | 3 | 1.24 | 1 | 1 | 7 | 6 | 7 | 10 | 5 | 53 |
| 434571 | SURI | MATTHEW | 2 | 2 | 3 | 2 | 2 | 2.80 | 1 | 1 | 7 | 6 | 9 | 8 | 8 | 60 |
| 506467 | SCARBROUGH | CYNTHIA | 1 | 4 | 3 | 2 | 2 | 1.33 | 1 | 2 | 8 | 5 | 6 | 4 | 7 | 58 |
| 546022 | HAMIDI | KIMBERLY | 1 | 5 | 3 | 2 | 1 | 2.96 | 1 | 1 | 7 | 7 | 6 | 9 | 8 | 61 |
| 498900 | HUANG | JOE | 2 | 5 | 3 | 2 | 3 | 2.47 | 1 | 1 | 0 | 5 | 0 | 2 | 5 | 40 |
| 781676 | WATKINS | YVONNE | 1 | 3 | 4 | 2 | 1 | 4.00 | 1 | 2 | 9 | 9 | 10 | 10 | 9 | 70 |
| 664653 | KHAN | JOHN | 2 | 4 | 3 | 2 | 3 | 1.24 | 1 | 2 | 3 | 8 | 5 | 2 | 7 | 59 |
| 908754 | MARQUEZ | CHYRELLE | 1 | 4 | 1 | 1 | 2 | 1.85 | 1 | 2 | 4 | 8 | 5 | 7 | 9 | 57 |
| 142630 | RANGFIO | TANIECE | 1 | 4 | 3 | 2 | 3 | 3.90 | 1 | 2 | 10 | 10 | 10 | 9 | 9 | 74 |
| 175325 | KHOURY | DENNIS | 2 | 4 | 3 | 2 | 1 | 2.45 | 1 | 1 | 8 | 8 | 10 | 10 | 6 | 69 |
| 378446 | SAUNDERS | TAMARA | 1 | 1 | 2 | 1 | 2 | 2.80 | 1 | 2 | 4 | 6 | 5 | 4 | 5 | 57 |

| ID | Lastname | Firstname | Gender | Ethnic | Year | Lowup | Section | GPA | Extrcredit | Review | Quiz1 | Quiz2 | Quiz3 | Quiz4 | Quiz5 | Final |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 289652 | BRADLEY | SHANNON | 1 | 4 | 3 | 2 | 1 | 2.46 | 1 | 2 | 6 | 9 | 8 | 9 | 9 | 68 |
| 466407 | PICKERING | HEIDI | 1 | 3 | 3 | 2 | 3 | 2.38 | 1 | 1 | 4 | 7 | 6 | 4 | 7 | 56 |
| 898766 | RAO | DAWN | 1 | 2 | 3 | 2 | 1 | 3.90 | 1 | 2 | 8 | 10 | 10 | 8 | 9 | 73 |
| 302400 | JONES | ROBERT | 2 | 3 | 4 | 2 | 3 | 1.14 | 1 | 2 | 2 | 5 | 4 | 5 | 6 | 43 |
| 157147 | BAKKEN | KREG | 2 | 4 | 3 | 2 | 1 | 3.95 | 2 | 2 | 10 | 10 | 10 | 10 | 9 | 74 |
| 519444 | RATHBUN | DAWN | 1 | 4 | 4 | 2 | 2 | 3.90 | 1 | 1 | 10 | 9 | 10 | 10 | 8 | 74 |
| 420327 | BADGER | SUSAN | 1 | 4 | 3 | 2 | 3 | 2.61 | 1 | 2 | 10 | 10 | 10 | 10 | 10 | 53 |
| 260983 | CUSTER | JAMES | 2 | 4 | 4 | 2 | 1 | 2.54 | 1 | 1 | 10 | 9 | 10 | 10 | 7 | 60 |
| 554809 | JONES | LISA | 1 | 3 | 3 | 2 | 3 | 3.35 | 1 | 1 | 7 | 8 | 8 | 9 | 6 | 69 |
| 777683 | ANDERSON | ERIC | 2 | 5 | 4 | 2 | 3 | 2.40 | 1 | 1 | 3 | 6 | 3 | 2 | 6 | 50 |
| 553919 | KWON | SHELLY | 1 | 2 | 3 | 2 | 1 | 3.90 | 2 | 2 | 10 | 10 | 10 | 10 | 8 | 75 |
| 479547 | LANGFORD | BLAIR | 2 | 3 | 3 | 2 | 1 | 3.42 | 2 | 2 | 10 | 10 | 10 | 9 | 10 | 75 |
| 755724 | LANGFORD | TREVOR | 2 | 4 | 3 | 2 | 2 | 2.96 | 1 | 2 | 8 | 9 | 9 | 9 | 8 | 62 |
| 337908 | UYEYAMA | VICTORINE | 1 | 1 | 3 | 2 | 2 | 2.34 | 2 | 1 | 10 | 8 | 10 | 10 | 7 | 63 |
| 798931 | ZUILL | RENEE | 1 | 4 | 3 | 2 | 1 | 2.22 | 2 | 2 | 10 | 9 | 10 | 10 | 8 | 62 |
| 843472 | PRADO | DON | 2 | 5 | 3 | 2 | 3 | 3.54 | 1 | 2 | 9 | 9 | 10 | 8 | 9 | 68 |
| 762308 | GOUW | BONNIE | 1 | 4 | 2 | 1 | 3 | 3.90 | 1 | 2 | 8 | 7 | 9 | 10 | 8 | 57 |
| 700978 | WEBSTER | DEANNA | 1 | 3 | 2 | 1 | 3 | 3.90 | 1 | 2 | 8 | 9 | 9 | 10 | 10 | 67 |
| 721311 | SONG | LOIS | 2 | 2 | 3 | 2 | 3 | 1.61 | 1 | 1 | 6 | 9 | 9 | 7 | 10 | 64 |
| 417003 | EVANGELIST | NIKKI | 1 | 2 | 3 | 2 | 2 | 1.91 | 1 | 2 | 9 | 8 | 10 | 10 | 6 | 66 |
| 153964 | TOMOSAWA | DANIEL | 2 | 2 | 3 | 2 | 3 | 2.84 | 2 | 1 | 10 | 9 | 10 | 10 | 10 | 63 |
| 765360 | ROBINSON | ERIC | 2 | 3 | 3 | 2 | 2 | 2.43 | 1 | 2 | 8 | 8 | 7 | 8 | 10 | 65 |
| 463276 | HANSEN | TIM | 2 | 4 | 3 | 2 | 1 | 3.84 | 2 | 2 | 10 | 10 | 10 | 9 | 10 | 74 |
| 737728 | BELTRAN | JIM | 2 | 3 | 3 | 2 | 1 | 2.57 | 1 | 1 | 6 | 8 | 9 | 5 | 7 | 62 |
| 594463 | CRUZADO | MARITESS | 1 | 4 | 4 | 2 | 2 | 3.05 | 1 | 2 | 9 | 8 | 10 | 8 | 8 | 65 |
| 938666 | SUAREZ-TAN | KHANH | 1 | 2 | 3 | 2 | 3 | 2.02 | 2 | 2 | 10 | 8 | 10 | 10 | 7 | 52 |
| 616095 | SPRINGER | ANNELIES | 1 | 4 | 3 | 2 | 1 | 3.64 | 1 | 2 | 10 | 10 | 10 | 10 | 10 | 72 |
| 219593 | POTTER | MICKEY | 1 | 5 | 3 | 2 | 3 | 2.54 | 1 | 2 | 5 | 8 | 6 | 4 | 10 | 61 |
| 473303 | PARK | SANDRA | 1 | 3 | 4 | 2 | 2 | 3.17 | 1 | 2 | 8 | 8 | 8 | 10 | 9 | 70 |
| 287617 | CUMMINGS | DAVENA | 1 | 5 | 3 | 2 | 3 | 2.21 | 1 | 2 | 9 | 10 | 9 | 9 | 9 | 52 |
| 681855 | GRISWOLD | TAMMY | 1 | 4 | 3 | 2 | 2 | 1.50 | 1 | 2 | 5 | 7 | 8 | 5 | 8 | 57 |
| 899529 | HAWKINS | CATHERINE | 1 | 3 | 4 | 2 | 2 | 2.31 | 1 | 1 | 10 | 8 | 9 | 10 | 7 | 49 |
| 576141 | MISHALANY | LUCY | 1 | 4 | 3 | 2 | 1 | 3.57 | 1 | 2 | 0 | 3 | 2 | 2 | 2 | 42 |
| 273611 | WU | VIDYUTH | 1 | 2 | 2 | 1 | 2 | 3.70 | 1 | 2 | 3 | 6 | 2 | 6 | 6 | 55 |
| 900485 | COCHRAN | STACY | 2 | 4 | 3 | 2 | 2 | 2.77 | 2 | 2 | 10 | 9 | 10 | 10 | 9 | 61 |
| 211239 | AUSTIN | DERRICK | 2 | 4 | 3 | 2 | 3 | 2.33 | 1 | 2 | 5 | 5 | 7 | 6 | 4 | 52 |
| 780028 | ROBINSON | CLAYTON | 2 | 4 | 3 | 2 | 1 | 3.90 | 1 | 2 | 10 | 10 | 10 | 9 | 10 | 73 |
| 896972 | HUANG | MIRNA | 1 | 2 | 3 | 2 | 1 | 2.56 | 1 | 1 | 7 | 6 | 10 | 8 | 7 | 57 |
| 280440 | CHANG | RENE | 1 | 2 | 3 | 2 | 2 | 3.90 | 1 | 2 | 10 | 8 | 10 | 10 | 8 | 68 |
| 920656 | LIAO | MICHELLE | 1 | 2 | 2 | 1 | 2 | 3.28 | 2 | 2 | 10 | 9 | 10 | 10 | 9 | 72 |
| 490016 | STEPHEN | LIZA | 1 | 5 | 3 | 2 | 2 | 2.72 | 1 | 2 | 8 | 9 | 9 | 8 | 10 | 60 |
| 140219 | GUADIZ | VALERIE | 1 | 2 | 4 | 2 | 1 | 1.84 | 1 | 1 | 7 | 8 | 9 | 8 | 10 | 66 |
| 972678 | KAHRS | JANNA | 1 | 4 | 4 | 2 | 2 | 2.37 | 1 | 2 | 10 | 10 | 10 | 10 | 10 | 53 |
| 988808 | MCCONAHA | CORA | 1 | 4 | 3 | 2 | 3 | 3.06 | 1 | 2 | 7 | 8 | 9 | 8 | 7 | 68 |
| 132931 | OSBORNE | ANN | 1 | 3 | 2 | 1 | 2 | 3.98 | 1 | 1 | 7 | 8 | 7 | 7 | 6 | 68 |
| 915457 | SHEARER | LUCIO | 2 | 3 | 3 | 2 | 1 | 2.22 | 1 | 2 | 10 | 10 | 10 | 9 | 8 | 52 |
| 762813 | DEAL | IVAN | 2 | 3 | 2 | 1 | 1 | 2.27 | 2 | 2 | 10 | 9 | 10 | 10 | 10 | 62 |

*Continued*

| ID | Lastname | Firstname | Gender | Ethnic | Year | Lowup | Section | GPA | Extrcredit | Review | Quiz1 | Quiz2 | Quiz3 | Quiz4 | Quiz5 | Final |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 897606 | GENOBAGA | JACQUELIN | 1 | 2 | 3 | 2 | 3 | 2.92 | 1 | 2 | 8 | 9 | 8 | 8 | 7 | 68 |
| 164842 | VALENZUELA | NANCY | 1 | 1 | 4 | 2 | 2 | 2.32 | 1 | 1 | 7 | 8 | 6 | 7 | 10 | 59 |
| 822485 | VALENZUELA | KATHRYN | 1 | 4 | 1 | 1 | 1 | 3.90 | 1 | 2 | 8 | 9 | 10 | 10 | 8 | 66 |
| 978889 | ZIMCHEK | ARMANDO | 2 | 4 | 4 | 2 | 1 | 3.90 | 1 | 2 | 4 | 8 | 6 | 6 | 9 | 64 |
| 779481 | AHGHEL | BRENDA | 1 | 5 | 3 | 2 | 1 | 3.01 | 1 | 2 | 3 | 5 | 3 | 2 | 4 | 49 |
| 921297 | KINZER | RICHARD | 2 | 4 | 3 | 2 | 2 | 2.73 | 1 | 2 | 7 | 9 | 9 | 7 | 8 | 67 |
| 983522 | SLOAT | AARON | 2 | 3 | 3 | 2 | 3 | 2.11 | 1 | 1 | 4 | 5 | 6 | 6 | 6 | 50 |
| 807963 | LEWIS | CARL | 2 | 3 | 2 | 1 | 1 | 2.56 | 2 | 1 | 8 | 5 | 6 | 4 | 7 | 62 |
| 756097 | KURSEE | JACKIE | 1 | 3 | 3 | 2 | 2 | 3.13 | 1 | 2 | 9 | 6 | 8 | 7 | 10 | 66 |
| 576008 | BULMERKA | HUSIBA | 1 | 4 | 4 | 2 | 3 | 3.45 | 2 | 1 | 10 | 8 | 7 | 9 | 7 | 68 |
| 467806 | DEVERS | GAIL | 1 | 3 | 3 | 2 | 1 | 2.34 | 1 | 1 | 7 | 6 | 8 | 7 | 9 | 59 |
| 307894 | TORRENCE | GWEN | 1 | 3 | 2 | 1 | 2 | 2.09 | 2 | 2 | 6 | 5 | 4 | 7 | 6 | 62 |

# Exercises

Answers to select exercises are downloadable at **www.spss-step-by-step.net**.

1. Set up the variables described above for the **grades.sav** file, using appropriate variable names, variable labels, and variable values. Enter the data for the first 20 students into the data file.

2. Perhaps the instructor of the classes in the **grades.sav** dataset teaches these classes at two different schools. Create a new variable in this dataset named **school**, with values of 1 and 2. Create variable labels, where 1 is the name of a school you like, and 2 is the name of a school you don't like. Save your dataset with the name **gradesme.sav**.

3. Which of the following variable names will SPSS accept, and which will SPSS reject? For those that SPSS will reject, how could you change the variable name to make it "legal"?

   - age
   - firstname
   - @edu
   - sex.
   - grade
   - not
   - anxeceu
   - date
   - iq

4. Using the **grades.sav** file, make the **gpa** variable values (which currently have two digits after the decimal point) have no digits after the decimal point. You should be able to do this without retyping any numbers. *Note that this won't actually round the numbers, but it will change the way they are displayed and how many digits are displayed after the decimal point for statistical analyses you perform on the numbers.*

5. Using **grades.sav**, search for a student with 121 total points. What is his or her name?

6. Why is each of the following variables defined with the measure listed? Is it possible for any of these variables to be defined as a different type of measure?

| | |
|---|---|
| ethnicity | Nominal |
| extrcred | Ordinal |
| quiz4 | Scale |
| grade | Nominal |

7. Ten people were given a test of balance while standing on level ground, and ten other people were given a test of balance while standing on a 30° slope. Their scores follow. Set up the appropriate variables, and enter the data into SPSS.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Scores of people standing on level ground: | 56 | 50 | 41 | 65 | 47 | 50 | 64 | 48 | 47 | 57 |
| Scores of people standing on a slope: | 30 | 50 | 51 | 26 | 37 | 32 | 37 | 29 | 52 | 54 |

8. Ten people were given two tests of balance, first while standing on level ground and then while standing on a 30° slope. Their scores follow. Set up the appropriate variables, and enter the data into SPSS.

| Participant: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Score standing on level ground: | 56 | 50 | 41 | 65 | 47 | 50 | 64 | 48 | 47 | 57 |
| Score standing on a slope: | 38 | 50 | 46 | 46 | 42 | 41 | 49 | 38 | 49 | 55 |

# Chapter 4
# Managing Data

THE MATERIAL covered in this chapter, when mastered, will provide a firm grounding in data management that will serve you well throughout the remaining chapters of this book. Rarely will you complete an analysis session without making use of some of the features in this chapter. Occasionally you may have a single data file and want to run a single operation. If so, reference to this chapter may be unnecessary; just use the data as you originally entered and formatted them. Chapter 3 explains how to create and format a data file, and Chapters 6–27 provide lucid step-by-step instructions about how to analyze data. However, it is common to want to analyze variables that are not *in* the original data file but are derived *from* the original variables (such as total points, percentage of possible points, or others). It is also sometimes desirable to display the data differently (e.g., alphabetically by student last names, by total points from high to low, or class records listed by sections); to code variables differently (e.g., 90–100 coded A, 80–89 coded B, and so forth); to perform analyses on only a certain portion of the data set (e.g., scores for females, GPAs for Whites, final percentages for Sophomores, final grades for Section 2, etc.); or if the sections were in three different files, to merge the files to gain a clearer picture of characteristics of the entire class.

All these operations may at times be necessary if you wish to accomplish more than a single analysis on a single set of data. Some operations presented here are complex, and it is possible that you may encounter difficulties that are not covered in this book. The different functions covered in this chapter are considered one at a time, and we present them as clearly and succinctly as space allows. You can also find additional help by selecting **Help** → **Topics** within SPSS; this will allow you to browse the documentation, or use the index to look up a particular key term that you are having problems with.

Despite the potential difficulty involved, a thorough understanding of these procedures will provide the foundation for fluency in the use of other SPSS operations. Once you have learned to manage data effectively, many statistical processes can be accomplished with almost ridiculous ease. We will present here seven different types of data management tools:

1. The judicious use of the **Case Summaries** option to assist in proofing and editing your data.
2. Using the **Replace Missing Values** subcommand for several options that deal with replacing missing data.
3. Computing and creating new variables from an already existing data set by using the **Transform** command and the **Compute Variable** subcommand.
4. Using the **Recode** subcommand to change the coding associated with certain variables.
5. Employing the **Select Cases** procedure to allow statistical operations on only a portion of your data.
6. Utilizing the **Sort Cases** subcommand to reorder your data.
7. Adding new cases or new variables to an already existing file using the **Merge Files** option.

Beginning with this chapter we introduce the convention that allows all chapters that follow to be self-contained. The three steps necessary to get as far as reading a data file and accessing a screen that will allow you to conduct analyses will begin the Step by Step section in each chapter. This will be followed by the steps necessary to run a particular analysis. A description of how to print the results and exit the program will conclude each section. The critical steps designed to perform a particular procedure for each chapter will typically involve a **Step 4** that accesses and sets up the procedure, and one or more versions of **Step 5** that describes the specifics of how to complete the procedure(s). If there is more than one Step 5 version, the steps are designated **5**, **5a**, **5b**, **5c**, and so forth.

This chapter will be formatted somewhat differently from the analysis chapters (6–27) because we are presenting seven different types of computations or manipulations of data. We will begin the Step by Step section with the three introductory steps mentioned in the previous page and then present each of the seven operations with its own title, introduction, and discussion. After the seventh of these presentations, we will conclude with the final steps. Unlike the analysis chapters, there are no Output sections in this chapter.

## 4.1 Step By Step: Manipulation of Data

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | ✳ ⊞ → ✳ All Programs ▷ → ✳ IBM SPSS Statistics → ✳ IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | ✳ 🚀 → ✳ ■ → ✳ Σ | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| | Step 2 |
|---|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* | |

*Screens 1 and 2 (displayed on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **grades.sav** *file for further analyses:*

| In Screen | Do This | Step 3 |
|---|---|---|
| Front1 | ✳ File → Open → ✳ Data [*or* ✳ 📂] | |
| Front2 | type `grades.sav` → ✳ Open [*or* ✳ ✳ grades.sav] | Data |

*Whether first entering SPSS or returning from earlier operations the standard menu of commands across the top is required. As long as it is visible you may perform any analyses. It is not necessary for the data window to be visible.*

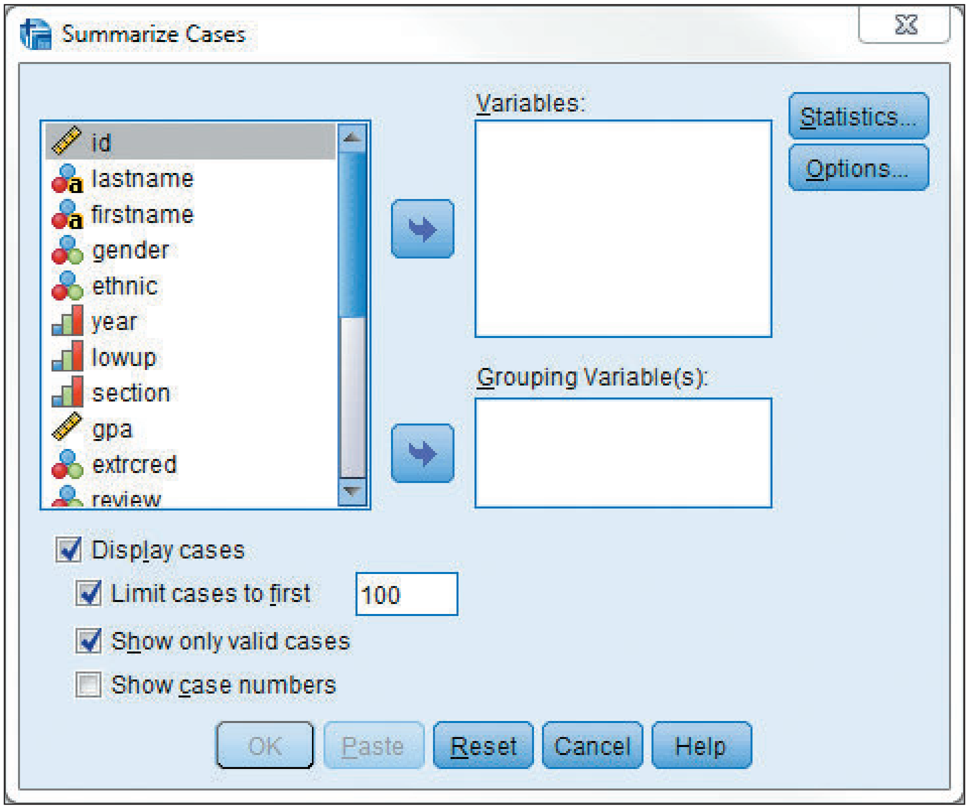
## 4.2 The Case Summaries Procedure

**LISTING ALL OR A PORTION OF DATA** After completion of Step 3, a screen with the desired menu bar appears. When you click a command (from the menu bar) a series of options emerges (usually) below the selected command. With each new set of options, click the desired item. To access the **Case Summaries** procedure, begin at any screen that shows the standard menu of commands.

| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | **Analyze** → **Reports** → **Case Summaries** | 4.1 |

One of the best ways to ensure that you have entered data correctly or that SPSS has followed your instructions and created, formatted, or arranged your data is to make generous use of the **Case Summaries** operation. The **Case Summaries** command allows you to list an entire data file or a subset of that file, either grouped or in the order of the original data. After executing the steps listed above, a new window opens (Screen 4.1) that allows you to specify which variables and/or cases you wish to have listed.

**Screen 4.1** The Summarize Cases Window



☑ Display cases

☑ Limit cases to first 100

☑ Show only valid cases

☐ Show case numbers

The variable list on the left allows you to select which variables you wish. You may, of course, select all of them. In selecting the variables you have the option of choosing the order in which they appear. The order of selection in the **Variables** box will be the same order presented in the output. Several options allow you to select and format both the content and structure of the output.

- **Variables**: The list of variables selected for listing all or a subset of values.
- **Grouping Variable(s)**: Will create an order for the listing of variables. If there are no grouping variables, cases will be listed in the order of the data file. To group variables you must first have one or more variables in the **Variables:** box. If, for instance, you included **gpa** in the **Variables:** box with **gender** in the **Grouping Variable(s):** box, then the output would list the GPAs for all women (in the order of the data file) and then GPAs for all men, also in order. If you selected **section**, the output would list GPAs for all students in section 1, then for students in section 2, and then for students in section 3. You may include more than one

grouping variable. If you select **section** and then **gender**, the output would first list GPAs for women in section 1, then GPAs for men in section 1, then GPAs for women in section 2, and so forth.

- **Display cases**: This is selected by default (and is the major function of the Case Summaries procedure). If deselected, then output will list only the number of data points in each category.
- **Limit Cases to first**: The first 100 cases is the default value. If you wish more or less, delete the **100** and type in the number you wish.
- **Show only valid cases**: This option is selected by default. It ensures that when variables are listed—if there are three subjects who did not include, for instance, their age—those three subjects would not be listed in the output. This option will often be deselected because we are often quite interested in where missing values occur.
- **Show case numbers**: The **Case Summaries** operation automatically provides numbers in sequence for each case/subject. Clicking **Show case numbers** simply provides a duplicate set of identical numbers—entirely superfluous. However, this feature can be very useful if you have a grouping variable so that your cases are not in sequence. Then the case number will assist you in finding a particular case.
- **Statistics**: A wide array of descriptive statistics is available. Most of these are described in Chapter 7. The **number of cases** is included by default. Any other statistics you wish may be clicked into the **Cell Statistics** box.
- **Options:** Allows you to exclude any subjects who have missing values in their data or to include or edit titles or subtitles. Since you may edit titles in the output anyway, this option is generally ignored.

*The steps below request a listing of all variables for the first 20 cases. We begin this procedure where sequence Step 4 ended, Screen 4.1.*

| In Screen | Do This | Step 5 |
|---|---|---|
| 4.1 | **id** *(or the first variable)* → **press** *and hold down the* **Shift** *key* → | |
| | **passfail** *(or the last variable). This highlights all variables from* **id** *to* **passfail** → | |
| | *top* *button* → **press** **TAB** *4 times [or highlight the* **100**] → | |
| | **type** 20 → **OK** | **Back1** |

The **Edit** → **Options** sequence allows you to select an alphabetic ordering or a variable list ordering. This sequence (shown above) will produce a listing of the first 20 subjects with all variables included in the order displayed in the original variable list.

*To select the variables* **id**, **total**, *and* **grade**, *grouped by* **section** *and* **gender**, *with cases numbered for all subjects, perform the following sequence of steps:*

| In Screen | Do This | Step 5a |
|---|---|---|
| 4.1 | *While holding down the* **Ctrl** *key* **id** → **total** → **grade** → | |
| | *top* *button* → **section** → *lower* *button* → **gender** → | |
| | *lower* *button* → **press** **TAB** *3 times [or highlight the* **100**] → **type** 105 → | |
| | **Show case numbers** → **OK** | **Back1** |

This operation will produce a listing of student **id**s, **total** points, and **grades**; first for women in section 1, then men in section 1, then women in section 2, and so forth; each case will be numbered. The **Show case numbers** function is useful this time because the subjects (students) are out of sequence (as described on the previous page).

## 4.2.1 The Replacing Missing Values Procedure

During the course of data entry you will discover soon enough that several subjects refused to answer the ethnicity question, three subjects did not take the first quiz, and that other gaps in your data file appear. These are called *missing values*. Missing values are not only an irritant but can influence your analyses in a number of undesirable ways. Missing values often make your data file more difficult to work with. There are a number of SPSS procedures that remove any case (or subject) with missing values for any analyses that follow. If, for instance, you are computing correlations, and 13 of your 35 subjects have one or more missing values in the data file, SPSS computes correlations (depending on your settings) for only the 22 subjects that have *no* missing values. This results in a distressing loss of legitimate information that should be available.

In many procedures SPSS allows you to delete subjects **listwise** (if one or more missing value(s) exist for a subject, *all* data for that subject are removed before any analyses), or to delete subjects **pairwise** (if for any calculation a necessary data point is missing for a subject, the *calculation* will be conducted without the influence of that subject). In other words, for pairwise deletion, a subject with missing values will be involved in all analyses except those for which critical values are missing. The authors suggest that you, to the greatest extent possible, deal with missing values during the data-coding and data-entry phase rather than relying on SPSS-missing-value defaults. For categorical data, this is easy, but for some continuous data, the SPSS procedures may be appropriate.

1. If you have missing values in *categorical data* (e.g., subjects didn't answer the ethnicity or level-of-income questions), you can create an additional level for that variable and replace the missing values with a different number. For instance, if you have five levels of ethnicity coded for five different ethnic groups, you could create a sixth level (coded 6) labeled "unknown." This would not interfere with any analyses you wished to conduct with the ethnicity variable, and there would no longer be missing values to contend with.
2. For *continuous* data a frequent procedure is to replace missing values with the mean score of all other subjects for that variable. SPSS has a procedure that allows this type of replacement to occur. Within this context, SPSS also has several other options (e.g., replace with the median, replace with the mean of surrounding points) that are presented and described on page 65. Although replacing *many* missing values by these techniques can sometimes bias the results, a small number of replacements has little influence on the outcome of your analyses. An often-used rule of thumb suggests that it is acceptable to replace up to 15% of data by the mean of values for that variable (or equivalent procedures) with little damage to the resulting outcomes. If a particular subject (or case) or a certain variable is missing more than 15% of the data, it is recommended that you drop that subject or variable from the analysis entirely.
3. A more sophisticated way to replace missing values in continuous data is to create a regression equation with the variable of interest as the dependent (or criterion) variable, and then replace missing values with the *predicted* values. This requires a fair amount of work (SPSS has no automatic procedures to accomplish this) and a substantial knowledge of multiple regression analysis (Chapter 16); but it is considered one of the better methods for dealing with missing data. If you make use of predicted values, you will not be able to use the procedure described in the following page. It is necessary to compute the new values separately and insert them into the original data file.

Concerning research ethics: If you replace missing data in research that you plan to publish, it is required that you state in the article how you handled any missing values that may influence your outcome. The reviewers and the readers then know what procedures you have used and can interpret accordingly.

If you spend time in the SPSS manuals, you will often see the phrases *system-missing values* and *user-missing values*. Just to clarify, system-missing values are simply omissions in your data set and are represented by a period in the cell where the missing value occurs. A user-missing value is a value that is specified *by the researcher* as a missing value for a particular level (or levels) of a variable. For instance, in an educational setting subjects might be coded 1, 2, or 3 based on which of three different forms of a test they took. Some subjects could be coded 8 for did-not-complete-test or 9 for did-not-attend. The researcher could specify codings 8 and 9 as user-missing values.

*Begin with any screen that shows the menu of commands across the top. To access the missing values procedure perform the following sequence of steps:*

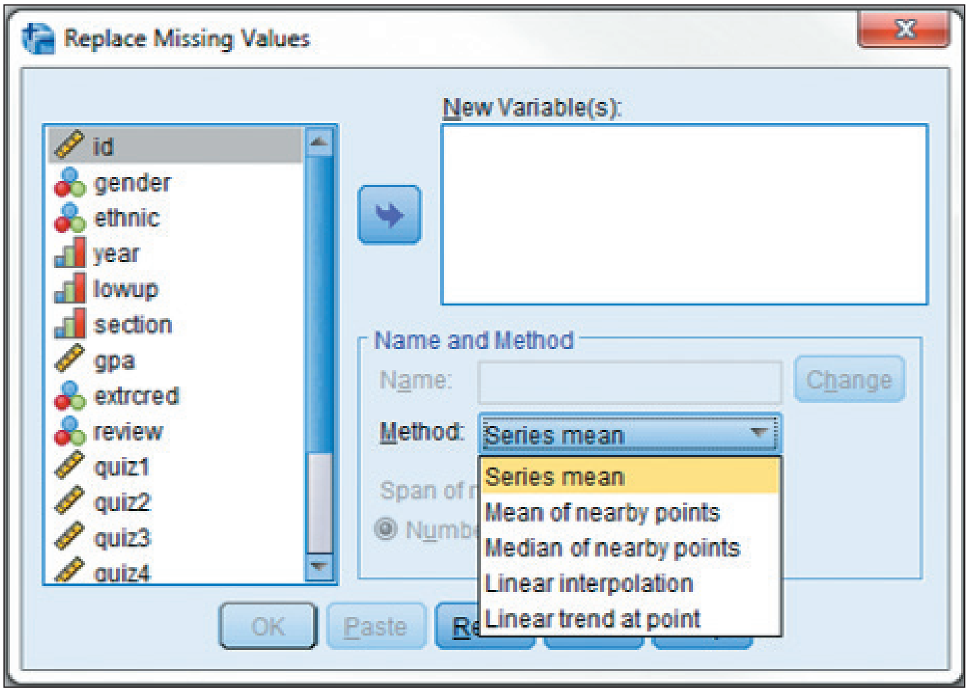
| In Screen | Do This | Step 4b |
|---|---|---|
| Menu | **Transform** → **Replace Missing Values** | 4.2 |

At this point a dialog box will open (Screen 4.2, below) that provides several options for dealing with missing values. The list of available variables appears to the left, and you will paste into the **New Variable(s)** box variables designated for missing-values replacement. You may designate more than one variable, and, by using the **Change** button, can select different methods for replacing missing values for each variable. We illustrate this procedure using the **grades.sav** file—even though this file contains no missing values. Under **Method** in the dialog box shown below, SPSS provides five different techniques for replacing missing values:

- **Series mean**: Missing values are replaced by the mean (or average) value of all other values for that variable. This is the default procedure.

Screen 4.2 The Replace Missing Values Window

*(Note that the drop-down menu to the right of* **Method** *is activated)*



- **Mean of nearby points**: Missing values are replaced by the mean of surrounding values (that is, values whose SPSS case numbers are close to the case with a missing value). You may designate how many values to use under **Span of nearby points**.
- **Median of nearby points**: Missing values are replaced by the median (see Glossary) of surrounding points. You may designate how many surrounding values to use.
- **Linear interpolation**: Missing values are replaced by the value midway between the surrounding two values.
- **Linear trend at point**: If values of the variable tend to increase or decrease from the first to the last case, then missing values will be replaced by a value consistent with that trend.

If you select one of the two options that is based on a span of points, you will designate the number of values by clicking **Number** and typing in the number of surrounding values you wish. Of the five techniques, **Series mean** is by far the most frequently used method. The other four are usually applied in time series analyses when there is some intrinsic trend of values for the cases over time. For instance, if your data were constructed so that the cases represented 10 different times rather than 10 different subjects, then you might anticipate changes in your variables over time. This does not frequently occur with subjects unless they are ordered based on some intrinsic quality that relates to the independent variables in the study, such as IQ from low to high (on a study of academic achievement) or body weight from low to high (for a study dealing with weight loss).

In the example that follows we will replace missing values for previous **GPA** and **quiz1** with the mean value.

*Beginning with Screen 4.2, perform the following sequence of steps.*

| In Screen | Do This | Step 5b |
|---|---|---|
| 4.2 | ✳ gpa → ✳ ➡ → ✳ quiz1 → ✳ ➡ → ✳ OK | Back1 |

Some comments concerning this very simple sequence of steps are required: (1) After clicking **OK**, the output will list the number of values replaced, the value that replaced the missing values, and the first and last case number of cases considered. (2) The sequence of steps is simple because the default option (**Series mean**) is the desired option, and usually will be. To select a different option: After you paste a variable into the **New Variable(s)** box, click the ▾ to the right of the **Method** label, click whichever option you desire from the menu, and then click the **Change** button. You may then click the **OK** button to run the program. (3) Despite the fact that SPSS can perform the operations illustrated above, there is little logical reason for doing so. If you wanted to replace a missing GPA you would probably do so based on other indicators of that person's academic ability (such as performance on tests and quizzes) rather than a class average. (4) When you replace missing values by standard SPSS procedures for a particular variable (or variables), SPSS creates a new variable with the old variable name followed by an underline (_) and a "1." If you wish to replace a missing quiz score, an average of the student's other quizzes makes more sense than a class average for that quiz. You may then keep both sets of variables (the original with missing values and the new variable without missing values) or you may cut and paste the data without missing values under the name of the original variable.

## 4.3 The Compute Procedure: Creating Variables

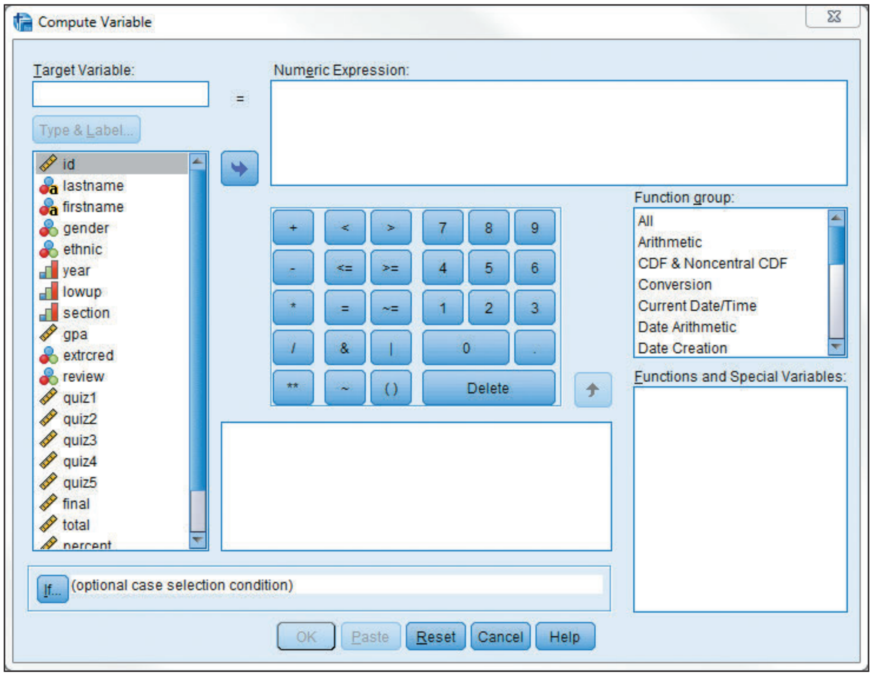
In the current data set, it would be quite normal for the teacher to instruct the computer to create new variables that calculate total points earned as the sum of the quizzes and final exam, and to determine the percent of total points for each student. The sequence of steps that follows will compute two new variables called **total** and **percent**. As noted earlier, the **grades.sav** file (available for download at *www.spss-step-by-step.net*) already contains the four new variables computed in this chapter: **total**, **percent**, **grade**, and **passfail**.

*Beginning with a screen that shows the menu of commands, perform the following step to access the* **Compute variable** *window.*

| In Screen | Do This | Step 4c |
|---|---|---|
| Menu | **Transform** → **Compute Variable...** | 4.3 |

At this point, a new window opens (Screen 4.3, below) that allows you to compute new variables. The list of variables is listed in a box to the left of the screen. Above these is the **Target Variable** box. In this box you will type the name of the new variable you wish to compute. To the right is the **Numeric Expression** box. In this box you will type or paste (usually a combination of both) the expression that will define the new variable. Three options are then provided to assist you in creating the expression defining the new variable.

**Screen 4.3** The Compute Variable Window



- *The calculator pad*: On this pad are all single-digit numbers and a decimal point. Each of these may be entered into the **Numeric Expression** by a mouse click on the screen button or by typing the same on the keyboard. In addition there are a number of *operation* buttons. If a similar button occurs on the keyboard (e.g., <, >, +, etc.) you may type that symbol rather than click the screen button. The operation keys and their meanings follow:

| Arithmetic Operations | | Rational Operations | | Logical Operations | |
|---|---|---|---|---|---|
| + | add | < | less than | & | and: both relations must be true |
| – | subtract | > | greater than | | |
| * | multiply | <= | less than or equal | \| | or: either relation may be true |
| / | divide | >= | greater than or equal | | |
| ** | raise to power | = | equal | ~ | negation: true becomes false, false becomes true |
| ( ) | order of operations | ~= | not equal | | |

- *The* **Functions** *boxes*: A terrifying array of over 180 different functions emerge. We present only nine of these functions; the nine we feel are most likely to be used. To help create order, SPSS has created three boxes in the screen (Screen 4.3) to assist:
  1. The box labeled **Function group** provides different categories of functions. For instance, in Screen 4.3 the Arithmetic function may be selected. This provides (in the window below) just 13 functions (rather than 188)—much easier to negotiate. The first seven functions in the chart on this page (below) are under **Arithmetic**, the final two are under **Random numbers**.
  2. The box labeled **Functions and Special Variables** lists the available options in each category. If you select **All** (in the box above), then all 188 functions will be listed alphabetically.
  3. The box to the left of the **Functions and Special Variables** box provides the term used in the chart on this page (below) and identifies what the function does. In this case **Abs** is highlighted (box to the right) and the definition of **ABS(numexpr)** is provided.

*An example of a sample* **Target Variable**, **Numeric Expression**, *and how it would compute in your file follows. Please refer back to Screen 4.3 for visual reference.*

| Expression | Illustration |
|---|---|
| **ABS(numexpr)**<br>*(absolute value)* | Target variable: **zpositiv** → Numeric expression: **ABS(zscore)**. Creates a variable named **zpositiv** that calculates the absolute value of a variable named **zscore** for each subject. |
| **RND(numexpr)**<br>*(round to the nearest integer)* | Target variable: **simple** → Numeric expression: **RND(gpa).** Computes a variable named **simple** by rounding off each subject's **gpa** to the nearest integer. Note: use "**Rnd(1)**" and (next entry) "**Trunc(1).**" |
| **TRUNC(numexpr)**<br>*(truncates decimal portion of a number)* | Target variable: **easy** → Numeric expression: **TRUNC(gpa)**. Computes a variable named **easy** that truncates each subject's **gpa** (truncate means "cut off"). This is like rounding, but it always rounds down. |
| **SQRT(numexpr)**<br>*(square root)* | Target variable: **scorert** → Numeric expression: **SQRT(score)**. Creates a new variable, **scorert**, by square rooting each **score** for each subject. |
| **EXP(numexpr)**<br>*(exponential (e) raised to a power)* | Target variable: **confuse** → Numeric expression: **EXP(gpa).** Computes a variable named **confuse** that calculates the value of e raised to each subject's **gpa's** power. e ≈ 2.721 (it is irrational). For a particular subject with a 3.49 **gpa**, **EXP**(3.49) = $(2.721\ldots)^{3.49} \approx 32.900506\ldots$ |
| **LG10(numexpr)**<br>*(base 10 logarithm)* | Target variable: **rhythm** → Numeric expression: **LOG10(total)**. Creates a new variable named **rhythm** that calculates the base 10 logarithm for the variable **total** for each subject. |
| **LN(numexpr)**<br>*(natural logarithm)* | Target variable: **natural** → Numeric expression: **LN(total)**. Creates a new variable named **natural** that calculates the natural logarithm for the variable **total** for each subject. |
| **RV.NORMAL(mean, stddev)** | Computes random numbers based on a normal distribution with a user-specified mean (**mean**) and standard deviation (**stddev**). Target variable: **randnorm** → Numeric expression: **RV. NORMAL(5,3).** This function will generate random numbers based on a normal distribution of values with a mean of 5 and a standard deviation of 3. One of these numbers will be assigned randomly to each subject (or case) in your data file under the variable name **randnorm**. |
| **RV.UNIFORM(min, max)** | Computes random numbers based on equal probability of selection for a range of numbers between a user-specified minimum (**min**) and maximum (**max**). Target variable: **random** → Numeric expression: **UNIFORM(1,100).** Random numbers from 1 to 100 will be assigned to each subject (or case) in your data file. |

- *The* **If** *pushbutton*: A click on this button opens a new screen so similar to Screen 4.3 that we will not reproduce it here. The only differences are an **Include all cases** option paired with an **Include if case satisfies condition**, and a **Continue** button at the bottom of the window.

These resources will be demonstrated in two examples that follow. The first example computes two new variables, **total** and **percent**. **Total** will be the sum of the five quizzes and the final exam. **Percent** will be each subject's **total** score, divided by 125 (possible points), multiplied by 100 (to make it a percent rather than a decimal), and rounded to the nearest integer. You will need to perform two separate sequences to create these two variables.

*The starting point for this operation is Screen 4.3. To compute the new variable* **total** *perform the following sequence of steps:*

| In Screen | Do This | Step 5c |
|---|---|---|
| 4.3 | type `total` → ✱ **quiz1** → ✱ ➡ [*or* ✱ ✱ **quiz1**] → press + → | |
| | ✱ ✱ **quiz2** → press + → ✱ ✱ **quiz3** → press + → ✱ ✱ **quiz4** → press | |
| | + → ✱ ✱ **quiz5** → press + → ✱ ✱ **final** → ✱ **OK** | Data |

With the click of the **OK**, the new variable is computed and entered in the data file in the last column position. If you wish to move this new variable to a more convenient location, you may cut and paste it to another area in the data file. The **total** variable is already in the downloaded file, so if you wish to practice this procedure use a different variable name, such as **total2**.

*The starting point for the next operation is also Screen 4.3. To compute the new variable* **percent** *and save in the data file the two variables we have created, perform the following sequence of steps: Remember: this procedure rounds the percent to the nearest integer.*

| In Screen | Do This | Step 5c' |
|---|---|---|
| 4.3 | type `percent` → ✱ **Arithmetic** *in the* **Function group** *box* → *scroll down in the* **Functions and** | |
| | **Special Variables** *box and* ✱ **Rnd(1)** → ✱ ⬆ → type `100*total/125` → ✱ **OK** | |
| Data | ✱ **File** → ✱ **Save** | |

The final step saves the newly created data. If you wish to create a variable only for the present session, then don't save the changes and when you next open the data file it will NOT include the new variable.

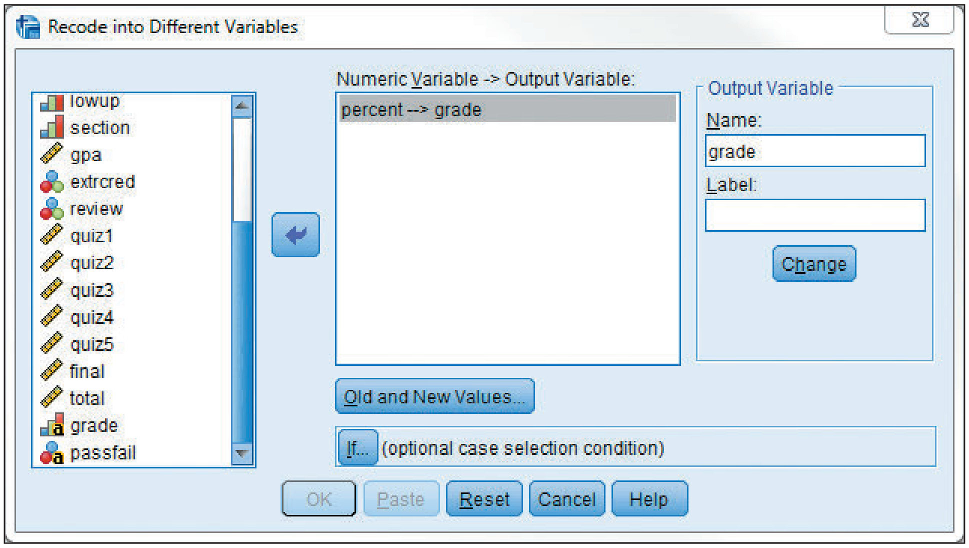
Upon completion of a compute operation, SPSS returns the screen to the data file and you may use the scroll bar to check the accuracy of the new variable(s). You may, of course, use several of these program functions in a single command. The example in Step 5c demonstrates a computation that includes rounding, multiplying by 100, and dividing by 125. In creating more complex computations, be sure to adhere strictly to the basic algebraic rules of order of operations. If you wish to do an operation on a complex expression, be sure to include it within parentheses.

## 4.4 The Recode into Different Variables Procedure Creating New Variables

This procedure also creates new variables, not by means of calculations with already existing variables but rather by dividing a preexisting variable into categories and coding each category differently. This procedure is ideally suited for a data file that includes class records because at the end of the term the teacher wants to divide scores into different grades. In the present example this coding is based on the class **percent** variable in which percents between 90 and 100 = A, between 80 and 89 = B, between 70 and 79 = C, between 60 and 69 = D, and between 0 and 59 = F. The **percent** variable can also be used to classify subjects into pass/fail categories with codings of greater than or equal to 60 = P and less than 60 = F.

Two different windows control the recoding-into-different-variables process. Screen 4.4 (below) allows you to identify the **Numeric Variable** (**percent** in the present example) and the **Output Variable** (**grade** in this case). A click on **Old and New Values** opens up the second window, Screen 4.5 (following page). This box allows you to identify the ranges of the old variable (90 to 100, 80 to 89, etc.) that code into levels of the new variable (A, B, C, D, and F).

**Screen 4.4** The Recode into Different Variables Window



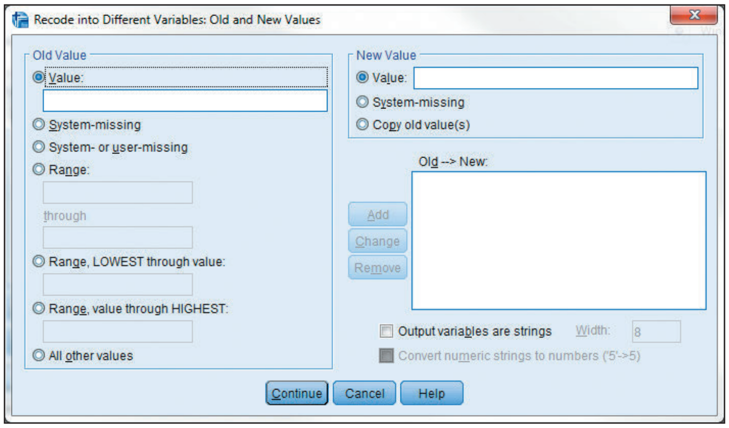
Numeric Variable -> Output Variable:
percent --> grade

Output Variable
Name: grade
Label:

Notice that in Screen 4.4 all numeric variables are listed to the left and the five function buttons are at the bottom of the dialog box. The initial stage (in the present example) is to click on **percent** and paste it into the **Numeric Variable** → **Output Variable** box. Then click in the box beneath the **Output Variable** title and type the name of the new variable, **grade**. A click on the **Change** button will move the **grade** variable back to the previous box. The entry in the active box will now read **percent** → **grade**. A click on the **Old and New Values** pushbutton opens the next window.

The dialog box represented by Screen 4.5 allows you to identify the values or range of values in the input variable (**percent**) that is used to code for levels of the Output variable (**grade**). Since we want a *range* of values to represent a certain grade, click on the top **Range** option and indicate the range of values (90–100) associated with the first level (A) of the new variable (**grade**). When the first range is entered, click in the **Value** box and type the letter A. A click on the **Add** button will paste that range and grade into the **Old** → **New** box. We are now ready to perform a similar sequence for each of the other four grades. When all five levels of the new variable are entered (with the associated ranges and letter grades), click **Continue** and Screen 4.4 will reappear. A click on the **OK** button will create the new variable.

### Screen 4.5 The Old and New Values Window


Recode into Different Variables: Old and New Values
Old Value
◉ Value:
○ System-missing
○ System- or user-missing
○ Range:
through
○ Range, LOWEST through value:
○ Range, value through HIGHEST:
○ All other values
New Value
◉ Value:
○ System-missing
○ Copy old value(s)
Old --> New:
Add
Change
Remove
□ Output variables are strings
Width: 8
□ Convert numeric strings to numbers ('5'->5)
Continue
Cancel
Help

A similar process will be conducted for creating the **passfail** variable. When these two new variables are created they will be listed in the last two columns of the data file. Once the new variable is created, you may type in value or variable labels, change the format, or move the variables to a new location in the data file by techniques described in Chapter 3.

The creation of the two new variables is described below in the Step by Step boxes. Remember that these variables are already in the data available by download. If you want to practice with that file, create variables with slightly different names (e.g., **grade2**) and then you can compare your results with those already present.

*To create the new variable* **grade** *the starting point is any menu screen. Perform the following sequence of steps to create this new variable.*

| In Screen | Do This | Step 5d |
|---|---|---|
| Menu | ✳ Transform → ✳ Recode into Different Variables | |
| 4.4 | ✳ percent → ✳ ➡ → press TAB → type `grade` → ✳ Change → | |
| | ✳ Old and New Values | |
| 4.5 | ✳ Output Variables are strings → *change 8 to 4 in the* **Width** *box* → | |
| | ✳ *circle to the left of* **Range** *(the top one, not the lower* **Range** *or* **Range***)* | |
| | *(Note: before each of the 5 lines that follow, first click in the top* **Range** *box)* | |
| | type `90` → press TAB → type `100` → ✳ *in* **Value** *box* → type `A` → ✳ Add | |
| | type `80` → press TAB → type `89.9` → ✳ *in* **Value** *box* → type `B` → ✳ Add | |
| | type `70` → press TAB → type `79.9` → ✳ *in* **Value** *box* → type `C` → ✳ Add | |
| | type `60` → press TAB → type `69.9` → ✳ *in* **Value** *box* → type `D` → ✳ Add | |
| | type `00` → press TAB → type `59.9` → ✳ *in* **Value** *box* → type `F` → ✳ Add | |
| | → ✳ Continue | |
| 4.4 | ✳ OK | Data |

*To create the new variable* **passfail** *the starting point is any menu screen. Perform the following sequence of steps to create this new variable.*

| In Screen | Do This | Step 5d' |
|---|---|---|
| Menu | ✳ **Transform** → ✳ **Recode into Different Variables** | |
| 4.4 | ✳ **percent** → ✳ ➡ → **press** **TAB** → **type** `passfail` → ✳ **Change** → | |
| | ✳ **Old and New Values** | |
| 4.5 | ✳ **Output Variables are strings** → *change 8 to 4 in the* **Width** *box* → | |
| | ✳ *circle to the left of* **Range** *(the top one, not the lower* **Range** *or* **Range***)* | |
| | *(Note: before each of the 2 lines that follow, first click in the top* **Range** *box)* | |
| | **type** `60` → **press** **TAB** → **type** `100` → ✳ *in* **Value** *box* → **type** `P` → ✳ **Add** | |
| | **type** `00` → **press** **TAB** → **type** `59.9` → ✳ *in* **Value** *box* → **type** `F` → ✳ **Add** | |
| | → ✳ **Continue** | |
| 4.4 | ✳ **OK** | Data |

## 4.4.1 The Recode Option Changing the Coding of Variables

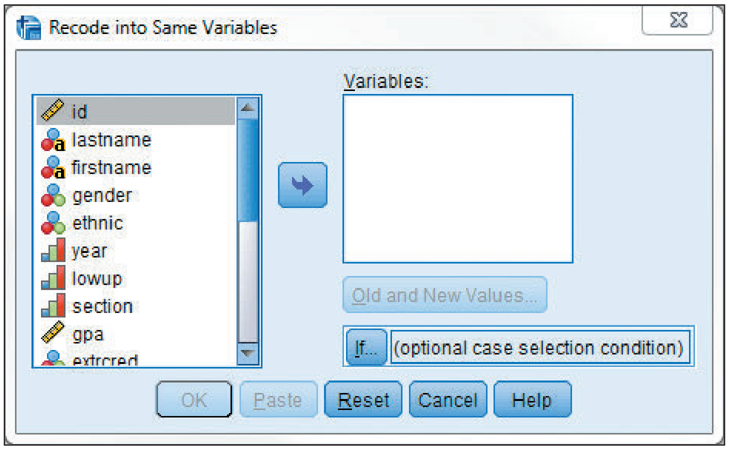
There are times when you may wish to change the coding of your variables. There are usually two reasons why it is desirable to do so. The first is when you wish to reverse coding to be consistent with other data files, or when one coding pattern makes more sense than another. For instance if you had a file in which the gender variable was coded male = 1 and female = 2, you may have other files where it is coded female = 1, male = 2. It would be desirable to recode the files so that they are consistent with each other. Or perhaps you have an **income** variable that has >$100,000 coded 1 and <$10,000 coded 7. If you felt that it would make more sense to have the lower incomes coded with the lower numbers, the **Recode** command could accomplish that easily for you.

The second reason for recoding is to group variables that are ungrouped in the original data. For instance, you might have a variable named **marital** that is coded 1 = never married, 2 = divorced, 3 = widowed, 4 = separated, 5 = married. You could use the **Recode** procedure to change the marital variable so that 1 = single (a combination of 1, 2, and 3) and 2 = married (a combination of 4 and 5).

*The starting point for recoding variables is a screen that shows the menu of commands across the top. Select the following options to access the Recode screen.*

| In Screen | Do This | Step 4e |
|---|---|---|
| Menu | ✳ **Transform** → ✳ **Recode into Same Variables** | 4.6 |

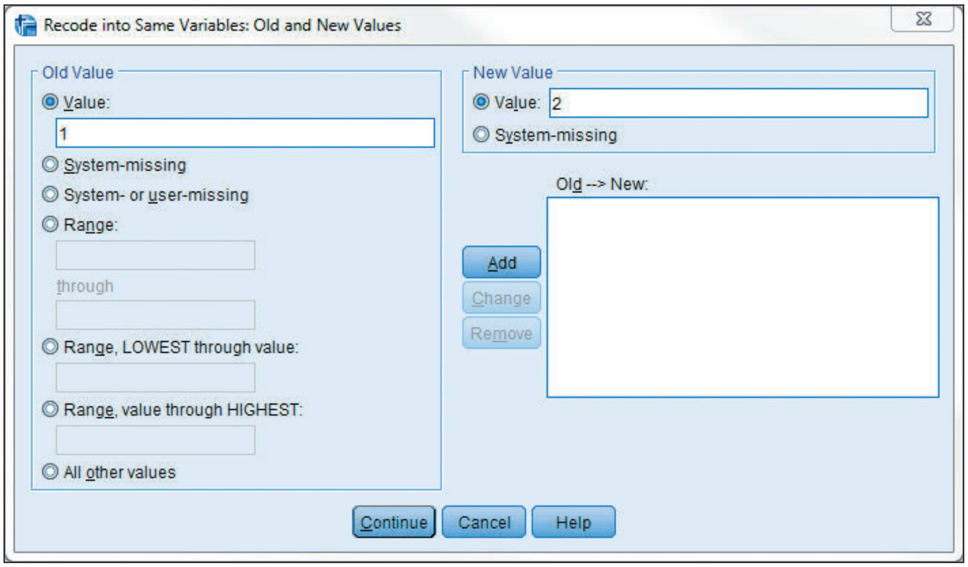
Screen 4.6 The Recode into Same Variables Window



Two different dialog boxes control (1) the selection of variable(s) to be recoded (Screen 4.6, above) and (2) clarification of old and new coding values (Screen 4.7, below). It is also possible to recode a string variable (such as **grade**) into numeric coding. A researcher might wish to recode the letters A, B, C, D, and F into 4, 3, 2, 1, and 0, so as to run correlations of **grade** with other variables. The **Recode** option can accommodate this comfortably. You can recode several variables in one setting but would need to switch back and forth between the two screens to do so.

In Screen 4.6 (above) the first step is to select the variable(s) of interest and paste them into the active box to the right (titled **Numeric Variables**). A click on **Old and New Values** opens up a new window (Screen 4.7).

Screen 4.7 The Old and New Values Window



Old Value: ● Value: 1

New Value: ● Value: 2

Options for **Old Value** include:

- A single value
- System-missing values (blanks in your data set)
- User-missing values (values you have designated as missing for various reasons)
- A range of values (internal range, lowest to specified value, specified value to highest)
- All other values (useful for complex combinations)

For each variable you specify the old value(s) and the new value(s), and then paste them into the **Old → New** box one level at a time. To demonstrate both a single switch recode and a recoding that involves a range of values, we will reverse the coding for the **gender** variable. Then, even though it's not in the **grades.sav** file, we recode the fictitious **marital** variable into levels 1, 2, and 3 representing unmarried and levels 4 and 5 representing married. If you wish to practice the latter procedure using the **grades.sav** file, simply use the **ethnicity** variable for recoding. Like the fictional *marital* variable, it also has five levels.

*The starting point is Screen 4.6. Perform the following sequence of steps to reverse the coding on the* **gender** *variable.*

| In Screen | Do This | Step 5e |
|---|---|---|
| 4.6 | ✳ **gender** ⟶ ◯ ➡ ⟶ ✳ **Old and New Values** | |
| 4.7 | ✳ *in* **Value** *box* ⟶ **type** 1 ⟶ **press** **TAB** *twice* ⟶ **type** 2 ⟶ ✳ **Add** | |
| | ✳ *in* **Value** *box* ⟶ **type** 2 ⟶ **press** **TAB** *twice* ⟶ **type** 1 ⟶ ✳ **Add** ⟶ | |
| | ✳ **Continue** | |
| 4.6 | ✳ **OK** | **Back1** |

*Perform the following sequence of steps to recode the mythical* **marital** *variable.*

| In Screen | Do This | Step 5e' |
|---|---|---|
| 4.6 | ✳ **marital** ⟶ ✳ ➡ ⟶ ✳ **Old and New Values** | |
| 4.7 | ✳ *first* **Range** *option* ⟶ ✳ *in* **Range** *box* ⟶ **type** 1 ⟶ **press** **TAB** ⟶ **type** 3 | |
| | ⟶ **press** **TAB** *twice* ⟶ **type** 1 ⟶ ◯ **Add** ⟶ | |
| | ✳ *in same* **Range** *box* ⟶ **type** 4 ⟶ **press** **TAB** ⟶ **type** 5 ⟶ | |
| | **press** **TAB** *twice* ⟶ **type** 2 ⟶ ✳ **Add** ⟶ ✳ **Continue** | |
| 4.6 | ✳ **OK** | **Back1** |

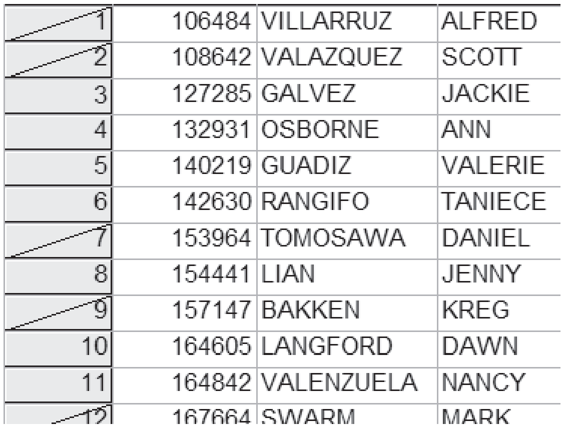
**Create new value labels!** To create new value labels for the recoded variables is an urgent concern. Otherwise Robert will be coded "female," Will Smith will be coded "Asian," and Elizabeth Taylor will be coded "never married." If you wish to create new value labels for the recoded variables, you will need to perform the appropriate operation (click on the **Variable View** tab and the appropriate cell in the **Values** column) described in Chapter 3. Further, if you want these changes to be permanent in your data file, you will need to go through the save data procedure (click **File** ⟶ click **Save**).

## 4.5 The Select Cases Option

**SELECTING A PORTION OF THE DATA FOR ANALYSIS** *To access the necessary initial screen, select the following options:*

| In Screen | Do This | Step 4f |
|---|---|---|
| Menu | ✳ **Data** ⟶ ✳ **Select Cases** | 4.8 |

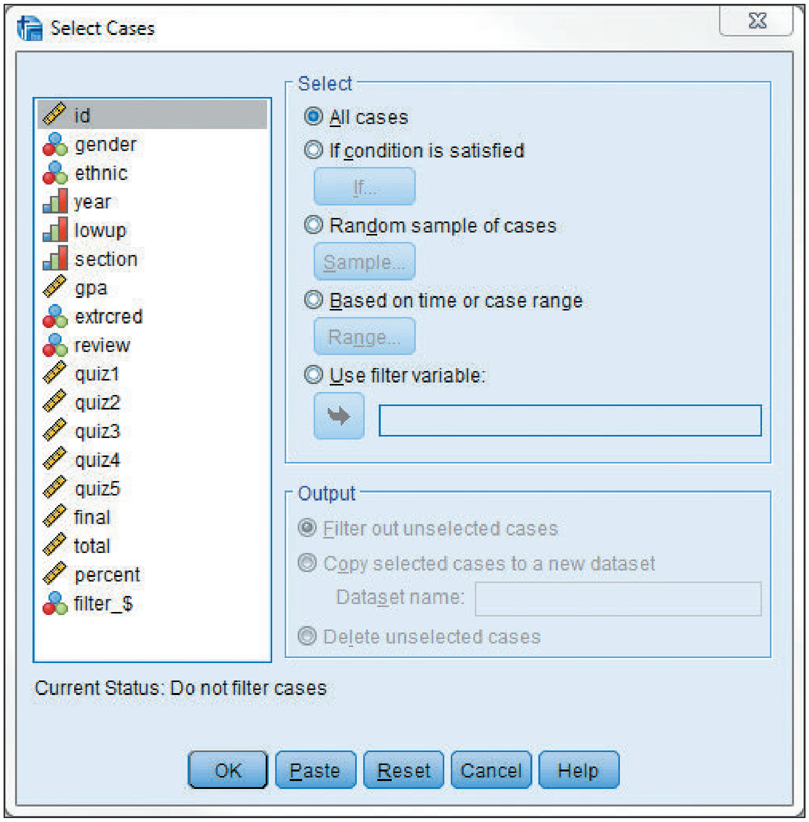
The purpose of the **Select Cases** option is to allow the user to conduct analyses on only a subset of data. This will be a frequently used function. Often we want to know what the mean **total** score is for females, or the average **total** score for Section 3, or the mean **total** score for sophomores. **Select Cases** enables you to accomplish this type of selection. It chooses from the data a subset of those data for analysis. Once analyses have been conducted with a subset of data, you may revert to the entire file by clicking **All cases** in the **Select Cases** dialog box. If you wish to create a file that consists only of the selected cases, you must first delete unselected cases and then save the file (click **File** → click **Save**). Then only the selected cases/subjects will remain. For ease of identifying cases that are selected and those that are not, in the data file a diagonal line is placed through the row number of nonselected cases. The example to the right illustrates females selected. The lines through case numbers represent males in the sample. Then, when certain cases are selected, SPSS creates a new variable named **filter_$** that codes selected cases as 1 and nonselected cases as 0. You may keep that variable for future reference (if you wish to make comparisons between the two groups), or you may delete it.



| | | | |
|---|---|---|---|
| 1 | 106484 | VILLARRUZ | ALFRED |
| 2 | 108642 | VALAZQUEZ | SCOTT |
| 3 | 127285 | GALVEZ | JACKIE |
| 4 | 132931 | OSBORNE | ANN |
| 5 | 140219 | GUADIZ | VALERIE |
| 6 | 142630 | RANGIFO | TANIECE |
| 7 | 153964 | TOMOSAWA | DANIEL |
| 8 | 154441 | LIAN | JENNY |
| 9 | 157147 | BAKKEN | KREG |
| 10 | 164605 | LANGFORD | DAWN |
| 11 | 164842 | VALENZUELA | NANCY |
| 12 | 167664 | SWARM | MARK |

In the **Select Cases** box (Screen 4.8) due to space constraints, we discuss only two options. The **All cases** option (mentioned above) simply turns off the select option, and further analyses will be conducted on all subjects. To access the screen for creating the conditional statements (Screen 4.9) click on the **If condition is satisfied** circle and then click the **If** button.

**Screen 4.8** The Select Cases Window



Select Cases

Select
● All cases
○ If condition is satisfied

○ Random sample of cases

○ Based on time or case range

○ Use filter variable:

Output
● Filter out unselected cases
○ Copy selected cases to a new dataset
Dataset name:
○ Delete unselected cases

Current Status: Do not filter cases

Screen 4.3 now opens. The actual screen that opens is not identical to Screen 4.3, but it is so similar that we'll not take up space by reproducing it here. In this screen is an active box where you may either type or click-and-paste relevant information. Below is the key pad, and to the right, the 188 functions lurking behind an innocent looking window. Quite a variety of conditional statements may be created. Several examples follow:

- *women*: **gender = 1**
- *third section*: **section = 3**
- *juniors and seniors*: **year >= 3**
- *sophomores and juniors*: **year >= 2 & year 1 & year > 4**
- *first-year students or seniors*: **year = 1 | year = 4**

To select a subset of variables the process is to click the desired variable and then paste it into the active box. You can then use the key pad (or simply type in the conditional statements) to create the selection you desire. In the sequences that follow we will demonstrate how to select women for an analysis. Then in another sequence we will show how to select sophomores and juniors.

*The starting point is Screen 4.8. Below we select females (***gender = 1***) for future analyses.*

| In Screen | Do This | Step 5f |
|---|---|---|
| 4.8 | **If condition is satisfied** → **If** | |
| 4.3 | *variable name* **gender** → ➡ → **type** = → **type** 1 → **Continue** | |
| 4.8 | **OK** | **Back1** |

*To select sophomores and juniors (***year = 2 and 3***) also begin at Screen 4.8.*

| In Screen | Do This | Step 5f' |
|---|---|---|
| 4.8 | **If condition is satisfied** → **If** | |
| 4.3 | **year** → ➡ → >= → **type** 2 → & → **year** → ➡ → | |
| | <= → **type** 3 → **Continue** | |
| 4.8 | **OK** | **Back1** |

## 4.6 The Sort Cases Procedure

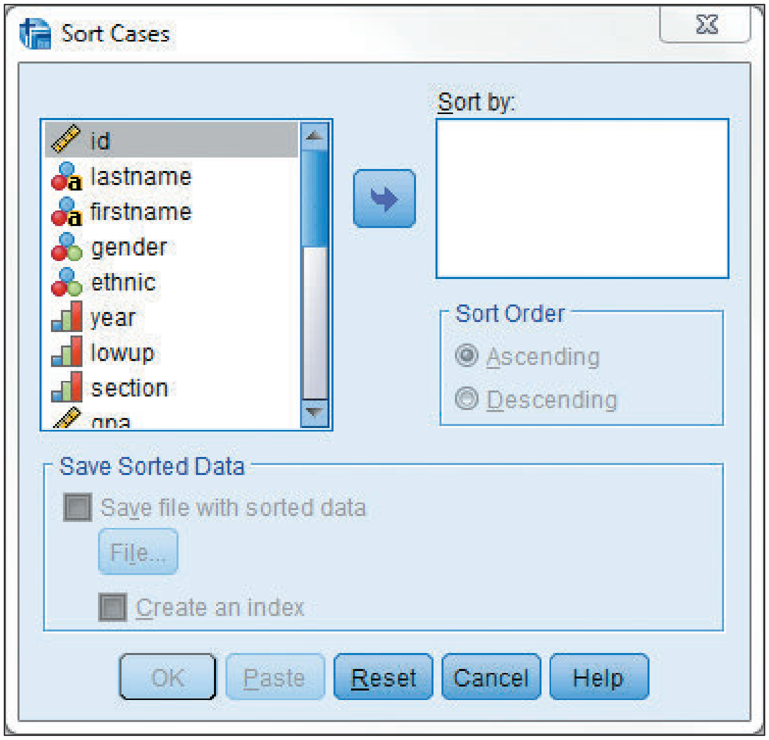
**REARRANGING YOUR DATA** There are many reasons one may wish to rearrange data. With the **grades.sav** file, the professor may wish to list final grades by **id** number. If so, it would be useful to be able to list scores with **id** numbers arranged from low to high. She may also be interested in the distribution of total scores of students in order to make grade breakdowns. Then it might be appropriate to list the file from high to low based on total points or total percentage. A class list is usually arranged alphabetically, but it is often necessary to list students alphabetically within each section. All these functions (and more) can be accomplished by the **Sort Cases** procedure.

*From any menu screen begin the process by selecting the following two options.*

| In Screen | Do This | Step 4g |
|---|---|---|
| Menu | ✳ **Data** → ✳ **Sort Cases** | 4.9 |

To sort cases a simple dialog box (Screen 4.9) handles the entire process. You simply select a variable of interest and specify if you wish the sort to be **Ascending** (small to large for numbers, alphabetic for string variables) or **Descending** (the opposite). An additional feature is that you can sort by more than one variable at a time. This is useful in sorting names. If you have several students with the same last name, then when sorting by **lastname**, **firstname**, SPSS will sort alphabetically initially by the last name and then by first name (for duplicate last names).

Screen 4.9 The Sort Cases Window



*To sort subjects by* **total** *points from high to low, execute the following steps:*

| In Screen | Do This | Step 5g |
|---|---|---|
| 4.9 | ✳ **total** → ○ ➡ → ✳ **Descending** → ✳ **OK** | Data |

*To sort subjects by* **lastname** *and* **firstname** *alphabetically, execute the following steps.*

| In Screen | Do This | Step 5g’ |
|---|---|---|
| 4.9 | ✳ **lastname** → ✳ ➡ → ✳ **firstname** → ✳ ➡ → ✳ **OK** | Data |

Note that **Ascending** is the default order so it doesn’t need to be clicked here.

# 4.7 Merging Files Adding Blocks of Variables or Cases

Merging files is a procedure that can baffle the most proficient of computer experts at times. Files can be constructed in a variety of different formats and be created in different data-analysis or word-processing programs. What we show here, however, are the simplest steps of merging files created in the IBM SPSS Statistics data editor, in the same format, in the same order, and with identical variable names (when we are adding more cases) or identical number and order of subjects (when we are adding more variables). While the last few sentences may sound apologetic, they actually contain sound advice for merging files: Prepare the foundation before you attempt the procedure. If possible perform the following steps before you begin:

- format files in the same data editor (i.e., IBM SPSS Statistics),
- create identical formats for each variable,
- make sure that matching variables have identical names,
- make sure that cases are in the same order (if you are adding new variables), and
- make sure that variables are in the same order (if you are adding new cases).

To be sure, there will be times that you are not able to do all of these things, but the more preparation beforehand, the less difficulty you will have in the merging process. For example, the present procedure allows you to merge an IBM SPSS Statistics file with an SPSS/PC+ file. Having used SPSS/PC+ many times I have a number of files in that format. If I wish to merge a PC+ file with a Windows file, even though I know it can be done, I will convert the PC+ file to IBM SPSS Statistics format before I attempt to merge them. Be especially careful to ensure that all variables have the same names and formats and are in the same order in the data file (the same order of cases is *required* when adding new variables). Adding new cases and adding new variables are two different procedures in SPSS. We will begin with adding cases.

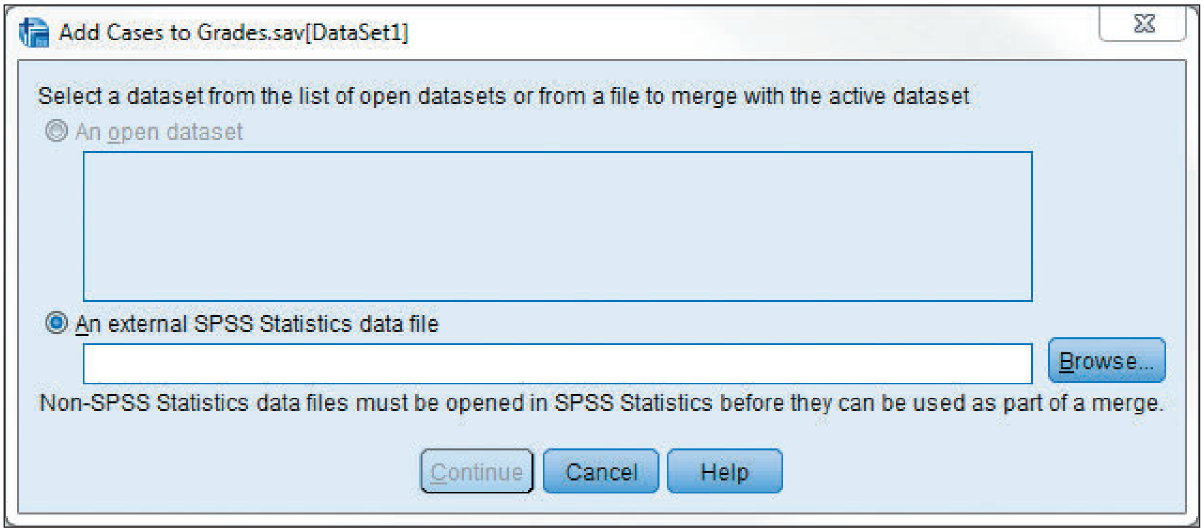
## 4.7.1 Adding Cases or Subjects

*After accessing the* **grades.sav** *file (Step 3 early in this chapter), from the menu of commands select the following options.*

| In Screen | Do This | Step 4h |
|---|---|---|
| Menu | Data → Merge Files → Add Cases | 4.10 |

At this point a new window will open (Screen 4.10, following page) titled **Add Cases to Grade.sav**. The **Grade.sav** file portion of the title identifies the open file that we wish to add cases to. This screen provides opportunity to read the file you wish to merge (called the *external data file*) with the already active file (**Grade.sav** in this case). For the sake of illustration, we have created a file named **graderow.sav** that contains 10 new subjects with identical variables and formats as the **grades.sav** file. **Graderow.sav** is the external file. Note the **Browse . . .** button. Click this button to find the external file you wish to merge. Note that the external file shown in the screen was read from a memory stick (E:\drive), and in addition to the file name it lists the location where it is found. When the correct file is identified, then click on the **Continue** button and Screen 4.11 will open (following page).

**Screen 4.10** The Add Cases: Read File Window

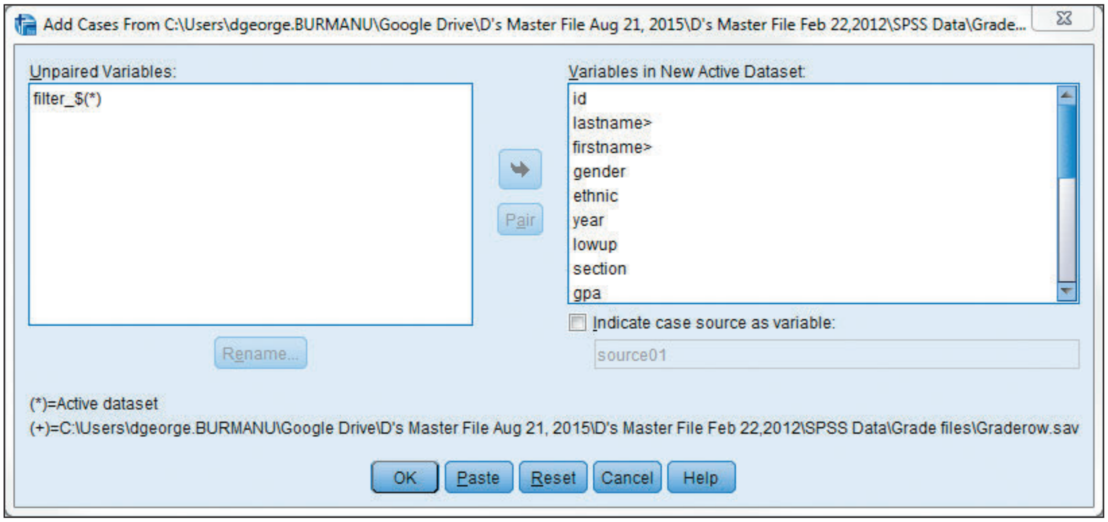

Add Cases to Grades.sav[DataSet1]
○ An open dataset
● An external SPSS Statistics data file
Browse...
Continue
Cancel
Help

Select a dataset from the list of open datasets or from a file to merge with the active dataset

Non-SPSS Statistics data files must be opened in SPSS Statistics before they can be used as part of a merge.

The next screen to appear is the **Add Cases from** box (Screen 4.11, below) with the name and source of your external file included in the title. Note in the title of Screen 4.11 the source of the external file (**E** drive), the directory (**SPSS Data**), and the name of the file name (**graderow.sav**) are all included. There are no **Unpaired Variables** in this example. If there were, unpaired variables would appear in the box to the left. Those from the original data file would be marked with an asterisk (*); those from the external data file would be marked with a plus (+). If you think the variable structures of the two files are identical and variable names *do* appear in the window to the left, then you need to examine those variables to make sure that names and formats are identical. Two simple operations are available before you click **OK** to approve the merger.

**Screen 4.11** The Add Cases from Drive, Program, and File Name Window


Add Cases From C:\Users\dgeorge.BURMANU\Google Drive\D's Master File Aug 21, 2015\D's Master File Feb 22,2012\SPSS Data\Grade...
Unpaired Variables:
filter_$(*)
Variables in New Active Dataset:
id
lastname>
firstname>
gender
ethnic
year
lowup
section
gpa
Pair
□ Indicate case source as variable:
source01
Rename...
OK
Paste
Reset
Cancel
Help

(*)=Active dataset
(+)=C:\Users\dgeorge.BURMANU\Google Drive\D's Master File Aug 21, 2015\D's Master File Feb 22,2012\SPSS Data\Grade files\Graderow.sav

- Any variables that are common to both files (paired variables, shown in the **Variables in the New Active Dataset** window) that you *do not* wish to have in the merged file you may click (to highlight) and delete by pressing the delete button.
- All unpaired variables are *excluded* from the merged file by default. If you wish any of these variables to be *included* click the desired variable(s) and press ➡ to paste them into the **Variables in New Active Dataset** box. Selected variable(s) will appear in the new file with blank cells where data are missing. These blank cells will be treated as missing values.

*To merge the* **graderow.sav** *file with the* **grades.sav** *file, perform the following sequence of steps. The starting point is Screen 4.10.*

| In Screen | Do This | Step 5h |
|---|---|---|
| 4.10 | ✳ **An external SPSS Statistics data file** → ✳ **Browse** → *after searching* | |
| | ✳ ✳ **graderow.sav** → ✳ **Continue** | |
| 4.11 | ✳ **OK** | Back1 |

What will result is a new data file that will need to be named and saved. This altered file can be saved with original name, or a **save as** function can save it as a new file and the original file may remain in its premerger form.

### 4.7.2 Adding Variables

*After accessing the* **grades.sav** *file (Step 3 early in this chapter), from the menu of commands select the following options.*

| In Screen | Do This | Step 4i |
|---|---|---|
| Menu | ✳ **Data** → **Merge Files** → ✳ **Add Variables** | 4.12 |

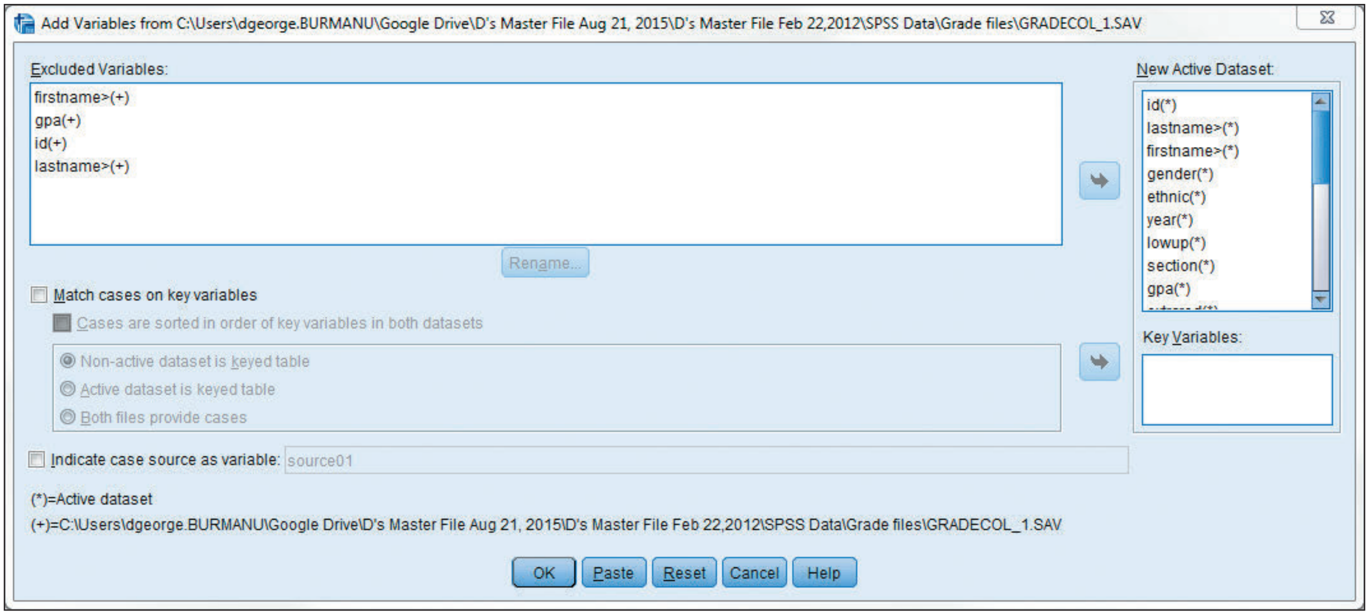
The procedure for adding variables is in many ways similar to that for adding cases or subjects. The initial entry procedure is the same. After clicking **Data**, **Merge Files**, *and* **Add Variables**, the same screen opens (Screen 4.10, page 78) as the screen used for adding cases. The purpose of this screen is to select an external file to merge with the working file. Upon clicking **Continue**, however, a new screen opens (Screen 4.12) that looks quite different from the screen for adding cases. The title is similar in that it identifies the drive, directory, and name of the external file to be merged. Note the new external file is **gradecol.sav**. This is a file that has the same subjects as the **grades.sav** file, includes several identical variables (**id**, **firstname**, **lastname**, **gpa**), and adds a new variable (**iq**) that identifies each subject's IQ.

- The matching variables (**firstname, gpa, id, lastname** in this case) will be listed in the **Excluded Variables** window and each will be followed by a "+" sign in parentheses.
- The variables that are in the original file that are not matched by the external file are listed in the **New Active Dataset** window and each one will be followed by an asterisk in parentheses.
- The new variable(s) to be added (**iq** in this case) is also in the **New Active Dataset** window and is followed by a "+" in parentheses.

Notice that Screen 4.12 (following page) has been created to show all this information in the appropriate boxes.

The procedure is to click on a matching variable in the external data file. It is required that this matching variable be in the same order in both files. Desirable matching variables might be **id** (ordered from low to high value) or **lastname** (ordered alphabetically). This requires first that you make sure that the matching variable is ordered identically in each file. Then, click the **Match cases on key variable in sorted files** option, select the matching variable in the box to the left, click the lower ➔ button, and click **OK**. Below is the step-by-step sequence for merging a working file (**grades.sav**) with an external file (**gradecol.sav**). In this example, the matching variable will be **id**, ordered from low to high value.

**Screen 4.12** The Add Variables: Read File Window


Add Variables from C:\Users\dgeorge.BURMANU\Google Drive\D's Master File Aug 21, 2015\D's Master File Feb 22,2012\SPSS Data\Grade files\GRADECOL_1.SAV

Excluded Variables:
firstname>(+)
gpa(+)
id(+)
lastname>(+)

New Active Dataset:
id(*)
lastname>(*)
firstname>(*)
gender(*)
ethnic(*)
year(*)
lowup(*)
section(*)
gpa(*)

Key Variables:

□ Match cases on key variables
■ Cases are sorted in order of key variables in both datasets
● Non-active dataset is keyed table
○ Active dataset is keyed table
○ Both files provide cases

□ Indicate case source as variable: source01

(*)=Active dataset
(+)=C:\Users\dgeorge.BURMANU\Google Drive\D's Master File Aug 21, 2015\D's Master File Feb 22,2012\SPSS Data\Grade files\GRADECOL_1.SAV

*The starting point for this sequence is Screen 4.10. Perform the following steps to merge the two files mentioned in the previous page.*

| In Screen | Do This | Step 5i |
|---|---|---|
| 4.10 | ✳ **An external SPSS data file** → ✳ **Browse** → *after searching* | |
| | ✳ ✳ **gradecol.sav** → ✳ **Continue** | |
| 4.12 | ✳ **Match cases on key variables in sorted files** → ✳ **id** *in variable list* | |
| | *to the left* → ✳ *lower* ➡ → ✳ **OK** | Back1 |

If the process doesn't work, check that the matching variables in the two files are in identical order.

## 4.8 Printing Results

Results of the analysis (or analyses) that have just been conducted require a window that displays the standard commands (**File Edit Data Transform Analyze . . .**) across the top. A typical print procedure is shown below beginning with the standard output screen (Screen 1, inside back cover).

*To print results, from the Output screen perform the following sequence of steps:*

| In Screen | Do This | Step 6 |
|---|---|---|
| Back1 | *Select desired output or edit (see pages 16–22)* → ✳ **File** → ✳ **Print** | |
| 2.9 | *Consider and select desired print options offered, then* → ✳ **OK** | |

*To exit you may begin from any screen that shows the* **File** *command at the top.*

| In Screen | Do This | Step 7 |
|---|---|---|
| Menu | 🖱 File → 🖱 Exit | 2.1 |

Note: After clicking **Exit**, there will frequently be small windows that appear asking if you wish to save or change anything. Simply click each appropriate response.

# Exercises

Answers to select exercises are downloaded at **www.spss-step-by-step.net**.

Some of the exercises that follow change the original data file. If you wish to leave the data in their original form, don't save your changes.

## Case Summaries

1. Using the **grades.sav** file, list variables (in the original order) from **id** to **quiz5**, first 30 students consecutive, fit on one page by editing.
2. Using the **helping3.sav** file, list variables **hclose**, **hseveret**, **hcontrot**, **angert**, **sympathi**, **worry**, **obligat**, **hcopet**, first 30 cases, fit on one page by editing.
3. List **ID**, **lastname**, **firstname**, **gender** for the first 30 students in the **grades.sav** file, with the lower division students listed first, followed by upper division students (**lowup** variable). Edit output to fit on one page.

## Missing Values

4. Using the **grades.sav** file delete the **quiz1** scores for the first 20 subjects. Replace the (now) missing scores with the average score for all other students in the class. Print out **lastname**, **firstname**, **quiz1** for the first 30 students. Edit to fit on one page. Don't save the file!

## Computing Variables

5. Using the **grades.sav** file calculate **total** (the sum of all five quizzes and the final) and **percent** (100 times the **total** divided by possible points, 125). Since **total** and **percent** are already present, name the new variables **total1** and **percent1**. Print out **id**, **total**, **total1**, **percent**, **percent1**, first 30 subjects. **Total** and **total1** and **percent** and **percent1** should be identical.
6. Using the **divorce.sav** file compute a variable named **spirit** (spirituality) that is the mean of **sp8** through **sp57** (there should be 18 of them). Print out **id, sex**, and the new variable **spirit**, first 30 cases, edit to fit on one page.
7. Using the **grades.sav** file, compute a variable named **quizsum** that is the sum of **quiz1** through **quiz5**. Print out variables **id**, **lastname**, **firstname**, and the new variable **quizsum**, first 30, all on one page.

## Recode Variables

8. Using the **grades.sav** file, compute a variable named **grade1** according to the instructions on page 70. Print out variables **id, lastname, firstname, grade** and the new variable **grade1**, first 30, edit to fit all on one page. If done correctly, **grade** and **grade1** should be identical.
9. Using the **grades.sav** file, recode a **passfail1** variable so that Ds and Fs are failing and As, Bs, and Cs are passing. Print out variables **id, grade, passfail1**, first 30, and edit to fit all on one page.
10. Using the **helping3.sav** file, redo the coding of the ethnic variable so that **Black** = 1, **Hispanic** = 2, **Asian** = 3, **Caucasian** = 4, and **Other/DTS** = 5. Now change the value labels to be consistent with reality (that is the coding numbers are different but the labels are consistent with the original ethnicity). Print out the variables **id** and **ethnic** (labels, not values), first 30 cases, fit on one page.

## Selecting Cases

11. Using the **divorce.sav** file select females (**sex** = 1), print out **id** and **sex**, first 30 subjects, numbered, fit on one page.
12. Select all the students in the **grades.sav** file with previous **GPA** less than 2.00, and **percent**ages for the class greater than 85. Print **id**, **GPA**, and **percent** on one page.
13. Using the **helping3.sav** file, select females (**gender** = 1) who spend more than the average amount of time helping (**thelplnz** > 0). Print out **id**, **gender**, **thelplnz**, first 30 subjects, numbered, fit on one page.

## Sorting Cases

14. Alphabetize the **grades.sav** file by **lastname**, **firstname**. Print out **lastname**, **firstname**, first 30 cases, edit to fit on one page.
15. Using the **grades.sav** file, sort by **id** (ascending order). Print out **id**, **total**, **percent**, and **grade**, first 30 subjects, fit on one page.

## Chapter 5

# Graphs and Charts: Creating and Editing

IBM SPSS Statistics possesses impressive and dramatic graphics capabilities. It produces high-quality graphs and charts, and editing and enhancement options are extensive. Furthermore, there are two complete sets of graphing procedures available in SPSS: **Legacy Dialogs** graphs and **Chart Builder** graphs. The Chart Builder option was first introduced in SPSS 14.0 and appears to be the format SPSS will stick with for the foreseeable future. From release to release they will continue to enhance the Chart Builder sequences. Therefore, all graphs (except for error bar charts) described in this section will be Chart Builder graphs.

## 5.1 Comparison of the Two Graphs Options

**Legacy Dialogs** graphs**:** These are a carryover from the style of graphs used in SPSS since version 7.5. The look of introductory dialog boxes is identical to earlier versions of SPSS. The legacy graphs are retained for the benefit of those who wish to remain with a familiar system. The only change (and a significant one) from earlier versions is that when you double click on the graph to edit, the edits are now identical to edits for the Chart Builder graphs.

**Chart Builder** graphs**:** Chart Builder provides the most professional looking graphs and a set of options that allows just about any edits you might desire. The dialog box screens allow the simplicity of click and drag for creating the structure of the original graph. We acknowledge that the edits are so extensive they can be confusing at times. We will assist you with what we consider the most frequently desired edits.

We now devote our attention to graphs created with the Chart Builder option. We present seven frequently used graphs (listed below) and then provide a limited number of editing options. SPSS, of course, has far more options than we present here. They devote over 300 pages (compared to our 16 pages) to the description of graphs and their edits in the SPSS manuals. Clearly we are obliged to present only a subset. The seven types of graphs follow.

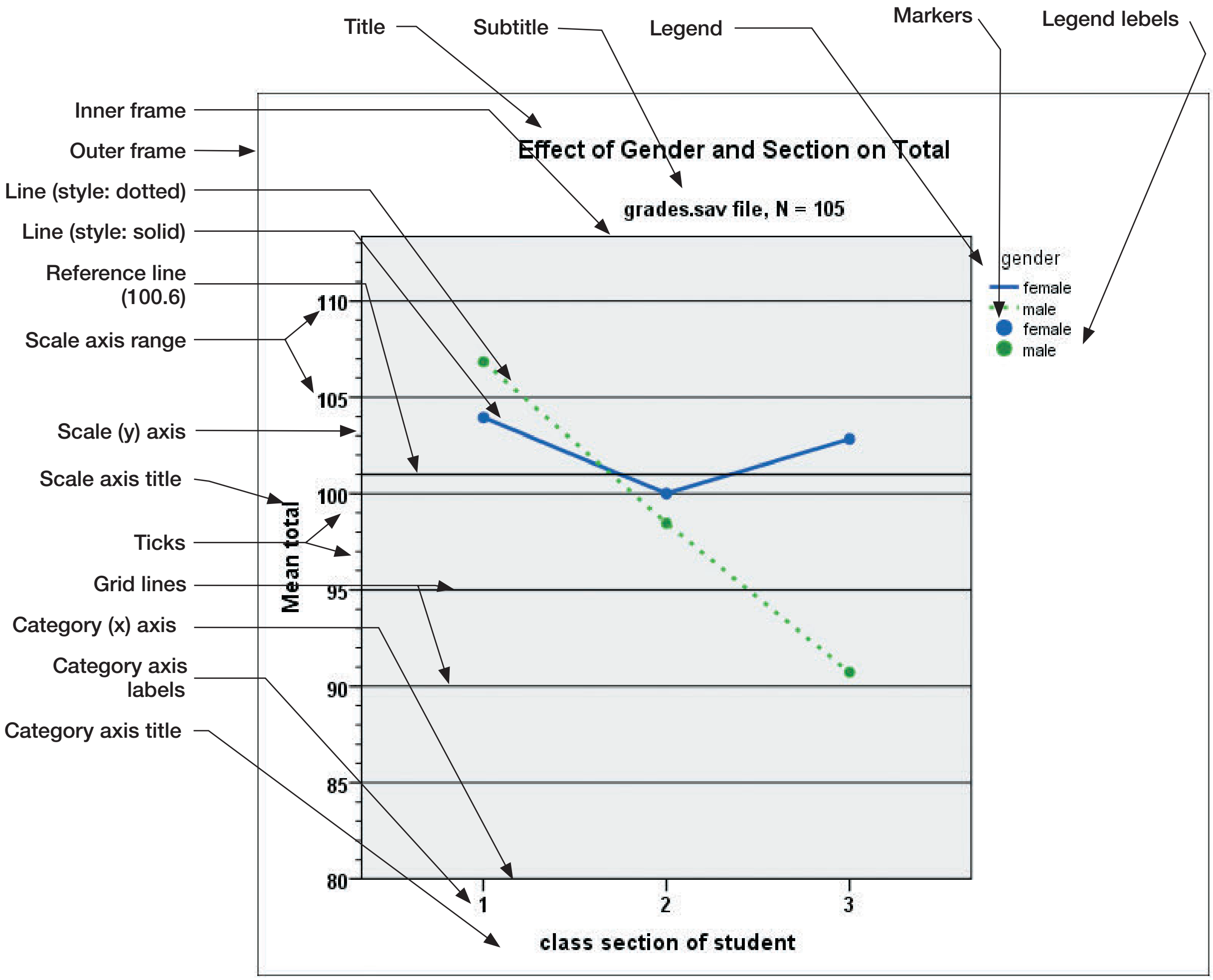
## 5.2 Types of Graphs Described

- **Bar graphs**: Bar graphs are used most often to display the distribution of subjects or cases in certain categories, such as the number of A, B, C, D, and F grades in a particular class.
- **Line graphs**: Line graphs may be used to display trends in data and multiple line charts are often employed to demonstrate ANOVA interactions.

- **Pie charts**: Pie charts, like bar charts, are another popular way of displaying the number of subjects or cases within different subsets of categorical data.
- **Box plots**: Box plots are based on percentiles and provide an excellent vehicle to display the distribution of your data.
- **Error bar charts**: Produces a chart that includes error bars indicating the standard error of measurement, or indicating the confidence interval. This is useful for visually clarifying which differences are significant and which are not significant.
- **Histograms**: Histograms look similar to bar graphs but are used more often to indicate the number of subjects or cases in ranges of values for a continuous variable, such as the number of students who scored between 90 and 100, between 80 and 89, between 70 and 79, and so forth on a final exam.
- **Scatter plots** (simple and overlay): Scatter plots are a popular way of displaying the nature of correlations between variables.

## 5.3 The Sample Graph

In the graph that follows, we attempt to crowd as much SPSS graphics terminology into one graph as possible. This provides a clear reference as you read through the chapter. When you encounter a word such as *category axis*, and don't know what it means, a quick glance back at this chart provides an example of a category axis. We cannot in one graph illustrate all terms, but we display most of them in the chart below. Be aware that some terminology is idiosyncratic to the SPSS graph editor.



## 5.4 Producing Graphs and Charts

Although there are a wide variety of graphs, there are certain features they have in common. They almost all have an X-axis and a Y-axis (with a certain variable graphed on the X-axis and a different variable graphed on the Y-axis), and something in the data region of the graph (dots, lines, bars) to indicate the relation between the X-axis variable and the Y-axis variable. The following steps indicate the general procedure to produce a graph. In this operation, a line graph is selected (similar to the sample graph, previous page). The steps here focus on the general operations involved in creating any graph; later in the chapter, a description of some of the key additional options for creating particular kinds of graphs are discussed. A note for advanced users: For many procedures, you can select certain values in the SPSS output, right-click, and select **Create Graph** and then select the kind of graph that you want to create. If you are not an advanced user already familiar with the SPSS graphing procedures, however, this can go very wrong or be misinterpreted. The usual (safer) steps to access the **Chart Builder** graphs follow:

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | ✳ → ✳ All Programs ▷ → ✳ IBM SPSS Statistics → ✳ IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | ✳ → ✳ → ✳ | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| | Step 2 |
|---|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* | |

*Screens 1 and 2 (displayed on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **grades.sav** *file for further analyses:*

| In Screen | Do This | Step 3 |
|---|---|---|
| Front1 | ✳ File → Open → ✳ Data [*or* ✳ ] | |
| Front2 | type `grades.sav` → ✳ Open [*or* ✳ ✳ grades.sav] | Data |

*You may select any kind of graph available in Step 4; for example, we have chosen a* **Line** *graph (below). You could also choose a* **Bar** *chart,* **Pie/Polar** *chart,* **Boxplot**, **Histogram**, *or* **Scatter/Dot** chart.

| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | ✳ Graphs → Chart Builder → ✳ Line | 5.1 |

Once Screen 5.1 appears (following page), the starting point to create the graph you wish emerges. The following features are of particular interest to us. First, in the window to the upper left are the variables available for creating graphs. In Screen 5.1, the variables come from the **grades.sav** file. The **Chart preview** window to the right (blank except for some instructions) is where you will select specifications for your graph. Directly beneath (under the title **Choose from**) are the basic classes of available graphs. Currently selected is the **Line** graph, and the two charts to the right identify two choices—either a single-line graph or a multiple-line chart. The horizontal tabs in the center (**Gallery**, **Basic Elements**, . . .) that provide certain editing options or alternative ways to access certain graphs will not be addressed in this chapter. Any edits described in this chapter will take place after the graph has been created.

To begin the graph sequence you will click and drag one of the pictures in the screen (to the lower right) into the **Chart preview** window just above. The line graph (shown) allows two options (single- or multiple-line charts). By contrast, the bar graph (**Bar**) provides eight options.

There are small icons to the left of the variable names. These icons identify the variables as:

- Numeric/scale (📏)
- Numeric/ordinal (📊)
- Numeric/nominal (●)
- String/nominal (●a)

### Screen 5.1 The Chart Builder Window (Line Chart Selected)

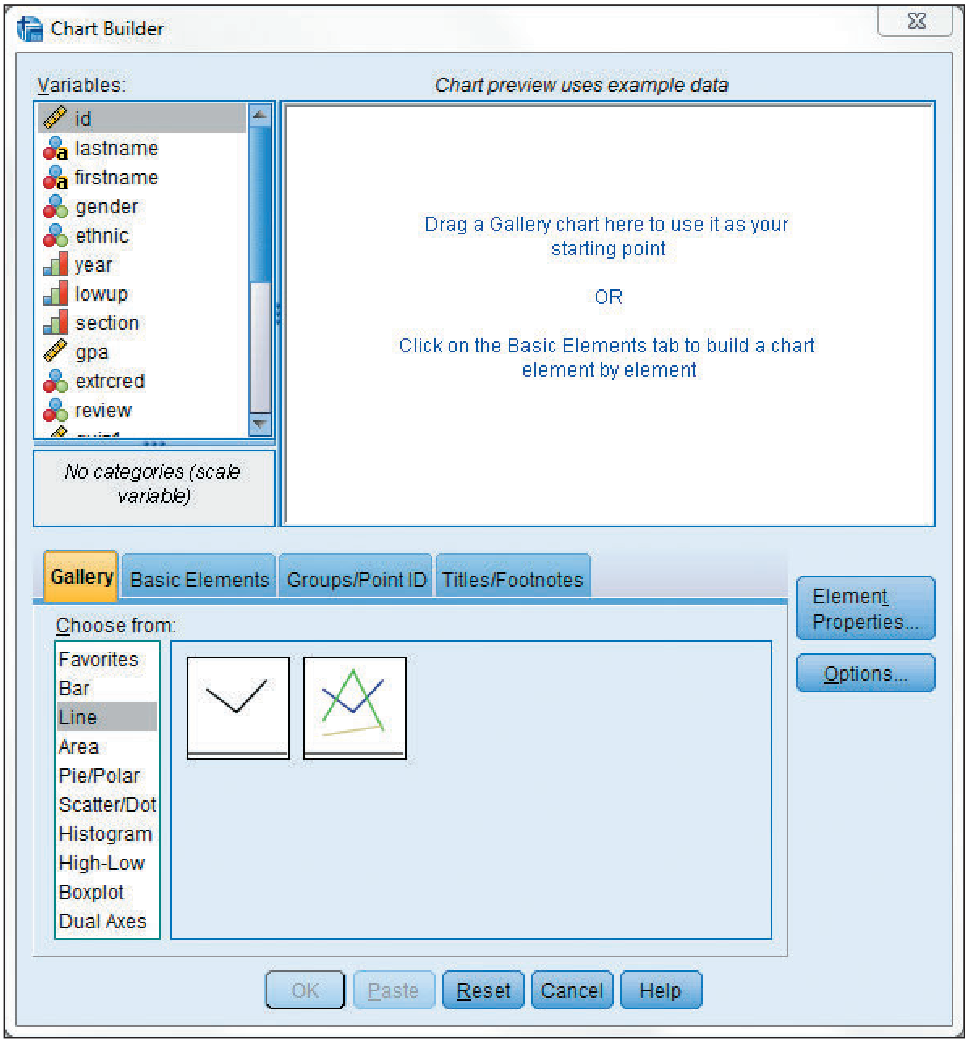


*Chart preview uses example data*

Drag a Gallery chart here to use it as your starting point

OR

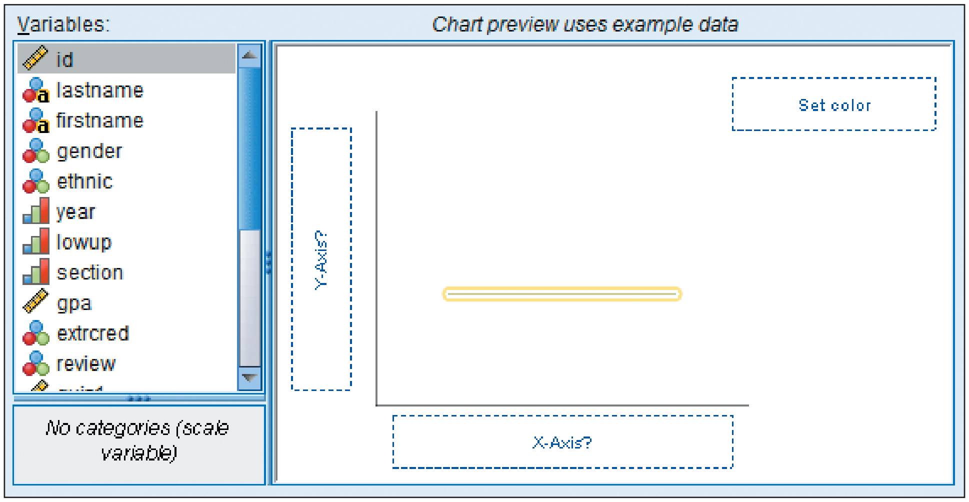
Click on the Basic Elements tab to build a chart element by element

*No categories (scale variable)*

If you try to produce a graph with, for instance, a scale variable when SPSS expects a nominal variable (or a nominal variable where a scale variable is expected), SPSS will ask if you want to change the variable from one kind to the other. When this happens, it usually means that you made a mistake (when selecting the variable, or when setting it up in your data file), or you want an unusual graph.

When you click and drag one of the chart icons into the **Chart preview** box (Screen 5.2), the window changes to allow you to specify characteristics of your desired chart. What appears here will differ substantially for different types of graphs. Each option, however, allows you to select the specifications necessary for that graph.

Screen 5.2 Chart Preview Window for a Multiple Line Chart



For a line chart Screen 5.2 allows you to specify the x-axis variable (**X-Axis?**), the y-axis variable (**Y-Axis?**), and the legend variable (**Set color**). The color will distinguish between different levels of the legend. If, for instance, gender was the legend variable, the graph would produce different color lines, markers, or bars for men and women.

If we wanted to create a multiple line chart that shows how men's and women's (**gender**) total points (**total**) differed in each of the three sections (**section**) you would perform the following steps: Click and drag **total** to the **Y-axis?** box; click and drag **section** to the **X-axis?** box; and then click and drag **gender** to the **Set color** box. A click on **OK** and that graph will be produced. Once produced, the graph can be edited to enhance its appearance or to increase clarity.

## 5.5 Bugs

For our book dealing with SPSS 17.0 (the 10th edition), we included a half page of warning about a number of bugs in the graphing program guaranteed to gray your hair and trigger your chocolate cravings. For the next several editions of the book a number of those bugs have been rectified. We were hoping to delete this paragraph for the present (14th) edition, but a few still remain—different bugs to be sure, but they are present: For example, the histogram in its original form is still meaningless without edits, the multiple regression lines (for a scatterplot with two predictor variables) look like they're on acid, and you will get frustrated from time to time attempting to edit lines or to fill bars with different designs or colors. Nevertheless we know that SPSS has and are certain that they will continue to improve the product, and, if you are persistent, you usually get the desired results.

## 5.6 Specific Graphs Summarized

The previous section describes in general terms how to create graphs. Those instructions apply to any graph you want to produce. This section focuses on certain graphs, addressing particularly useful or effective editing options that are specific to those graphs.

### 5.6.1 Bar Charts

*In the sequence that follows, we create a clustered bar chart showing how men and women differed in total points scored in each of the three sections.*

| In Screen | Do This | Step 5 |
|---|---|---|
| Menu | ✳ **Graphs** → ✳ **Chart Builder** → ✳ **Bar** | |
| 5.1 | ✳ 📊 → *drag to* **Chart preview** *box (directly above)* | |
| 5.2 | ✳ **total** → *drag to* **Y-Axis?** *box* → ✳ **section** → *drag to* **X-Axis?** *box* → | |
| | ✳ **gender** → *drag to* **Cluster on X: set color** *box* → ✳ **OK** | Back1 |

The graph now appears on the Output screen, ready to edit or print. Common editing options follow: Be aware that access to editing options always begins with a double click on the graph itself. We begin with changing the scale and increment values. Small boxes (following the description of each editing option) provide a mini version of a step-by-step sequence.

**Changing the scale and increment values**: We may decide that we don't want such a tall graph (scaled from 0 to 120) and would prefer to have more frequent increments (every 10 points rather than the default every 20). Notice that we leave the maximum value automatic.

✳✳ on graph → ✳✳ y-axis values → ✳ **Scale** tab → ✳✳ inside **Minimum** box → **type** 80 → ✳✳ inside **Major Increment** box → **type** 10 → ✳✳ inside **Origin** box → **type** 80 → ✳ **Apply** → ✳ ☒ to return

**Adding grid lines**: To enhance clarity we may wish to add grid lines to the chart. The following sequence will place horizontal grid lines at 5-point intervals.

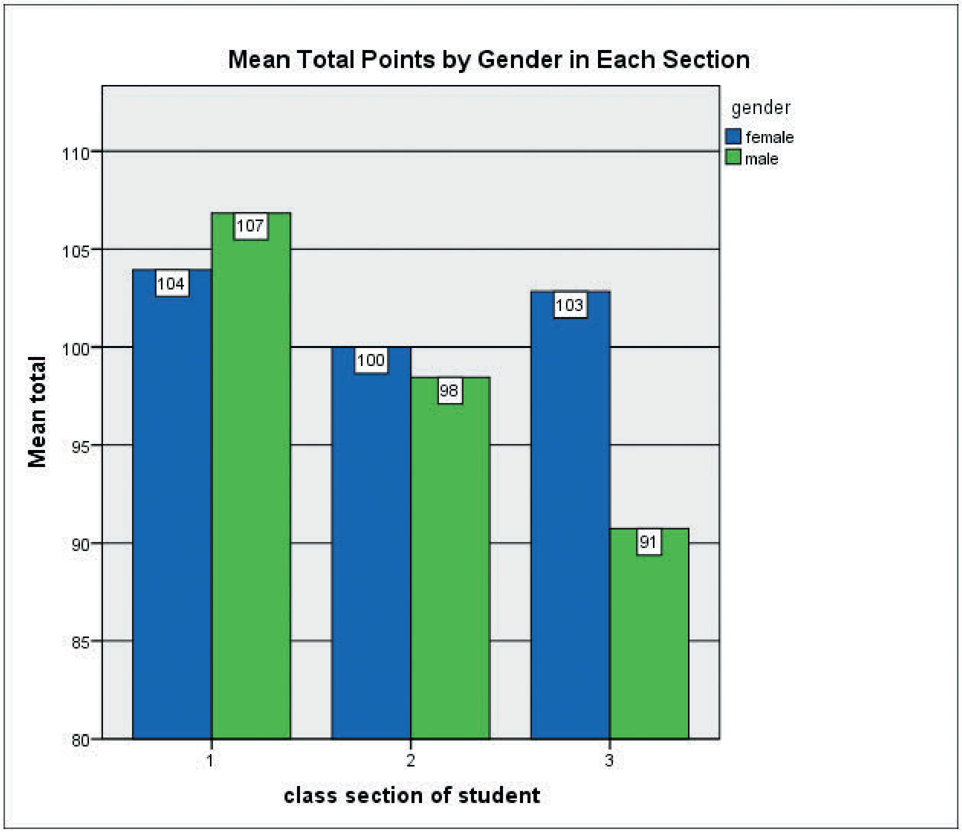
✳✳ on graph → ✳ y-axis values → ✳ ▦ icon (above) → ✳ **Both major and minor ticks** → ✳ **Apply** → ✳ ☒ to return

**Adding labels inside the bars**: We may wish to include the actual values (mean value of **total** points in this case) inside each of the bars. The following sequence will provide the mean **total** points, accurate to one decimal, in bordered boxes at the top of each bar.

✳✳ on graph → ✳ 📊 icon → ✳ **Custom** → ✳ ▣ icon → ✳ **Apply** → ✳ **Number format** tab → ✳✳ inside **Decimal Places** box → **type** 1 → ✳ **Apply** → ✳ ☒ to return

**Adding a Title and Changing the Font**: The following sequence will create a title for the Chart and then change the font size to 14.

on graph → icon → highlight **Title** (using the red cursor line) → **type** Mean Total Points by Gender in each Section → press **Enter** → on the (just created) title → next to **Preferred Size** → 14 → **Apply** → to return



The steps listed above produce the bar graph shown here. Some sequences are common to all graphs (such as adding a title and changing the font). Others are unique to bar graphs. Additional options include:

**1**) Changing the format, font, or size of axis labels or values along each axis ( on the label or value you wish to edit and then edit using the tabbed table to the right);

**2**) Changing the wording of axis labels or values along each axis ( on the label or value you wish to edit, a second time and a red cursor appears that allows you to edit the text); or

**3**) Change the orientation of the graph ( if you want the bars to be horizontal; again if you wish to change orientation back to vertical).

## 5.6.2 Line Graphs

Line graphs may be used to display trends in data and multiple line charts are often employed to demonstrate ANOVA interactions. To create a line graph, select **Graphs** → **Chart Builder** → **Line**. The dialog box in Screen 5.1 appears as described on pages 85–86. We now provide the step-by-step sequence to create the graph.

*The sequence step below produces a line graph (similar to the Sample Chart) with total points (**total**) on the Y-axis, class section (**section**) on the X-axis, and separate lines (with different colors) for females and males (**gender**).*

| In Screen | Do This | Step 5a |
|---|---|---|
| Menu | ✳ **Graphs** → ✳ **Chart Builder** → ✳ **Line** | |
| 5.1 | ✳ 📈 → *drag to* **Chart preview** *box (directly above)* | |
| 5.2 | ✳ **total** → *drag to* **Y-Axis?** *box* → ✳ **section** → *drag to* **X-Axis?** *box* → | |
| | ✳ **gender** → *drag to* **Set color** *box* → ✳ **OK** | Back1 |

The graph appears on the Output screen, ready to edit or print. Certain editing options follow:

**Changing the scale and increment values**: We may decide that we don't want such a tall graph (scaled from 0 to 120) and would prefer to have more frequent increments (every 10 points rather than the default every 20). Notice that we leave the maximum value automatic.

✳✳ on graph → ✳✳ y-axis values → ✳ **Scale** tab → ✳✳ inside **Minimum** box → **type** 80 → ✳✳ inside **Major Increment** box → **type** 10 → ✳✳ inside **Origin** box → **type** 80 → ✳ **Apply** → ✳ ☒ to return

**Adding grid lines**: To enhance clarity we may wish to add grid lines to the chart. The following sequence will place horizontal grid lines at 5-point intervals.

✳✳ on graph → ✳ y-axis values → ✳ ▦ icon (above) → ✳ **Both major and minor ticks** → ✳ **Apply** → ✳ ☒ to return

**Adding markers on the lines**: Since the color of the lines may not show up well (on a black & white printer) it is desirable to have different markers for men and women at each value point.

✳✳ on graph → ✳ 📈 icon → ✳ **O** next to **female** (legend) → ✳ **Marker** tab → ✳ desired color → ✳ **Apply** → ✳ **O** next to **male** → ✳ ⌄ next to **Type** → ✳ desired shape → ✳ desired color → ✳ **Apply** → ☒ to return

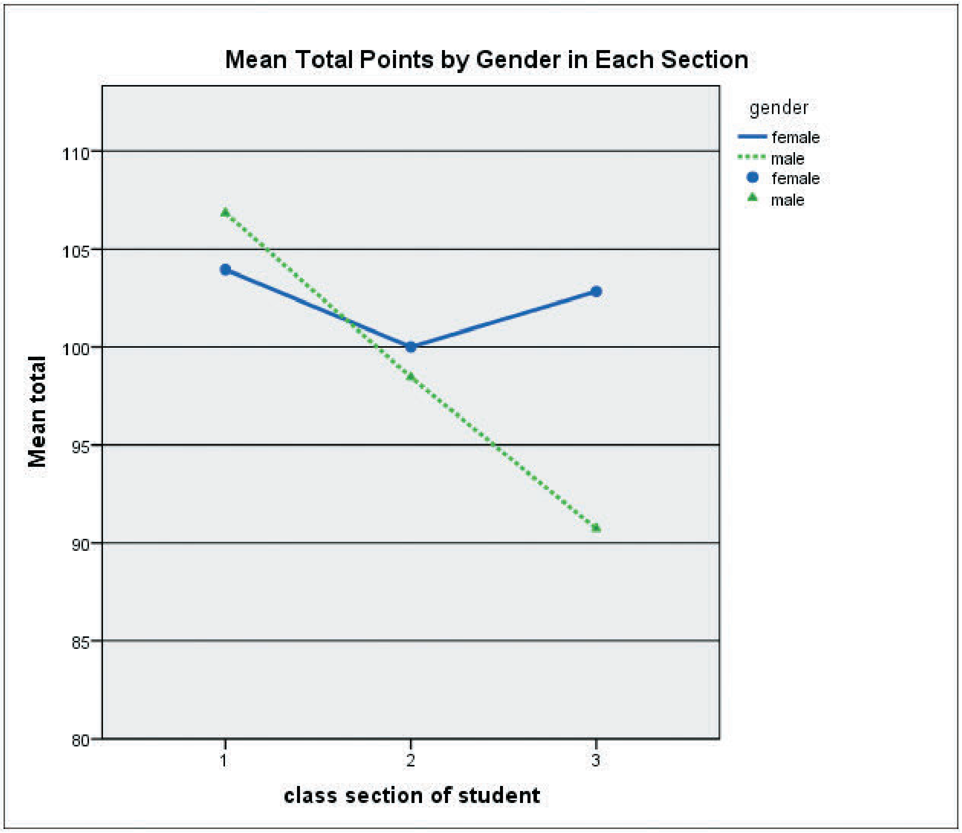
**Changing line styles and weight**: This will allow you to change the style and weight of the two lines so that you can distinguish between them when you have no color enhancement.

✳✳ on graph → ✳ on the line to the left of **female** → ✳ ⌄ to the right of **weight** → ✳ 2.0 → ✳ **Apply** → ✳ on the line to the left of **male** → ✳ ⌄ to the right of **weight** → ✳ 2.0 → ✳ ⌄ to the right of **style** → ✳ a "dotted" option → ✳ **Apply** → ✳ ☒ to return

**Adding a Title and Changing the Font**: The following sequence will create a title for the Chart and then change the font size to 14.

🖱🖱 on graph → 🖱 icon → highlight **Title** (using the red cursor line) → **type** Mean Total Points by Gender in each Section → press **Enter** → 🖱 on the (just created) title → 🖱 ▾ next to **Preferred Size** → 🖱 14 → 🖱 **Apply** → 🖱 ☒ to return

The graph produced by the sequences listed above is displayed here. Additional editing options include:



**1**) Changing the format, font, or size of axis labels or values along each axis (🖱🖱 on the label or value you wish to edit and then edit using the tabbed table to the right);

**2**) Changing the wording of axis labels or values along each axis (🖱 on the label or value you wish to edit, 🖱 a second time and a red cursor appears that allows you to edit the text);

**3**) Change the orientation of the graph: (🖱 if you want the axes switched; 🖱 again if you wish to change back to the original); or

**4**) It is also possible to add a text box (🖱), add sub headings (🖱), add foot notes (🖱), or include a reference line (perhaps a horizontal line representing the grand mean) (🖱), and a variety of other options.

## 5.6.3 Pie Charts

Pie charts, like bar charts, are another popular way of displaying the number of subjects or cases within different subsets of categorical data. To create a pie chart, select **Graphs** → **Chart Builder** → **Pie/Polar**. A dialog box similar to Screen 5.2 appears. Because pie charts do not have X- and Y-axes, it is very easy to create a pie chart: Just drag the categorical variable that you desire to the **Slice by?** box. A click on **OK** creates a pie chart with frequencies for each level of the selected variable.

*The sequence step below produces a pie chart where each slice of the pie represents the number of different grades in the class. The meaning of each slice is identified in the legend.*

| In Screen | Do This | Step 5b |
|---|---|---|
| Menu | ✳ **Graphs** → ✳ **Chart Builder** → ✳ **Pie/Polar** | |
| 5.1 | ✳ ◕ → *drag to* ***Chart preview*** *box (directly above)* | |
| 5.2 | ✳ **grade** → *drag to* ***Slice by?*** *box* → ✳ **OK** | Back1 |

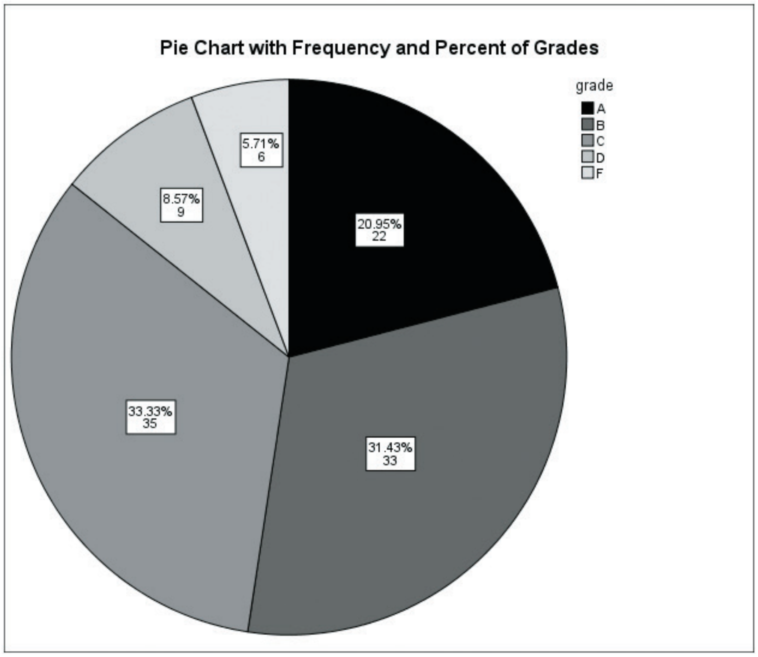
The graph appears on the Output screen, ready to edit or print. Certain editing options follow:

**Inserting percents and frequencies, changing the color of slices and inserting a title**: We employ a single sequence box of several edits to produce the graph shown below.

✳ ✳ on graph → ✳ icon → ✳ and drag **Count** to **Displayed** box → ✳ **Apply** → ✳ □ A → ✳ black → ✳ **Apply** → ✳ □ B → ✳ dark gray → ✳ **Apply** → ✳ □ C → ✳ med gray → ✳ **Apply** → ✳ □ D → ✳ light gray → ✳ **Apply** → ✳ F → ✳ white → ✳ **Apply** → ✳ icon → highlight **Title** (using the red cursor line) → **type** Pie Chart with Frequency and Percent of Grades → ✳ to return

The reason for the many clicks to change the color of the slices is to be able to distinguish between each slice, hence, which slice is associated with which grade. Another option is to click the ▼ next to **Pattern** and then select different textures for each pie slice. The legend will then reflect the color or pattern that is used to designate each grade.

If you wish to "explode" a pie slice (fortunately this does not damage your computer) click on the desired slice, click on the ◕ icon and the pie slice will separate. If you wish it to return to its original position, click once again on the ◕ icon.



## 5.6.4 Boxplots

Boxplots are based on percentiles and provide an excellent vehicle to display the distribution of your data. Boxplots also label outliers farther than 1.5 times the interquartile range away from the median so that you can give these cases special treatment (for example, by examining these cases and be sure they are not mistakes). To create a boxplot, select **Graphs** → **Chart Builder** → **Boxplot**. A dialog box very similar to Screen 5.2 appears.

*The sequence step below produces a graph with total points (**total**) on the Y-axis, class section (**section**) on the X-axis, and separate boxplots for females and males (**gender**).*

| In Screen | Do This | Step 5c |
|---|---|---|
| Menu | ✳ **Graphs** → ✳ **Chart Builder** → 🖱 **Boxplot** | |
| 5.1 | ✳ → *drag to* **Chart preview** *box (directly above)* | |
| 5.2 | ✳ **total** → *drag to* **Y-Axis?** *box* → ✳ **section** → *drag to* **X-Axis?** *box* → | |
| | ✳ **gender** → *drag to* **Cluster on X: set color** *box* → ✳ **OK** | Back1 |

The graph appears on the Output screen, ready to edit or print. Certain editing options follow:

**Using pattern rather than color to distinguish between bars**: Unless you print out your graphs on a color printer, patterns within bars often provides a better resource for differentiating between groups than does color.

✳ ✳ on graph → ✳ on □ next to **male** in the legend → right-✳ on a "male" bar in the chart → move cursor to **Select** → ✳ **This Group of Box** → ✳ **Fill & Border** tab → ✳ ∨ to the right of **pattern** → ✳ desired pattern → ✳ **Apply** → ✳ ☒ to return

**Changing the thickness of the bars and clustering**: The original graph is adequate, but looks better with thicker bars and more distinct clustering.

✳ ✳ on graph → ✳ ✳ one of the bars → ✳ **Bar options** tab → ✳ ▽ (under **Bars**) and drag right until 60% shows → ✳ ▽ (under **Clusters**) and drag left until 75% shows → ✳ **Apply** → ✳ ☒ to return

**Changing the scale values**: Although this graph has no crying need to be changed, it is frequently an important step. We change the maximum value to 130 to make a shorter graph.

✳ ✳ on graph → ✳ ✳ y-axis values → ✳ **Scale** tab → ✳ ✳ inside **Maximum** box → **type** 130 → ✳ **Apply** → ✳ ☒ to return

**Adding grid lines**: To enhance clarity we may wish to add grid lines to the chart. The following sequence will place horizontal grid lines at 10-point intervals.

✳ ✳ on graph → ✳ y-axis values → ✳ ▦ icon (above) → ✳ **Both major and minor ticks** → ✳ **Apply** → ✳ ☒ to return

**Adding a Title and Changing the Font**: The following sequence will create a title for the Chart and then change the font size to 14.

🖱 🖱 on graph ⟶ 🖱 icon ⟶ highlight **Title** (using the red cursor line) ⟶ **type** Boxplot of Total Points by Gender in each Section ⟶ press **Enter** ⟶ 🖱 on the (just created) title ⟶ 🖱 ▾ next to **Preferred Size** ⟶ 🖱 14 ⟶ 🖱 **Apply** ⟶ 🖱 ☒ to return

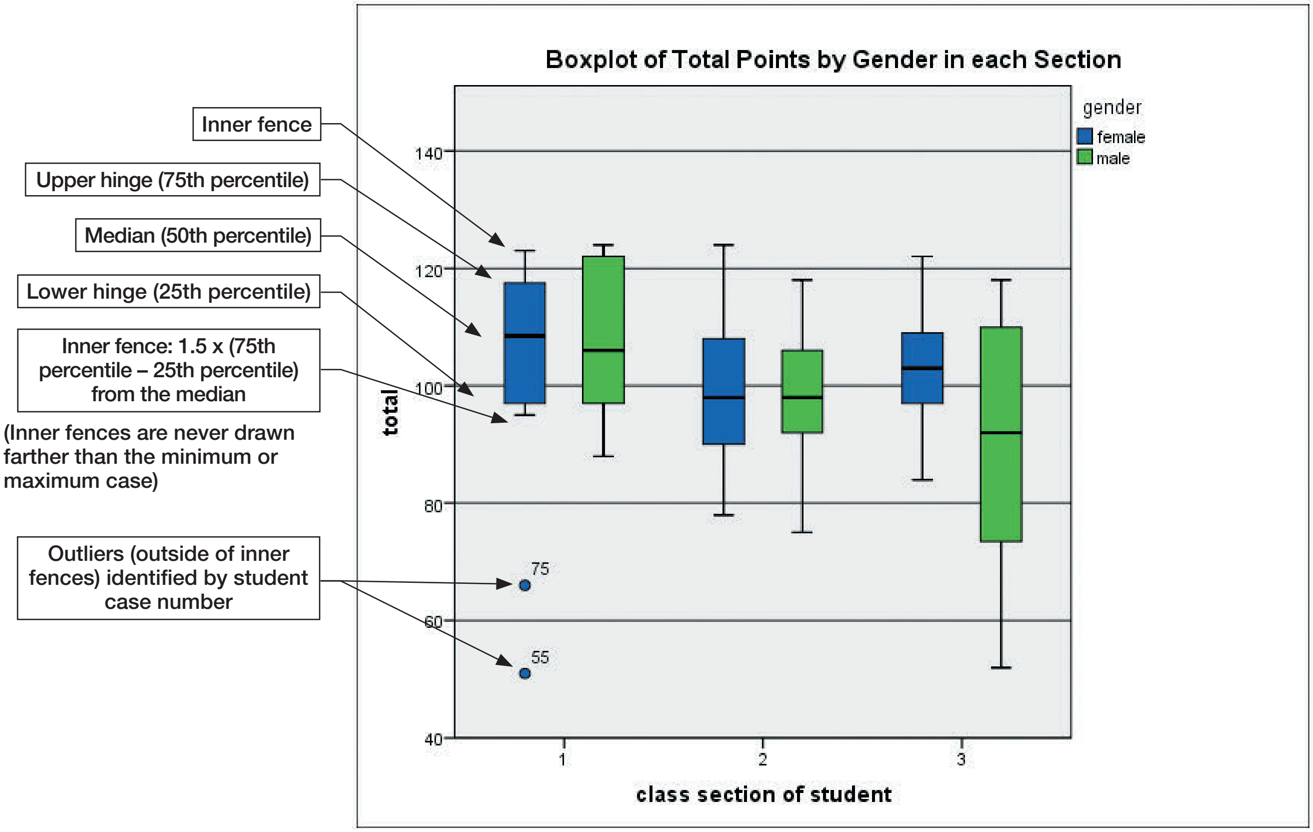
The graph produced by the sequences listed above is displayed below.

Here are some additional editing options:

**1**) Changing the format, font, or size of axis labels or values along each axis (🖱🖱 on the label or value you wish to edit and then edit using the tabbed table to the right);

**2**) Changing the wording of axis labels or values along each axis (🖱 on the label or value you wish to edit, 🖱 a second time and a red cursor appears that allows you to edit the text);

**3**) Change the orientation of the graph: (🖱 if you want the axes switched; 🖱 again if you wish to change back to the original).

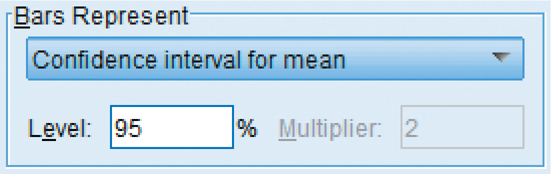


## 5.6.5 Error Bar Charts

Error bars indicate the confidence interval, the standard deviation, or the standard error of the mean. This is useful for visually clarifying which differences are or are not significant.

For Error Bar Charts we choose to go with **Legacy Dialogs** graphs rather than **Chart Builder**. These graphs provide a number of important options not easily available with Chart Builder. We will also not discuss editing options for this very simple chart.

To create an error bar chart, select **Graphs** ⟶ **Legacy Dialogs** ⟶ **Error Bar**. A dialog box somewhat similar to Screen 5.1 appears. Options specific to error bar charts let you specify what type of error bars you want to produce. You can request error bars that are long enough to represent a certain confidence interval of the mean (typically, a 95% confidence interval; this is the default), a certain number of standard deviations around the mean (for example, the error bars could be one or two standard deviations long), or a certain number of standard errors around the mean.

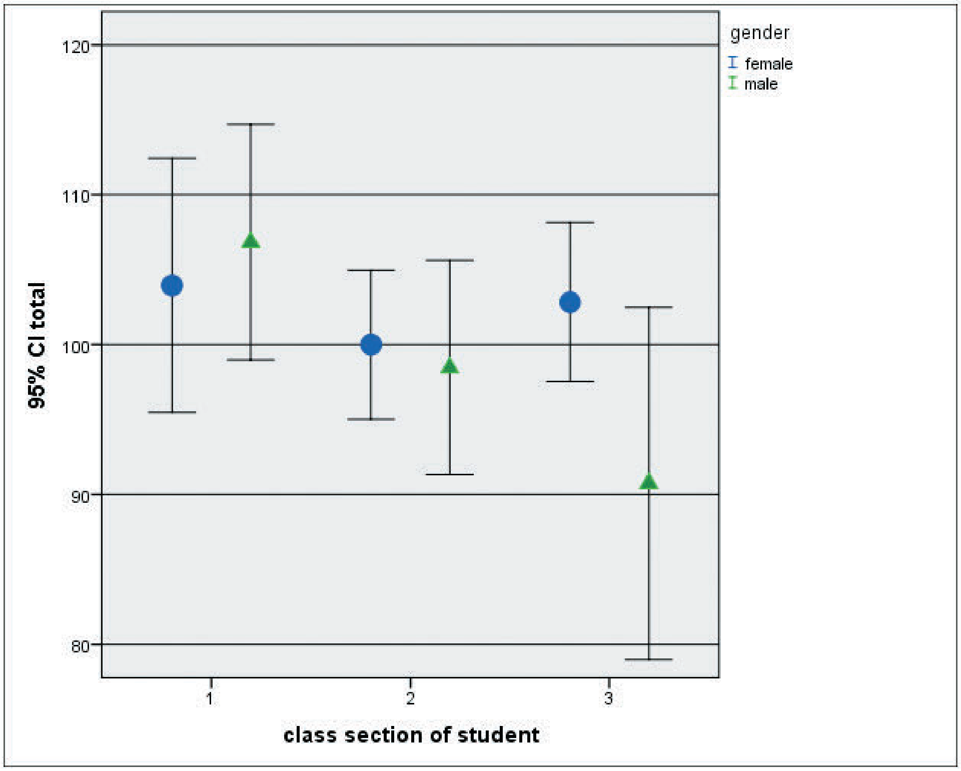


Bars Represent
Confidence interval for mean ▾
Level: 95 % Multiplier: 2

To select a different confidence interval, select the "95.0" and type a new confidence interval. To have SPSS produce error bars based on a certain number of standard deviations or standard errors of the mean, click on the ▾, choose the type of error bar you want, select the "2.0" that will appear to the right of **Multiplier**, and type the number of standard deviations or confidence intervals you desire.

*The sequence step below produces a graph with total points on the Y-axis, class section on the X-axis, and separate 95% confidence interval (95% is the default; you can select different percentages if you wish) error bars for females and males.*

| In Screen | Do This | Step 5d |
|---|---|---|
| **Menu** | ✳ **Graphs** → **Legacy Dialog** → ✳ **Error Bar** | |
| | ✳ **Clustered** → ✳ **Define** | |
| | ✳ **total** → ✳ *top* ➡ → ✳ **section** → ✳ *second* ➡ → ✳ **gender** → ✳ | |
| | *third* ➡ → ✳ **OK** | **Back1** |



This is the graph (previous page) produced (with minor editing) by sequence Step 5d. Notice that men are represented by green triangular markers (positioned at the mean for each group) with solid lines showing the range for 95% confidence intervals. Women are represented by blue round markers (also at the mean) with solid lines showing the 95% confidence interval range. Additional editing options are available such as Titles, subtitles, and footnotes.

### 5.6.6 Histograms

Histograms look similar to bar graphs but are used more often to indicate the number of subjects or cases in ranges of values for a continuous variable, such as the number of students who scored between 90 and 100, between 80 and 89, between 70 and 79, and so forth for a final class percentage. To create a histogram, select **Graphs** → **Chart Builder** → **Histogram**. A dialog box very similar to Screen 5.2 appears. There is one frequently desired option on this initial screen. In the **Element Properties** dialog box to the right is the **Display normal curve** option. If this box is checked, a line representing a normal curve with the mean and standard deviation of your data will be drawn in front of the histogram bars. This makes it possible to easily examine how normally your data are distributed.

*The sequence step below produces a graph with total points* (**total**) *on the X-axis and provides a graphical display of how many subjects scored in each range of values.*

| In Screen | Do This | Step 5e |
|---|---|---|
| 5.1 | ✱ **Graphs** → ✱ **Chart Builder** → ✱ **Histogram** | |
| 5.2 | ✱ ▥ → *drag to* **Chart preview** *box (above)* → ✱ **Display normal curve** | |
| | → ✱ **Apply** → ✱ **total** → *drag to* **X-Axis?** *box* → ✱ **OK** | Back1 |

**Creating a usable graph**: The original histogram generated by SPSS is almost meaningless due to problems cited in the "Bugs" section. To create meaningful scales (in this case 5-point intervals) and frequencies (for each bar), conduct the following sequence of operations:

✱ ✱ on graph → ✱ ✱ on any bar in the graph → ✱ **Custom** in box to the right → ✱ **Interval width** → ✱ inside **Interval width** box → **type** 5 → ✱ **Apply** → ✱ on any number on the X axis → ✱ **Scale** tab in box to the right → ✱ ✱ inside **Major increment** box → **type** 10 → ✱ ✱ inside **Maximum** box → **type** 130 → ✱ **Apply** → ✱ ☒ to return

**Adding grid lines**: To enhance clarity we may wish to add grid lines to the chart. The following sequence will place horizontal grid lines at 5-point intervals.

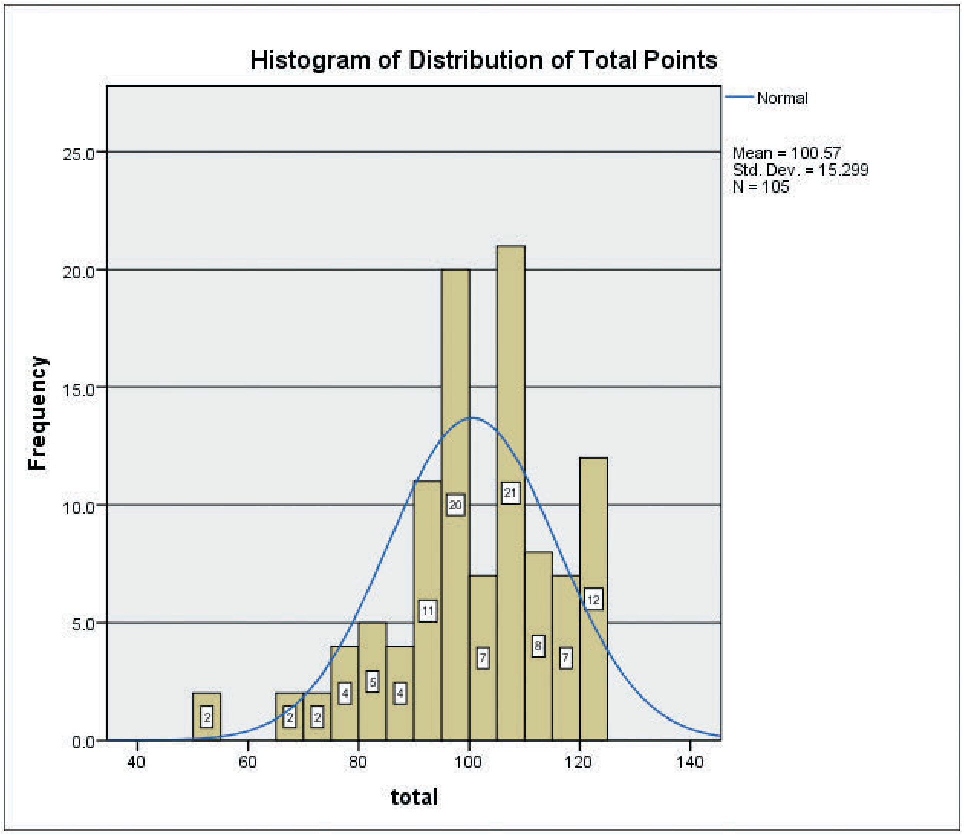
✱ ✱ on graph → ✱ ✱ y-axis values → ✱ ▦ icon (above) → ✱ ☒ to return

**Adding labels inside the bars**: We may wish to include the actual values (number of subjects who scored in each range of total points) inside the bars. The following sequence will provide the frequency inside each bar.

✱ ✱ on graph → ✱ ▥ icon → ✱ ☒ to return

**Adding a Title and Changing the Font**: The following sequence will create a title for the Chart and then change the font size to 14.

✱ ✱ on graph → ✱ ▭ icon → highlight **Title** (using the red cursor line) → **type** Histogram of Distribution of Total Points → press **Enter** → ✱ on the (just created) title → ✱ ▾ next to **Preferred Size** → ✱ 14 → ✱ **Apply** → ✱ ☒ to return



The above graph reflects all the edits described above. A number of different editing options are available, such as:

**1**) changing the format, font, or size of axis labels or values along each axis (✱✱ on the label or value you wish to edit and then edit using the tabbed table above);

**2**) changing the wording of axis labels or values along each axis (✱ on the label or value you wish to edit, ✱ a second time and a red cursor appears that allows you to edit the text); or

**3**) change the orientation of the graph (✱ ▤ if you want the bars to be horizontal; ✱ ▤ again if you wish to change orientation back to vertical).

## 5.6.7 Scatterplots

Scatterplots are a popular way of displaying the nature of correlations between variables. One dot is drawn for each case in the data file; both the X- and Y-axis variables are scale variables. To create a scatterplot, select **Graphs** → **Chart Builder** → **Scatter/Dot**. Once you select and drag the desired graph type to the box above, a dialog box very much like Screen 5.2 appears where you can select which variables you want plotted on the X- and Y-axes.

*The sequence step below produces a scatterplot with total points on the Y-axis, and previous GPA on the X-axis. Different shapes will be used for dots representing females and males.*

| In Screen | Do This | Step 5f |
|---|---|---|
| 5.1 | ✱ Graphs → ✱ Chart Builder → ✱ Scatter/Dot | |
| 5.2 | ✱ ⋰ → *drag to* **Chart preview** *box (above)* → ✱ **total** → *drag to* **Y-Axis?** *box* → ✱ **gpa** → *drag to* **X-Axis?** *box* → ✱ **gender** → *drag to* **Set color** *box* → ✱ **OK** | **Back1** |

The graph appears on the Output screen, ready to edit or print. Certain editing options follow:

**Changing the style and color of markers for men and women**: Since there are two different sets of dots (one set for men, one for women) plotted on the same chart, it is urgent that we create markers that can easily be distinguished one from the other.

✱✱ on graph → ✱✱ **O** next to **female** (in the legend) → ✱ ▾ next to **Type** → ✱ △ → ✱ ⧄ Fill icon → ✱ medium gray → ✱ **Apply** → ✱ **O** next to **male** → ✱ ⧄ Fill icon → ✱ black → ✱ **Apply** → ✱ ☒ to return

**Inserting regression lines for both men and women**: Since we have two sets of points plotted, it will be interesting to compare the regression lines. A steeper line suggests a stronger association between **gpa** and **total**. Our efforts suggest that it is possible to change the style of each regression line, but not the color.

✱✱ on graph → ✱ ⋰ icon → ✱ ☒ to return

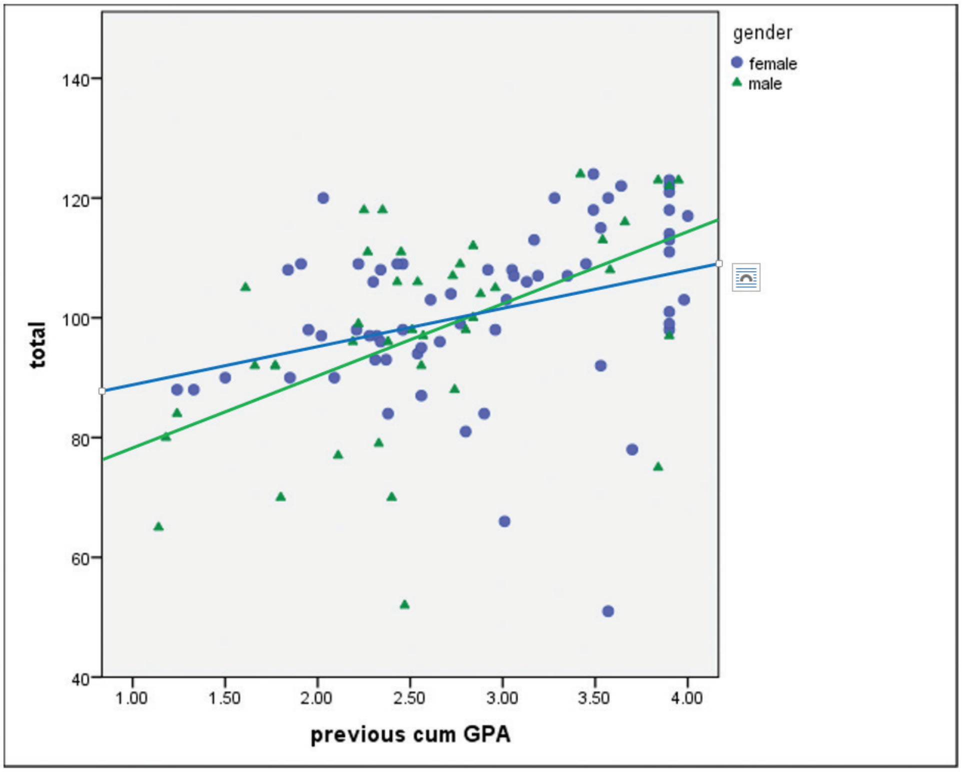
**Adding a Title and Changing the Font**: The following sequence will create a title for the Chart and then change the font size to 14.

✱✱ on graph → ✱ ⊏ icon → highlight **Title** (using the red cursor line) → **type** Correlation between Total Points and GPA for Men and Women → press **Enter** → ✱ on the (just created) title → ✱ ▾ next to **Preferred Size** → ✱ 14 → ✱ **Apply** → ✱ ☒ to return

The graph (on the following page) is the result of Step 5f and all the edits that follow. The markers for men and women are sufficiently different that they are easy to distinguish. The regression lines suggest that the association between total points and GPA is stronger for men than it is for women. Other options:

**1**) Changing the format, font, or size of axis labels or values along each axis (✱✱ on the label or value you wish to edit and then edit using the tabbed table to the right);

**2**) Changing the wording of axis labels or values along axes (✱ on the label or value you wish to edit, ✱ a second time and a red cursor appears that allows you to edit the text).



## 5.7 Printing Results

Results of the analysis (or analyses) that have just been conducted require a window that displays the standard commands (**File Edit Data Transform Analyze**. . .) across the top. A typical print procedure is shown below beginning with the standard output screen (Screen 1, inside back cover).

*To print results, from the Output screen perform the following sequence of steps:*

| In Screen | Do This | Step 6 |
|---|---|---|
| Back1 | *Select desired output or edit (see pages 16–22)* → ✳ **File** → ✳ **Print** | |
| 2.9 | *Consider and select desired print options offered, then* → ✳ **OK** | |

*To exit you may begin from any screen that shows the* **File** *command at the top.*

| In Screen | Do This | Step 7 |
|---|---|---|
| Menu | ✳ **File** → ✳ **Exit** | 2.1 |

Note: After clicking **Exit**, there will frequently be small windows that appear asking if you wish to save or change anything. Simply click each appropriate response.

# Exercises

Answers to selected exercises are downloadable at **www.spss-step-by-step.net**.

All of the following exercises use the **grades.sav** sample data file.

1. Using a bar chart, examine the number of students in each section of the class along with whether or not students attended the review session. Does there appear to be a relation between these variables?
2. Using a line graph, examine the relationship between attending the **review** session and **section** on the **final** exam score. What does this relationship look like?
3. Create a boxplot of **quiz 1** scores. What does this tell you about the distribution of the quiz scores? Create a boxplot of **quiz 2** scores. How does the distribution of this quiz differ from the distribution of quiz 1? Which case number is the outlier?
4. Create an error bar graph highlighting the 95% confidence interval of the mean for each of the three **sections**' **final** exam scores. What does this mean?
5. Based on the examination of a histogram, does it appear that students' previous GPAs are normally distributed?
6. Create the scatterplot described in Step 5f (page 98). What does the relationship appear to be between **gpa** and academic performance (**total**)? Add regression lines for both men and women to this scatterplot. What do these regression lines tell you?
7. By following all steps on pages 88 and 89, reproduce the bar graph shown on page 89.
8. By following all steps on pages 90 and 91, reproduce the line graph shown on page 91.
9. By following all steps on pages 92, reproduce the pie chart shown on page 92.
10. By following all steps on pages 93 and 94, reproduce the Boxplot shown on page 94.
11. By following all steps page 95, reproduce the Error Bar Chart shown on page 95. Note that the edits are not specified on page 95. See if you can perform the edits that produce an identical chart.
12. By following all steps on pages 96 and 97, reproduce the histogram shown on page 97.
13. By following all steps on page 98, reproduce the scatterplot shown on page 99.

# Chapter 6
# Frequencies

THIS CHAPTER deals with frequencies, graphical representation of frequencies (bar charts and pie charts), histograms, and percentiles. Each of these procedures is described below. **Frequencies** is one of the SPSS commands in which it is possible to access certain graphs directly (specifically, bar charts, pie charts, and histograms) rather than accessing them through the **Graphs** command. More information about editing these graphs is treated in some detail in Chapter 5. Bar charts or pie charts are typically used to show the number of cases ("frequencies") in different categories. As such they clearly belong in a chapter on frequencies. Inclusion of histograms and percentiles seems a bit odd because they are most often used with a continuous distribution of values and are rarely used with categorical data. They are included here because the **Frequencies** command in SPSS is configured in such a way that, in addition to frequency information, you can also access histograms for continuous variables, certain descriptive information, and percentiles. The **Descriptives** command and descriptive statistics are described in Chapter 7; however, that procedure does not allow access to histograms or percentiles.

## 6.1 Frequencies

Frequencies is one of the simplest yet one of the most useful of all SPSS procedures. The **Frequencies** command simply sums the number of instances within a particular category: There were 56 males and 37 females. There were 16 Whites, 7 Blacks, 14 Hispanics, 19 Asians, and 5 others. There were 13 A's, 29 B's, 37 C's, 7 D's, and 3 F's. Using the **Frequencies** command, SPSS will list the following information: value labels, the value code (the number associated with each level of a variable, e.g., female = 1, male = 2), the frequency, the percent of total for each value, the valid percent (percent after missing values are excluded), and the cumulative percent. These are each illustrated and described in the Output section.

## 6.2 Bar Charts

The **Bar chart(s)** option is used to create a visual display of frequency information. A bar chart should be used only for categorical (not continuous) data. The gender, ethnicity, and grade variables listed in the previous paragraph represent categorical data. Each of these variables divides the data set into distinct categories such as male, female; A, B, C, D, F; and others. These variables can be appropriately displayed in a bar chart. Continuous data contain a series of numbers or values such as scores on the final exam, total points, finishing times in a road race, weight in pounds of individuals in your class, and so forth. Continuous variables are typically represented graphically with histograms, our next topic.

## 6.3 Histograms

For continuous data, the **Histograms** option will create the appropriate visual display. A histogram is used to indicate frequencies of a *range* of values. A histogram is used when the number of instances of a variable is too large to want to list all of them. A good example is the breakdown of the final point totals in a class of students. Since it would be too cumbersome to list *all* scores on a graph, it is more practical to list the number of subjects within a *range* of values, such as how many students scored between 60 and 69 points, between 70 and 79 points, and so forth.

## 6.4 Percentiles

The **Percentile Values** option will compute any desired percentiles for continuous data. Percentiles are used to indicate what percent of a distribution lies below (and above) a particular value. For instance if a score of 111 was at the 75th percentile, this would mean that 75% of values are lower than 111 and 25% of values are higher than 111. Percentiles are used extensively in educational and psychological measurement.

The file we use to illustrate frequencies, bar charts, histograms, and percentiles (pie charts are so intuitive we do not present them here) is the example described in the first chapter. The file is called **grades.sav** and has an $N = 105$. This analysis computes frequencies, bar charts, histograms, and percentiles utilizing the **gender**, **ethnic**, **grade**, and **total** variables.

## 6.5 Step by Step

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps*:

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | → All Programs → IBM SPSS Statistics → IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons*:

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | → → | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| | Step 2 |
|---|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3)* | |

*Screens 1 and 2 (displayed on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **grades.sav** *file for further analyses*:

| In Screen | Do This | Step 3 |
|---|---|---|
| Front1 | **File** → **Open** → **Data** [*or* ] | |
| Front2 | **type** `grades.sav` → **Open** [*or* **grades.sav**] | Data |

Whether first entering SPSS or returning from earlier operations the standard menu of commands across the top is required. As long as it is visible you may perform any analyses. It is not necessary for the data window to be visible.

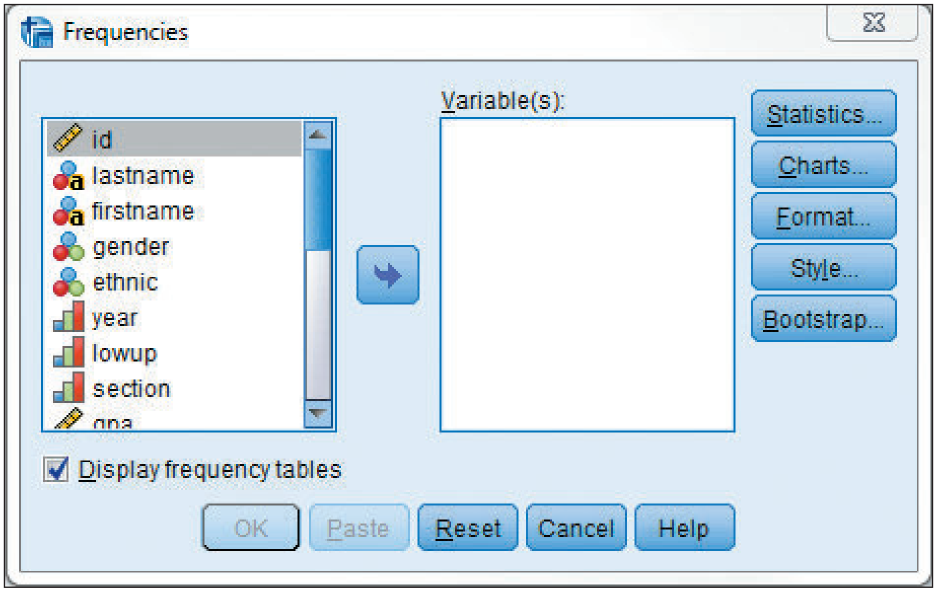
*After completion of Step 3 a screen with the desired menu bar appears. When you click a command (from the menu bar), a series of options will appear (usually) below the selected command. With each new set of options, click the desired item. The sequence to access frequencies begins at any screen with the menu of commands visible*:

| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | ✳ **Analyze** → **Descriptive Statistics** → ✳ **Frequencies** | 6.1 |

## 6.5.1 Frequencies

A screen now appears (below) that allows you to select variables for which you wish to compute frequencies. The procedure involves clicking the desired variable name in the box to the left and then pasting it into the **Variables(s)** (or "active") box to the right by clicking the right arrow (➡) in the middle of the screen. If the desired variable is not visible, use the scroll bar arrows (▲ ▼) to bring it to view. To deselect a variable (to move it from the **Variable(s)** box back to the original list), click on the variable in the active box and the ➡ in the center will become a ⬅. Click on the left arrow to move the variable back. To clear all variables from the **Variable(s)** box, click the **Reset** button.

**Screen 6.1** The Frequencies window



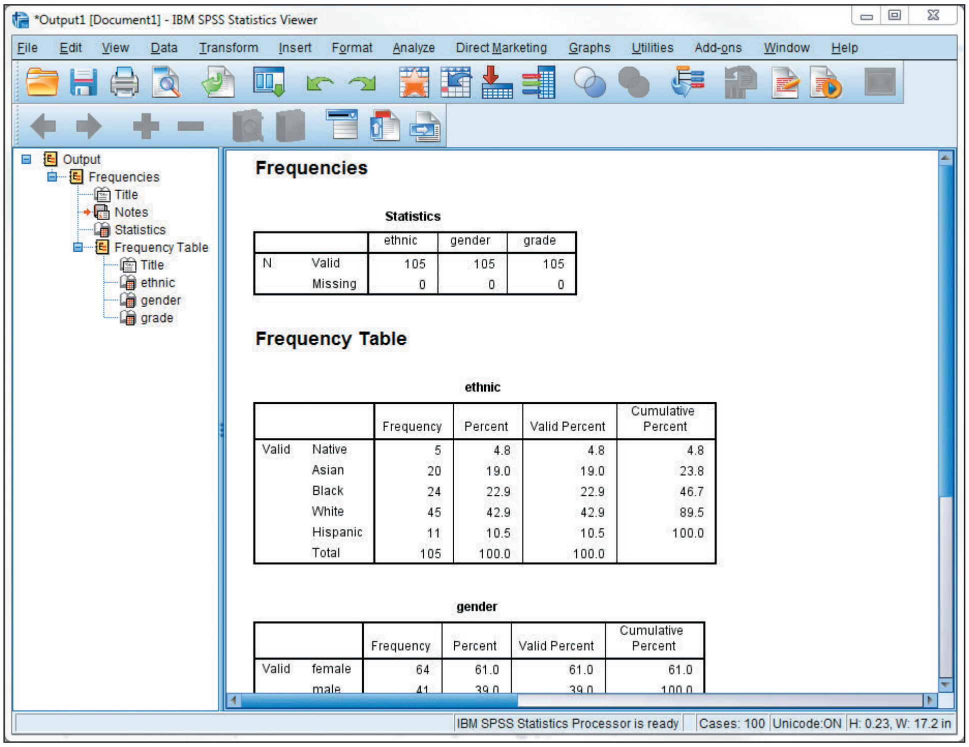
*The following sequence of steps will allow you to compute frequencies for the variables* **ethnic**, **gender**, *and* **grade**.

| In Screen | Do This | Step 5 |
|---|---|---|
| 6.1 | *In the box to the left* ✳ **ethnic** → ✳ ➡ → ✳ **gender** → ✳ ➡ → ✳ **grade** → ✳ ➡ → ✳ **OK** | 6.2 |

You have now selected the three variables associated with gender, ethnicity, and grades. By clicking the **OK** button, SPSS proceeds to compute frequencies. After a few moments the output will be displayed on the screen. The Output screen will appear every time an analysis is conducted (labeled Screen 6.2), and appears on the following page.

The results are now located in a window with the title **Output# [Document#] – IBM SPSS Statistics Viewer** at the top. To view the results, make use of the up and down arrows on the scroll bar (▲▼). Partial results from the procedure described above are found in the Output section. More complete information about output screens, editing output, and pivot charts are included in Chapter 2 (pages 17–22, 34–39 for Mac users). If you wish to conduct further analyses with the same data set, the starting point is again Screen 6.1. Perform whichever of Steps 1–4 (usually Step 4 is all that is necessary) are needed to arrive at this screen.

**Screen 6.2** SPSS Output Viewer



**Frequencies**

**Statistics**

| | | ethnic | gender | grade |
|---|---|---|---|---|
| N | Valid | 105 | 105 | 105 |
| | Missing | 0 | 0 | 0 |

**Frequency Table**

**ethnic**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Native | 5 | 4.8 | 4.8 | 4.8 |
| | Asian | 20 | 19.0 | 19.0 | 23.8 |
| | Black | 24 | 22.9 | 22.9 | 46.7 |
| | White | 45 | 42.9 | 42.9 | 89.5 |
| | Hispanic | 11 | 10.5 | 10.5 | 100.0 |
| | Total | 105 | 100.0 | 100.0 | |

**gender**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | female | 64 | 61.0 | 61.0 | 61.0 |
| | male | 41 | 39.0 | 39.0 | 100.0 |

## 6.5.2 Bar Charts

To create bar charts of categorical data, the process is identical to sequence Step 5 (in previous page), except that instead of clicking the final OK, you will click the **Charts** option (see Screen 6.1). At this point a new screen (Screen 6.3, below) appears: **Bar charts**, **Pie charts**, and **Histograms** are the types of charts offered. For categorical data you will usually choose **Bar charts**. You may choose **Frequencies** (the number of instances within each category) or **Percentages** (the percent of total for each category). After you click **Continue**, the Charts box disappears leaving Screen 6.1. A click on **OK** completes the procedure.

**Screen 6.3** The Frequencies: Charts window

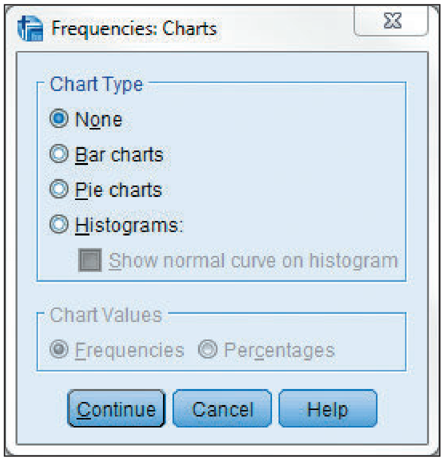


Chart Type

● None

○ Bar charts

○ Pie charts

○ Histograms:

□ Show normal curve on histogram

Chart Values

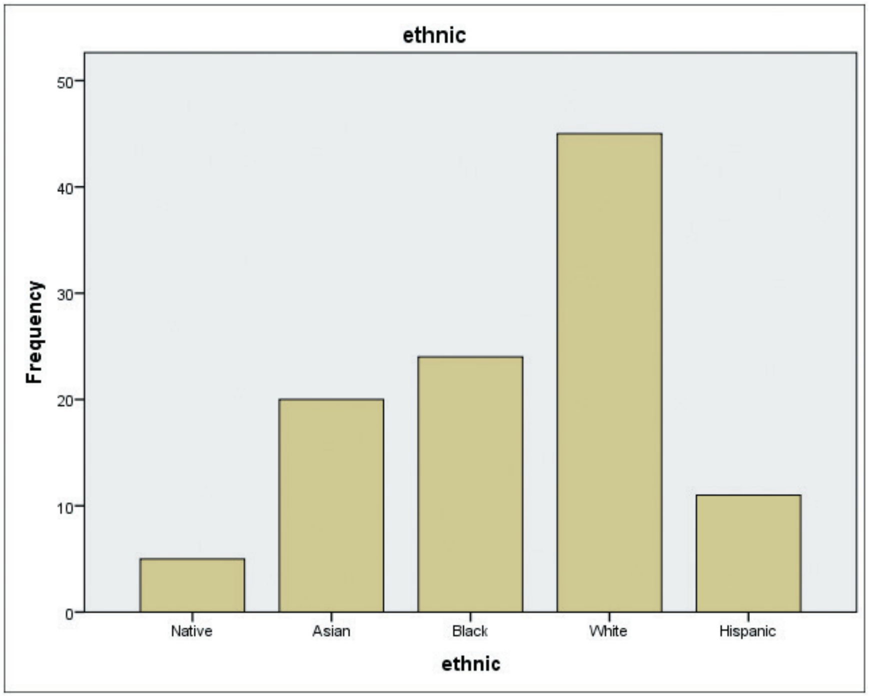
● Frequencies ○ Percentages

*Screen 6.1 is the starting point for this procedure. Notice that we demonstrated a double click of the variable name to paste it into the active box (rather than a click on the* ➡ *button).*

| In Screen | Do This | Step 5a |
|---|---|---|
| 6.1 | ✱ ✱ **ethnic** → ✱ ✱ **gender** → ✱ ✱ **grade** → ✱ **Charts** | |
| 6.3 | ✱ **Bar charts** → ✱ **Percentages** *(if you do not want frequencies)*→ ✱ **Continue** | |
| 6.1 | ✱ **OK** | 6.2 |

After a few moments of processing time (several hours if you are working on a typical university network) the output screen will emerge. A total of three bar charts have been created, one describing the ethnic breakdown, another describing the gender breakdown, and a third dealing with grades. To see these three graphs simply requires scrolling down the output page until you arrive at the desired graph. If you wish to edit the graphs for enhanced clarity, double click on the graph and then turn to Chapter 5 to assist you with a number of editing options. The chart that follows (Screen 6.4) shows the bar chart for ethnicity.

**Screen 6.4** A Sample Bar Chart



## 6.5.3 Histograms

Histograms may be accessed in the same way as bar charts. The distinction between bar charts and histograms is that histograms are typically used for display of continuous (not categorical) data. For the variables used above (gender, ethnicity, and grades), histograms would not be appropriate. We will here make use of a histogram to display the distribution for the total points earned by students in the class. Perform the following sequence of steps to create a histogram for total. Refer, if necessary, to Screens 6.1, 6.2, and 6.3 on previous pages for visual reference. The histogram for this procedure is displayed in the Output section.

*This procedure begins at Screen 6.1. Perform whichever of Steps 1–4 (pages 102–103) are necessary to arrive at this screen. You may also need to click the* **Reset** *button before beginning.*

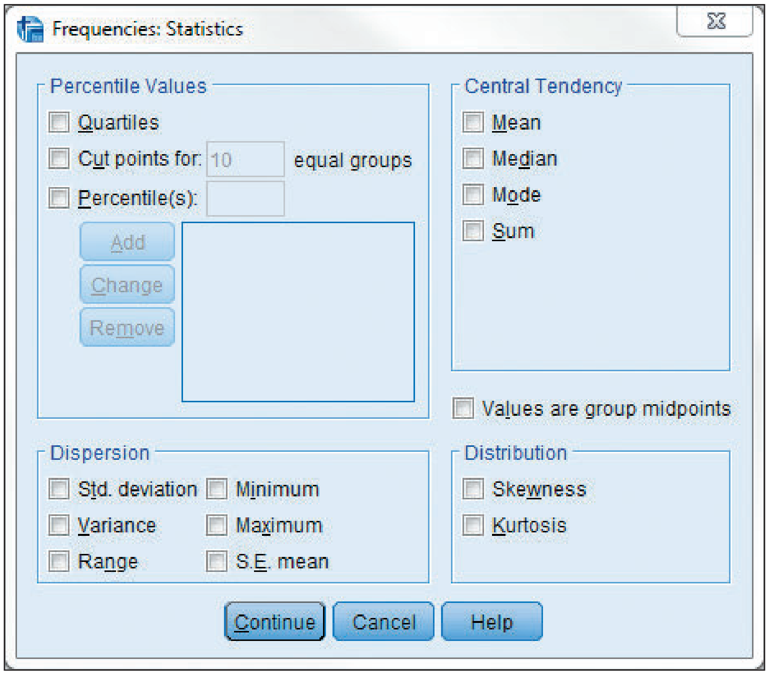
| In Screen | Do This | Step 5b |
|---|---|---|
| 6.1 | 🖱 **total** → 🖱 ➡ → 🖱 **Charts** | |
| 6.3 | 🖱 **Histograms** → 🖱 **Show normal curve on histogram** → 🖱 **Continue** | |
| 6.1 | 🖱 **Display frequency tables** (*so ✓ does not show in box*) → 🖱 **OK** | 6.2 |

Note the step where you click **Display frequency tables** to deselect that option. For categorical data, you will always keep this option since it constitutes the entire non-graphical output. For continuous data (the **total** variable in this case), a display of frequencies would be a list about 70 items, indicating that 1 subject scored 45, 2 subjects scored 47, and so on up to the number of subjects who scored 125. This is rarely desired. If you click this option prior to requesting a histogram, a warning will flash indicating that there will be no output. The **Show normal curve on histogram** allows a normal curve to be superimposed over the histogram.

## 6.5.4 Percentiles and Descriptives

Descriptive statistics are explained in detail in Chapter 7. Using the **Frequencies** command, under the **Statistics** option (see Screen 6.1), descriptive statistics and percentile values are available. When you click on the **Statistics** option, a new screen appears (Screen 6.5, below) that allows access to this additional information. Three different step sequences (on the following page) will explain (a) how to create a histogram and access descriptive data, (b) how to calculate a series of percentiles with equal spacing between each value, and (c) how to access specific numeric percentiles. All three sequences will utilize the **total** points variable.

**Screen 6.5** The Frequencies: Statistics Window



*For any of the three procedures below, the starting point is Screen 6.1. Perform whichever of Steps 1–4 (pages 102–103) are necessary to arrive at this screen. Step 5c gives steps to create a histogram for total points and also requests the mean of the distribution, the standard deviation, the skewness, and the kurtosis. Click the* **Reset** *button before beginning if necessary.*

| In Screen | Do This | Step 5c |
|---|---|---|
| 6.1 | ✳ **total** → ✳ ➡ → ✳ **Charts** | |
| 6.3 | ✳ **Histograms** → ✳ **Continue** | |
| 6.1 | ✳ **Statistics** | |
| 6.5 | ✳ **Mean** → ✳ **Skewness** → ✳ **Kurtosis** → ✳ **Std. deviation** → ✳ **Continue** | |
| 6.1 | ✳ **Display frequency tables** *(so ✓ does not show in box)* → ✳ **OK** | 6.2 |

*To calculate percentiles of the* **total** *variable for every 5th percentile value (e.g., 5th, 10th, 15th, etc.) it's necessary to divide the percentile scale into 20 equal parts. Click* **Reset** *if necessary.*

| In Screen | Do This | Step 5d |
|---|---|---|
| 6.1 | ✳ **total** → ✳ ➡ → ✳ **Statistics** | |
| 6.5 | ✳ **Cut points for □ eq...** → **press** **TAB** → **type** 20 → ✳ **Continue** | |
| 6.1 | ✳ **Display frequency tables** *(deselect)* → ✳ **OK** | 6.2 |

Note that when you type the 20 (or any number) it automatically writes over the default value of 10, already showing.

*Finally, to designate particular percentile values (in this case, 2, 16, 50, 84, 98) perform the following sequence of steps. Click* **Reset** *if necessary.*

| In Screen | Do This | Step 5e |
|---|---|---|
| 6.1 | ✳ **total** → ✳ ➡ → ✳ **Statistics** | |
| 6.5 | ✳ **Percentile(s)** → ✳ *in* **Percentile(s)** *box* → **type** 2 → ✳ **Add** → | |
| | ✳ *in* **Percentile(s)** *box* → **type** 16 → ✳ **Add** → | |
| | ✳ *in* **Percentile(s)** *box* → **type** 50 → ✳ **Add** → | |
| | ✳ *in* **Percentile(s)** *box* → **type** 84 → ✳ **Add** → | |
| | ✳ *in* **Percentile(s)** *box* → **type** 98 → ✳ **Add** → ✳ **Continue** | |
| 6.1 | ✳ **Display frequency tables** *(deselect)* → ✳ **OK** | 6.2 |

Note: Quartile values (the 25th, 50th, and 75th percentiles) may be obtained quickly by clicking the **Quartiles** box (see Screen 6.5), clicking **Continue**, and then clicking **OK** (see Screen 6.1).

## 6.6 Printing Results

Results of the analysis (or analyses) that have just been conducted require a window that displays the standard commands (**File Edit Data Transform Analyze . . .**) across the top. A typical print procedure is shown below beginning with the standard output screen (Screen 1, inside back cover).

*To print results, from the Output screen perform the following sequence of steps:*

| In Screen | Do This | Step 6 |
|---|---|---|
| Back1 | *Select desired output or edit (see pages 16–22)* ⟶ ✳ **File** ⟶ ✳ **Print** | |
| 2.9 | *Consider and select desired print options offered, then* ⟶ ✳ **OK** | |

*To exit you may begin from any screen that shows the* **File** *command at the top.*

| In Screen | Do This | Step 7 |
|---|---|---|
| Menu | ✳ **File** ⟶ ✳ **Exit** | 2.1 |

Note: After clicking **Exit**, there will frequently be small windows that appear asking if you wish to save or change anything. Simply click each appropriate response.

## 6.7 Output

### Frequencies, Histograms, Descriptives, and Percentiles

In the output, due to space constraints, we often present results of analyses in a more space-conserving format than is typically done by SPSS. We use identical terminology as that used in the SPSS output and hope that minor formatting differences do not detract from understanding.

### 6.7.1 Frequencies

What follows is partial results (and a slightly different format) from sequence Step 5, page 103.

| Variable | Value Label | Frequency | Percent | Valid Percent | Cum Percent |
|---|---|---|---|---|---|
| ETHNIC | Native | 5 | 4.8 | 4.8 | 4.8 |
| | Asian | 20 | 19.0 | 19.0 | 23.8 |
| | Black | 24 | 22.9 | 22.9 | 46.7 |
| | White | 45 | 42.9 | 42.9 | 89.5 |
| | Hispanic | 11 | 10.5 | 10.5 | 100.0 |
| | TOTAL | 105 | 100.0 | 100.0 | |
| GENDER | Female | 64 | 61.0 | 61.0 | 61.0 |
| | Male | 41 | 39.0 | 39.0 | 100.0 |
| | TOTAL | 105 | 100.0 | 100.0 | |

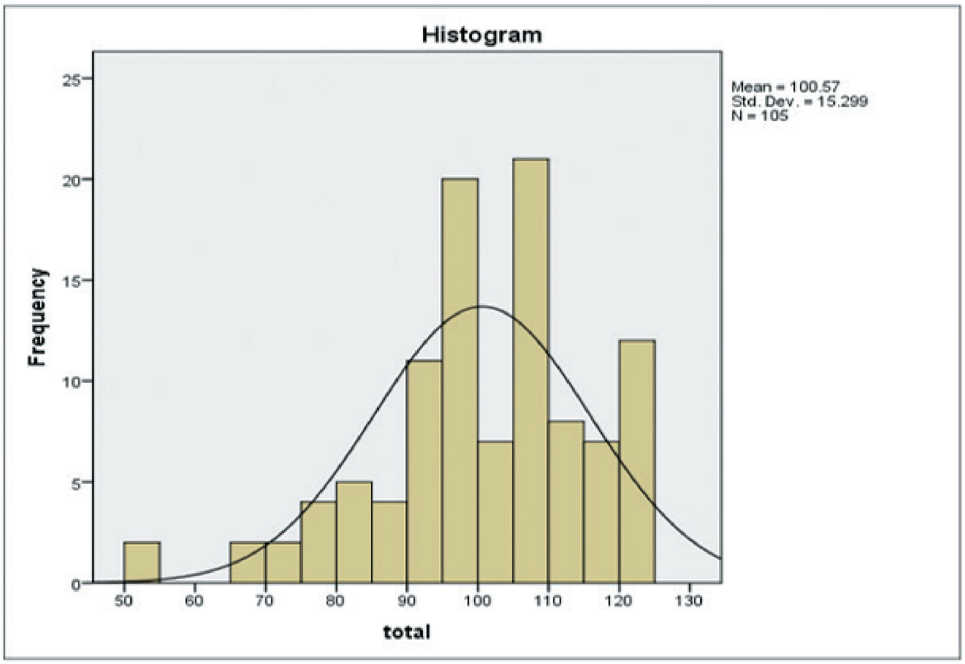
The number of subjects in each category is self-explanatory. Definitions of other terms follow:

| Term | Definition/Description |
|---|---|
| **Value label** | Names for levels of a variable. |
| **Value** | The number associated with each level of the variable (just in front of each label). |
| **Frequency** | Number of data points for a variable or level. |
| **Percent** | The percent for each component part, including missing values. If there were missing values, they would be listed in the last row as **missing** along with the frequency and percent of missing values. The total would still sum to 100.0%. |
| **Valid percent** | Percent of each value excluding missing values. |
| **Cum percent** | Cumulative percentage of the **Valid percent.** |

## 6.7.2 Histograms

What follows is output from sequence Step 5b on page 106. To produce an identical graph you will need to perform the edits described on pages 96–97.

Note that on the horizontal axis (graph below) the border values of each of the bars are indicated. This makes for clear interpretation since it is easy to identify that, for instance, 11 students scored between 90 and 95 points, 20 students scored between 95 and 100 points, and 8 students scored between 100 and 105 points. The graph has been edited to create the 5-point increments for bars. For creation of an identical graph several of the editing options would need to be applied. Please see Chapter 5 to assist you with this. A normal curve is superimposed on the graph due to selecting the **Show normal curve on histogram** option.



## 6.7.3 Descriptives and Percentiles

*What follows is complete output (slightly different format) from sequence Step 5d on page* 107.

**Descriptives**

| Variable | Mean | Std Deviation | Kurtosis | S E Kurtosis | Skewness | S E Skewness |
|---|---|---|---|---|---|---|
| TOTAL | 100.571 | 15.299 | .943 | .467 | –.837 | .236 |

### Percentiles

| Percentile | Value | Percentile | Value | Percentile | Value | Percentile | Value |
|---|---|---|---|---|---|---|---|
| 5.00 | 70.00 | 30.00 | 95.80 | 55.00 | 105.00 | 80.00 | 113.00 |
| 10.00 | 79.60 | 35.00 | 97.00 | 60.00 | 106.60 | 85.00 | 118.00 |
| 15.00 | 86.70 | 40.00 | 98.00 | 65.00 | 108.00 | 90.00 | 120.00 |
| 20.00 | 90.00 | 45.00 | 98.70 | 70.00 | 109.00 | 95.00 | 122.70 |
| 25.00 | 92.00 | 50.00 | 103.00 | 75.00 | 111.00 | | |

For Percentiles: For the **total** points variable, 5% of values fall below 70 points and 95% of values are higher than 70 points; 10% of values fall below 79.6 points and 90% are higher, and so forth.

Descriptive information is covered in Chapter 7 so we will not discuss those terms here. Note that when the skewness and kurtosis are requested, the standard errors of those two measures are also included.

# Exercises

Answers to selected exercises can be downloaded at **www.spss-step-by-step.net**.

Notice that data files other than the **grades.sav** file are being used here. Please refer to the **Data Files** section starting on page 364 to acquire all necessary information about these files and the meaning of the variables. As a reminder, all data files are downloadable from the web address shown above.

1. Using the **divorce.sav** file display frequencies for **sex**, **ethnic**, and **status**. Print output to show frequencies for all three; edit output so it fits on one page. On a second page, include three bar graphs of these data and provide labels to clarify what each one means.
2. Using the **graduate.sav** file display frequencies for **motive**, **stable**, and **hostile**. Print output to show frequencies for all three; edit output so it fits on one page. Note: This type of procedure is typically done to check for accuracy of data. Motivation (**motive**), emotional stability (**stable**), and hostility (**hostile**) are scored on 1- to 9-point scales. You are checking to see if you have, by mistake, entered any 0s or 99s.
3. Using the **helping3.sav** file compute percentiles for **thelplnz** (time helping, measured in z scores) and **tqualitz** (quality of help measured in z scores). Use percentile values 2, 16, 50, 84, 98. Print output and circle values associated with percentiles for **thelplnz**; box percentile values for **tqualitz**. Edit output so it fits on one page.
4. Using the **helping3.sav** file compute percentiles for **age**. Compute every 10th percentile (10, 20, 30, etc.). Edit (if necessary) to fit on one page.
5. Using the **graduate.sav** file display frequencies for **gpa**, **areagpa**, and **grequant**. Compute quartiles for these three variables. Edit (if necessary) to fit on one page.
6. Using the **grades.sav** file create a histogram for **final**. Include the normal curve option. Create a title for the graph that makes clear what is being measured. Perform the edits on pages 96–97 so the borders for each bar are clear.

# Chapter 7
# Descriptive Statistics

**Descriptives** is another frequently used SPSS procedure. Descriptive statistics are designed to give you information about the distributions of your variables. Within this broad category are measures of central tendency (**Mean**, **Median**, **Mode**), measures of variability around the mean (**Std deviation** and **Variance**), measures of deviation from normality (**Skewness** and **Kurtosis**), information concerning the spread of the distribution (**Maximum**, **Minimum**, and **Range**), and information about the stability or sampling error of certain measures, including standard error (S.E.) of the mean (**S.E. mean**), S.E. of the kurtosis, and S.E. of the skewness (included by default when skewness and kurtosis are requested). Using the **Descriptives** command, it is possible to access all of these statistics or any subset of them. In this introductory section of the chapter, we begin with a brief description of statistical significance (included in all forms of data analysis) and the normal distribution (because most statistical procedures require normally distributed data). Then each of the statistics identified above is briefly described and illustrated.

## 7.1 Statistical Significance

All procedures in the chapters that follow involve testing the significance of the results of each analysis. Although statistical significance is not employed in the present chapter it was thought desirable to cover the concept of statistical significance (and normal distributions in the section that follows) early in the book.

Significance is typically designated with words such as “significance,“ “statistical significance,“ or “probability.“ The latter word is the source of the letter that represents significance, the letter “p.“ The *p* value identifies the likelihood that a particular outcome may have occurred by chance. For instance, group A may score an average of 37 on a scale of depression while group B scores 41 on the same scale. If a *t* test determines that group A differs from group B at a $p = .01$ level of significance, it may be concluded that there is a 1 in 100 probability that the resulting difference happened by chance, and a 99 in 100 probability that the discrepancy in scores is a reliable finding.

Regardless of the type of analysis the *p* value identifies the likelihood that a particular outcome occurs by chance. A Chi-square analysis identifies whether observed values differ significantly from expected values; a *t* test or ANOVA identifies whether the mean of one group differs significantly from the mean of another group or groups; correlations and regressions identify whether two or more variables are significantly related to each other. In all instances a significance value will be calculated identifying the likelihood that a particular outcome is or is not reliable. Within the context of research in the social sciences, nothing is ever “proved.“ It is demonstrated or supported at a certain level of likelihood or significance. The smaller the *p* value, the greater the likelihood that the findings are valid.

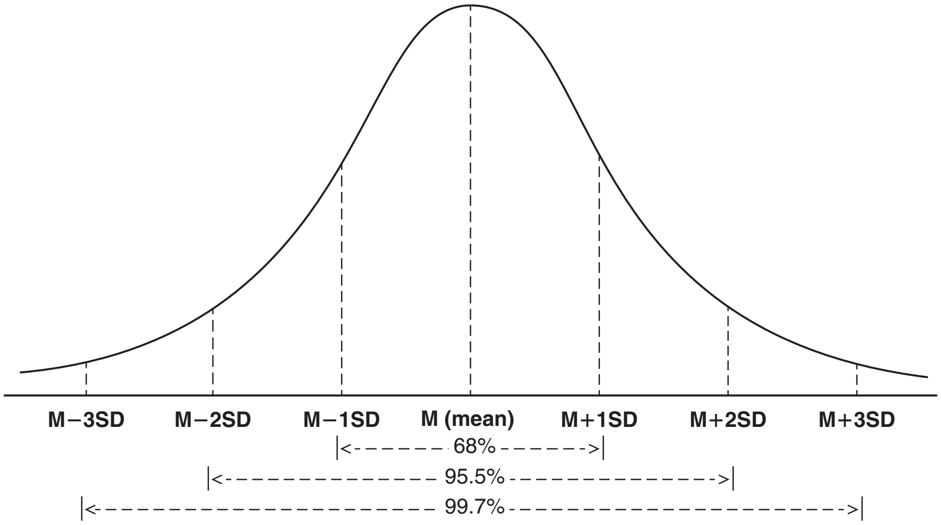
Social scientists have generally accepted that if the *p* value is less than .05 then the result is considered **statistically significant**. Thus, when there is less than a 1 in 20 probability that a certain outcome occurred by chance, then that result is considered statistically significant. Another frequently observed convention is that when a significance level falls between .05 and .10, the result is considered **marginally significant**. When the significance level falls far below .05 (e.g., .001, .0001, etc.) the smaller the value, the greater confidence the researcher has that his or her findings are valid.

When one writes up the findings of a particular study, certain statistical information and *p* values are always included. Whether or not a significant result has occurred is the key focus of most studies that involve statistics.

## 7.2 The Normal Distribution

Many naturally occurring phenomena produce distributions of data that approximate a normal distribution. Some examples include the height of adult humans in the world, the weight of collie dogs, the scoring averages of players in the NBA, and the IQs of residents of the United States. In all of these distributions, there are many mid-range values (e.g., 60–70 inches, 22–28 pounds, 9–14 points, 90–110 IQ points) and few extreme values (e.g., 30 inches, 80 pounds, 60 points, 12 IQ points). There are other distributions that approximate normality but deviate in predictable ways. For instance, times of runners in a 10-kilometer race will have few values less than 30 minutes (none less than 26:17), but many values greater than 40 minutes. The majority of values will lie above the mean (average) value. This is called a *negatively skewed distribution*. Then there is the distribution of ages of persons living in the United States. While there are individuals who are 1 year old and others who are 100 years old, there are far more 1-year-olds, and in general the population has more values below the mean than above the mean. This is called a *positively skewed distribution*. It is possible for distributions to deviate from normality in other ways, some of which are described in this chapter.

A normal distribution is symmetric about the mean or average value. In a normal distribution, 68% of values will lie between plus-or-minus (±) 1 standard deviation (described below) of the mean, 95.5% of values will lie between ± 2 standard deviations of the mean, and 99.7% of values will lie between ± 3 standard deviations of the mean. A normal distribution is illustrated in the figure below.



A final example will complete this section. The average (or mean) height of an American adult male is 69 inches (5' 9") with a standard deviation of 4 inches. Thus, 68% of American men are between 5' 5" and 6' 1" (69 ± 4); 95.5% of American men are between 5' 1" and 6' 5" (69 ± 8), and 99.7% of American men are between 4' 9" and 6'9" (69 ± 12) in height (don't let the NBA fool you!).

## 7.3 Measures of Central Tendency

The **Mean** is the average value of the distribution, or, the sum of all values divided by the number of values. The mean of the distribution [3 5 7 5 6 8 9] is:

$$\mathbf{(3 + 5 + 7 + 5 + 6 + 8 + 9)/7 = \underline{6.14}}$$

The **Median** is the middle value of the distribution. The median of the distribution [3 5 7 5 6 8 9], is 6, the middle value (when reordered from small to large, 3 5 5 **6** 7 8 9). If there is an even number of values in a distribution, then there will be two middle values. In that case the average of those two values is the median.

The **Mode** is the most frequently occurring value. The mode of the distribution [3 **5** 7 **5** 6 8 9] is 5, because 5 occurs most frequently (twice, all other values occur only once).

## 7.4 Measures of Variability Around the Mean

The **Variance** is the sum of squared deviations from the mean divided by $N - 1$. The variance for the distribution [3 5 7 5 6 8 9] (the same numbers used above to illustrate the mean) is:

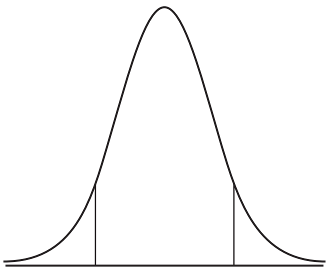
$$\mathbf{[(3–6.14)^2 + (5–6.14)^2 + (7–6.14)^2 + (5–6.14)^2 + (6–6.14)^2 + (8–6.14)^2 + (9–6.14)^2]/6 = \underline{4.1429}}$$

Variance is used mainly for computational purposes. Standard deviation is the more commonly used measure of variability.

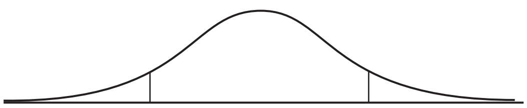
The **Standard deviation** is the positive square root of the variance. For the distribution [3 5 7 5 6 8 9], the standard deviation is the square root of 4.1429, or 2.0354.

## 7.5 Measures of Deviation from Normality

**Kurtosis** is a measure of the “peakedness“ or the “flatness“ of a distribution. A kurtosis value near zero (0) indicates a shape close to normal. A positive value for the kurtosis indicates a distribution more peaked than normal. A negative kurtosis indicates a shape flatter than normal. An extreme negative kurtosis (e.g., < –5.0) indicates a distribution where more of the values are in the tails of the distribution than around the mean. A kurtosis value between ±1.0 is considered excellent for most psychometric purposes, but a value between ±2.0 is in many cases also acceptable, depending on the particular application. Remember that these values are only guidelines. In other settings different criteria may arise, such as significant deviation from normality (outside ±2 × the standard error). Similar rules apply to skewness.

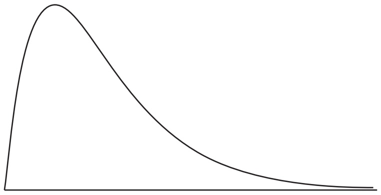


Example of positive kurtosis

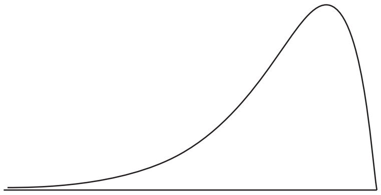


Example of negative kurtosis

**Skewness** measures to what extent a distribution of values deviates from symmetry around the mean. A value of zero (0) represents a symmetric or evenly balanced distribution. A positive skewness indicates a greater number of *smaller* values (sounds backward, but this is correct). A negative skewness indicates a greater number of *larger* values. As with kurtosis, a skewness value between ±1.0 is considered excellent for most psychometric purposes, but a value between ±2.0 is in many cases also acceptable, depending on your application.



Example of positive skewness



Example of negative skewness

## 7.6 Measures for Size of the Distribution

For the distribution [3 5 7 5 6 8 9], the **Maximum** value is 9, the **Minimum** value is 3, and the **Range** is 9 – 3 = 6. The **Sum** of the scores is 3 + 5 + 7 + 5 + 6 + 8 + 9 = 43.

## 7.7 Measures of Stability: Standard Error

SPSS computes the **Standard errors** for the mean, the kurtosis, and the skewness. Standard error is designed to be a measure of stability or of sampling error. The logic behind standard error is this: If you take a random sample from a population, you can compute the mean, a single number. If you take another sample of the same size from the same population you can again compute the mean—a number likely to be slightly different from the first number. If you collect many such samples, the standard error of the mean is the standard deviation of this sampling distribution of means. A similar logic is behind the computation of standard error for kurtosis or skewness. A small value (what is "small" depends on the nature of your distribution) indicates *greater* stability or *smaller* sampling error.

The file we use to illustrate the **Descriptives** command is our example described in the first chapter. The data file is called **grades.sav** and has an *N* = 105. This analysis computes descriptive statistics for variables **gpa**, **total**, **final**, and **percent**.

## 7.8 Step by Step

### 7.8.1 Descriptives

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps*:

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | All Programs → IBM SPSS Statistics → IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons*:

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | → → | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| Step 2 |
|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* |

*Screens 1 and 2 (displayed on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **grades.sav** *file for further analyses*:

| In Screen | Do This | Step 3 |
|---|---|---|
| Front1 | ✳ File → Open → ✳ Data [*or* ✳ 📂] | |
| Front2 | type `grades.sav` → ✳ Open [*or* ✳ ✳ grades.sav] | Data |

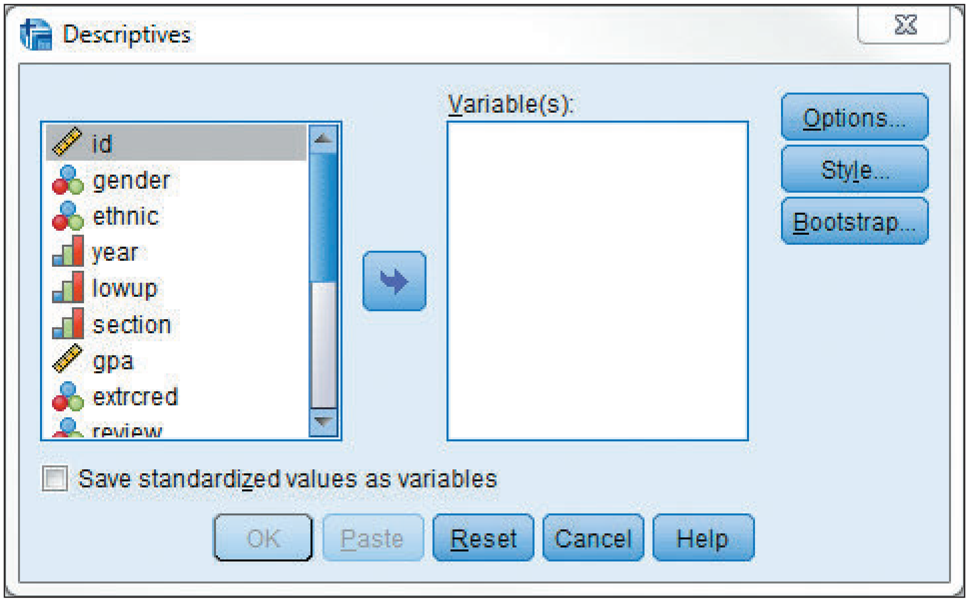
Whether first entering SPSS or returning from earlier operations the standard menu of commands across the top is required. As long as it is visible you may perform any analyses. It is not necessary for the data window to be visible.

*After completion of Step 3 a screen with the desired menu bar appears. When you click a command (from the menu bar), a series of options will appear (usually) below the selected command. With each new set of options, click the desired item. The sequence to access Descriptive Statistics begins at any screen with the menu of commands visible*:

| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | ✳ Analyze → Descriptive Statistics → ✳ Descriptives | 7.1 |

A new screen now appears (below) that allows you to select variables for which you wish to compute descriptives. The procedure involves clicking the desired variable name in the box to the left and then pasting it into the **Variable(s)** (or "active") box to the right by clicking the right arrow (➡) in the middle of the screen. If the desired variable is not visible, use the scroll bar arrows (▲ ▼) to bring it to view. To **de**select a variable (that is, to move it from the **Variable(s)** box back to the original list), click on the variable in the active box and the ➡ in the center will become a ⬅. Click on the left arrow to move the variable back. To clear all variables from the active box, click the **Reset** button.

Screen 7.1 The Descriptives Window



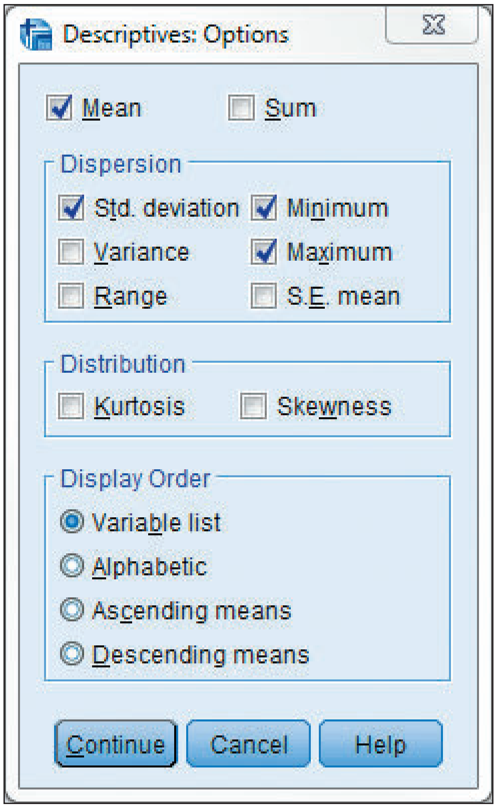
The only check box on the initial screen, **Save standardized values as variables**, will convert all designated variables (those in the **Variable(s)** box) to *z* scores. The original variables will remain, but new variables with a "z" attached to the front will be included in the list of variables. For instance, if you click the **Save standardized values as variables** option, and the variable **final** was in the **Variable(s)** box, it would be listed in two ways: **final** in the original scale and **zfinal** for the same variable converted to *z* scores. You may then do analyses with either the original variable or the variable converted to *z* scores. Recall that *z* scores are values that have been mathematically transposed to create a distribution with a mean of zero and a standard deviation of one. See the glossary for a more complete definition. Also note that for non-mouse users, the SPSS people have cleverly underlined the "z" in the word "standardized" as a gentle reminder that standardized scores and *z* scores are the same thing.

*To create a table of the default descriptives (mean, standard deviation, maximum, minimum) for the variables* **gpa** *and* **total**, *perform the following sequence of steps:*

| In Screen | Do This | Step 5 |
|---|---|---|
| 7.1 | ✱ **gpa** → ✱ ➡ → ✱ **total** → ✱ ➡ → ✱ **OK** | 7.3 |

If you wish to calculate more than the four default statistics, after selecting the desired variables, before clicking the **OK**, it is necessary to click the **Options** button (at the bottom of Screen 7.1). Here every descriptive statistic presented earlier in this chapter is included with a couple of exceptions: Median and mode are accessed through the Frequencies command only. See Chapter 6 to determine how to access these values. Also, the standard errors ("S.E.") of the kurtosis and skewness are not included. This is because when you click either kurtosis or skewness, the standard errors of those values are automatically included. To select the desired descriptive statistics, the procedure is simply to click (so as to leave a ☑ in the box to the left of the desired value) the descriptive statistics you wish. This is followed by a click of **Continue** and **OK**. The **Display order** options include (a) **Variable list** (the default—in the same order as displayed in the data editor), (b) **Alphabetic** (names of variables ordered alphabetically), (c) **Ascending means** (ordered from smallest mean value to largest mean value in the output), and (d) **Descending means** (from largest to smallest).

Screen 7.2 The Descriptives: Options Window



Descriptives: Options

☑ Mean ☐ Sum

Dispersion

☑ Std. deviation ☑ Minimum

☐ Variance ☑ Maximum

☐ Range ☐ S.E. mean

Distribution

☐ Kurtosis ☐ Skewness

Display Order

● Variable list

○ Alphabetic

○ Ascending means

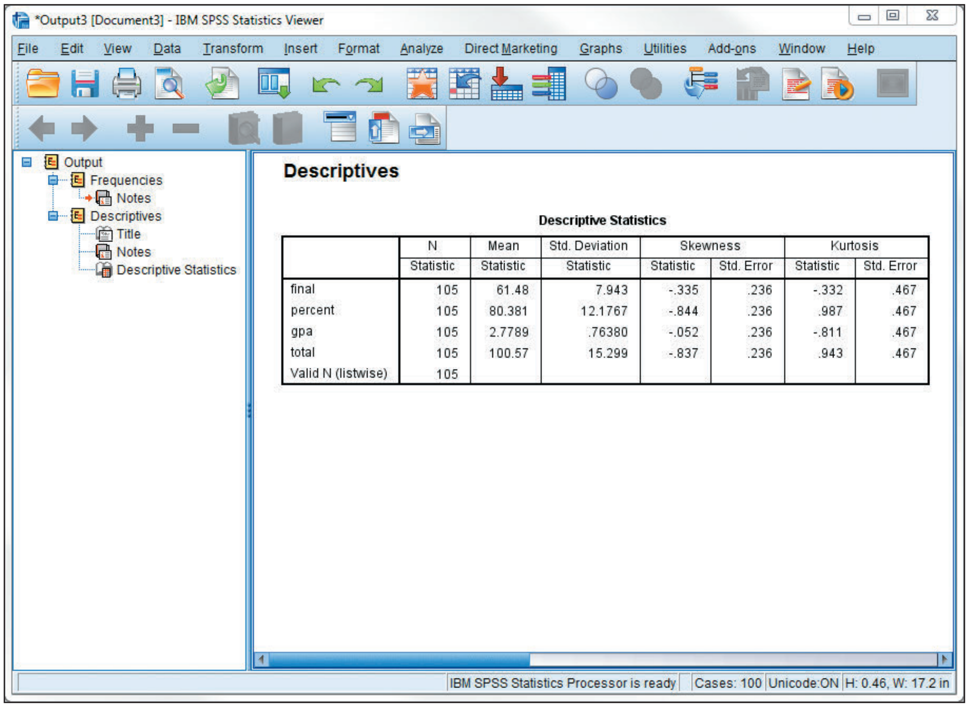
○ Descending means

*To select the variables* **final**, **percent**, **gpa**, *and* **total**, *and then select all desired descriptive statistics, and perform the following sequence of steps. Press the* **Reset** *button if there are undesired variables in the active box.*

| In Screen | Do This | Step 5a |
|---|---|---|
| 7.1 | ✳ **final** → ✳ ➡ → ✳ **percent** → ✳ ➡ → ✳ **gpa** → ✳ ➡ → ✳ **total** → ✳ ➡ → ✳ **Options** | |
| 7.2 | ✳ *all desired descriptive statistics (so a ✓ appears in each box)* → ✳ **Continue** | |
| 7.1 | ✳ **OK** | 7.3 |

Upon completion of either Step 5 or Step 5a, Screen 7.3 will appear (below). The results of the just-completed analysis are included in the top window labeled **Output# [Document#] – IBM SPSS Statistics Viewer**. Click on the ▣ to the right of this title if you wish the output to fill the entire screen and then make use of the arrows on the scroll bar (▲▼▶◀) to view the results. Even when viewing output, the standard menu of commands is still listed across the top of the window. Further analyses may be conducted without returning to the data screen. Partial output from this analysis is included in the Output section.

**Screen 7.3** SPSS Output Viewer Window


*Output3 [Document3] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Output
Frequencies
Notes
Descriptives
Title
Notes
Descriptive Statistics
IBM SPSS Statistics Processor is ready
Cases: 100
Unicode:ON
H: 0.46, W: 17.2 in

**Descriptives**

**Descriptive Statistics**

| | N | Mean | Std. Deviation | Skewness | | Kurtosis | |
|---|---|---|---|---|---|---|---|
| | Statistic | Statistic | Statistic | Statistic | Std. Error | Statistic | Std. Error |
| final | 105 | 61.48 | 7.943 | -.335 | .236 | -.332 | .467 |
| percent | 105 | 80.381 | 12.1767 | -.844 | .236 | .987 | .467 |
| gpa | 105 | 2.7789 | .76380 | -.052 | .236 | -.811 | .467 |
| total | 105 | 100.57 | 15.299 | -.837 | .236 | .943 | .467 |
| Valid N (listwise) | 105 | | | | | | |

## 7.9 Printing Results

Results of the analysis (or analyses) that have just been conducted require a window that displays the standard commands (**File Edit Data Transform Analyze...**) across the top. A typical print procedure is shown in the following page beginning with the standard output screen (Screen 1, inside back cover).

*To print results, from the Output screen perform the following sequence of steps*:

| In Screen | Do This | Step 6 |
|---|---|---|
| Back1 | *Select desired output or edit (see pages 16–22)* → ✳ **File** → ✳ **Print** | |
| 2.9 | *Consider and select desired print options offered, then* → ✳ **OK** | |

*To exit you may begin from any screen that shows the* **File** *command at the top.*

| In Screen | Do This | Step 7 |
|---|---|---|
| Menu | ✳ **File** → ✳ **Exit** | 2.1 |

Note: After clicking **Exit**, there will frequently be small windows that appear asking if you wish to save or change anything. Simply click each appropriate response.

# 7.10 Output

## 7.10.1 Descriptive Statistics

What follows is output from sequence Step 5a, page 118. Notice that the statistics requested include the N, the Mean, the Standard Deviation, the Variance, the Skewness, and the Kurtosis. The Standard Errors of the Skewness and Kurtosis are included by default.

**IBM SPSS Statistics: Descriptive Statistics**

| | N | Mean | Std. Deviation | Variance | Skewness | | Kurtosis | |
|---|---|---|---|---|---|---|---|---|
| | Statistic | Statistic | Statistic | Statistic | Statistic | Std. Error | Statistic | Std. Error |
| FINAL | 105 | 61.48 | 7.94 | 63.098 | –.335 | .236 | –.332 | .467 |
| PERCENT | 105 | 80.381 | 12.177 | 148.272 | –.844 | .236 | .987 | .467 |
| GPA | 105 | 2.779 | .763 | .583 | –.052 | .236 | –.811 | .467 |
| TOTAL | 105 | 100.57 | 15.30 | 234.074 | –.837 | .236 | .943 | .467 |
| Valid N (listwise) | 105 | | | | | | | |

First observe that in this display the entire output fits neatly onto a single page or is entirely visible on the screen. This is rarely the case. When more extensive output is produced, make use of the up, down, left, and right scroll bar arrows to move to the desired place. You may also use the index in the left window to move to particular output more quickly. Notice that all four variables fall within the "excellent" range as acceptable variables for further analyses; the skewness and kurtosis values all lie between ±1.0. All terms are identified and described in the introductory portion of the chapter. The only undefined word is **listwise**. This means that any subject that has a missing value for any variable has been deleted from the analysis. Since in the **grades.sav** file there are no missing values, all 105 subjects are included.

# Exercises

Answers to selected exercises may be downloaded at **www.spss-step-by-step.net.**

Notice that data files other than the **grades.sav** file are being used here. Please refer to the **Data Files** section starting on page 362 to acquire all necessary information about these files and the meaning of the variables. As a reminder, all data files are downloadable from the web address shown above.

1. Using the **grades.sav** file select all variables except **lastname**, **firstname**, **grade**, and **passfail**. Compute descriptive statistics, including **mean**, **standard deviation**, **kurtosis**, and **skewness**. Edit so that you eliminate **Std. Error** (Kurtosis) and **Std. Error** (Skewness) making your chart easier to interpret. Edit the output to fit on one page.
   - Draw a line through any variable for which descriptives are meaningless (either they are categorical or they are known to not be normally distributed).
   - Place an "*" next to variables that are in the ideal range for both skewness and kurtosis.
   - Place an **X** next to variables that are acceptable but not excellent.
   - Place a ψ next to any variables that are not acceptable for further analysis.
2. Using the **divorce.sav** file select **all** variables <u>except</u> the indicators (for spirituality, **sp8–sp57**, for cognitive coping, **cc1–cc11**, for behavioral coping, **bc1–bc12**, for avoidant coping, **ac1–ac7**, and for physical closeness, **pc1–pc10**). Compute descriptive statistics, including **mean**, **standard deviation**, **kurtosis**, and **skewness**. Edit so that you eliminate **Std. Error** (Kurtosis) and **Std. Error** (Skewness) and your chart is easier to interpret. Edit the output to fit on <u>two</u> pages.
   - Draw a line through any variable for which descriptives are meaningless (either they are categorical or they are known to not be normally distributed).
   - Place an "*" next to variables that are in the ideal range for both skewness and kurtosis.
   - Place an **X** next to variables that are acceptable but not excellent.
   - Place a ψ next to any variables that are not acceptable for further analysis.
3. Create a practice data file that contains the following variables and values:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| • VAR1: | 3 | 5 | 7 | 6 | 2 | 1 | 4 | 5 | 9 | 5 |
| • VAR2: | 9 | 8 | 7 | 6 | 2 | 3 | 3 | 4 | 3 | 2 |
| • VAR3: | 10 | 4 | 3 | 5 | 6 | 5 | 4 | 5 | 2 | 9 |

   Compute: the **mean**, the **standard deviation**, and **variance** and print out on a single page.

# Chapter 8
# Crosstabulation and $\chi^2$ Analyses

THE PURPOSE of crosstabulation is to show in tabular format the relationship between two or more categorical variables. Categorical variables include those in which distinct categories exist such as gender (female, male), ethnicity (Asian, White, Hispanic), place of residence (urban, suburban, rural), responses (yes, no), grades (A, B, C, D, F), and many more. Crosstabulation can be used with continuous data only if such data are divided into separate categories, such as age (0–19 years, 20–39 years, 40–59 years, 60–79 years, 80–99 years), total points (0–99, 100–149, 150–199, 200–250), and so on. While it is acceptable to perform crosstabulation with continuous data that has been categorized, it is rare to perform chi-square analyses with continuous data because a great deal of useful information about the distribution is lost by the process of categorization. For instance, in the total points distribution (above), two persons who scored 99 and 100 points, respectively, would be in the first and second categories and would be considered identical to two persons who scored 0 and 149 points, respectively. Nonetheless, crosstabulation with continuous data is often used for purposes of data description and display. The SPSS command **Crosstabs** and the subcommands **Cells** and **Statistics** are used to access all necessary information about comparisons of frequency data.

## 8.1 Crosstabulation

While the **Frequencies** command can tell us there are 5 Natives, 20 Asians, 24 Blacks, 45 Whites, and 11 Hispanics (and that there are 64 females and 41 males) in our **grades.sav** file, it cannot give us the number of female Asians or male Whites. This is the function of the **Crosstabs** command. It would be appropriate to "cross" two variables (**ethnic** by **gender**) to answer the questions posed above. This would produce a table of 10 different cells with associated frequencies inserted in each cell by crossing two (2) levels of **gender** with five (5) levels of **ethnic**. It is possible to cross three or more variables, although only with a very large data set would a researcher be likely to perform a crosstabulation with three variables because there would be many low-count and empty cells if the number of subjects was not sufficient. For the present sample, an **ethnic** by **gender** by **grade** crosstabulation would probably not be recommended. This procedure would create a 5 (**ethnic**) × 2 (**gender**) × 5 (**grade**) display of frequencies—a total of 50 cells to be filled with only 105 subjects. A large number of low-count or empty cells would be guaranteed. If such a crosstabulation were created with a larger *N*, SPSS would produce five different 5 × 2 tables to display these data.

## 8.2 Chi-Square ($\chi^2$) Tests of Independence

In addition to frequencies (or the *observed values*) within each cell, SPSS can also compute the expected value for each cell. *Expected value* is based on the assumption that the two variables are independent of each other. A simple example demonstrates the derivation of expected value. Suppose there is a group of 100 persons in a room and that 30 are male and 70 are female. If there are 10 Asians in the group, it would be anticipated (expected)—if the two variables are independent of each other—that among the 10 Asians there would be 3 males and 7 females (the same proportion as is observed in the entire group). However, with the same group of 100, if 10 of them were football players we would *not* expect 3 male football players and 7 female football players. In American society, most football players are male, and the two categories (gender and football players) are *not* independent of each other. If there were no additional information given, it would be expected that all 10 of the players would be male. The purpose of a chi-square test of independence is to determine whether the observed values for the cells deviate significantly from the corresponding expected values for those cells.

You may wish to test levels of a *single* variable. For instance, if you wish to determine whether actual values differ significantly from expected values for the five levels of ethnicity included in the **ethnic** variable you need to use a different chi-square test. The actual test you use is described on pages 221–223 of the Nonparametric Procedures chapter. However, with only one variable SPSS cannot compute expected values. You need to specify which expected values you wish. In the **ethnic** variable in the **Grades.sav** file, among the 105 subjects there are actually 5 Natives, 20 Asians, 24 Blacks, 45 Caucasians, and 11 Hispanics. The default expected value (explained on page 222) is equal numbers in each group—that is 105/5, or 21 in each of the five groups. You may specify different expected values if you wish. For instance you may want to see if your sample differs significantly from the actual ethnic breakdown of your community. Then you enter numbers that represent the proportion of each ethnic group as your expected values.

The chi-square statistic is computed by summing the squared deviations [observed value ($f_o$) minus expected value ($f_e$)] divided by the expected value for each cell:

$$\chi^2 = \Sigma[(f_o - f_e)^2/f_e]$$

For every *inferential* statistic (this is the first of many in this book!), there are three questions that you need to ask as a researcher. The first question is, "Are you fairly certain that the difference tested by this statistic is real instead of random?" In this case, if there is a large discrepancy between the observed values and the expected values, the $\chi^2$ statistic would be large, suggesting a significant difference between observed and expected values. ("Significant difference" is statistics-speak for "fairly certain that this difference is real.") Along with this statistic, a probability value is computed. The value of $p$ is the probability that the difference between observed and expected values tested by the $\chi^2$ statistic is due to chance. With $p < .05$, it is commonly accepted that the observed values differ significantly from the expected values and that the two variables are *NOT* independent of each other. More complete descriptions and definitions are included in the Output section of this chapter.

The second question that you need to ask for every inferential statistic is, "How big is the difference tested by the statistic?" Depending on what your data look like, it is possible to have a small $p$ value (and be very confident that the difference is real) even if the difference is very small, or to have a large $p$ value (and not be confident that the difference is real) even if the difference in your data appears to be large. The $\chi^2$ statistic itself isn't a good measure of how big the difference between the observed and expected values because it is largely dependent on the number of dimensions and sample size. Thus comparisons of one chi-square value with another are often misleading.

To control for this difficulty, Pearson suggested the phi ($\varphi$) statistic, which divides the chi-square value by $N$ and then takes the positive square root of the result. The purpose was to standardize a measure of association (also known as *effect size*) to values between 0 and 1 (with 0 indicating completely independent variables and a value close to 1 indicating a strong association between variables). However, if one of the dimensions of the crosstabulation is larger than 2, $\varphi$ may attain a value larger than 1.0. To control for this, Cramér's V was introduced (the positive square root of $\chi^2/[N(k - 1)]$, where $k$ is the smaller of the number of rows and columns). This effect size measure *does* vary between 0 and 1.0 and is a commonly used measure of the strength of association between variables in a $\chi^2$ analysis.

The final question that you need to ask for every inferential statistic is, “Is this finding important?” The answer to that question will depend on whether you’re confident that the finding isn’t the result of randomness, how big the effect is, and how much you care. Example: you may not care about a small decrease in the likelihood of blinking because of taking a certain medication, but you may care a lot about a small decrease in the likelihood of dying because of the same medication.

The file we use to illustrate **Crosstabs** is our example described in the first chapter. The data file is called **grades.sav** and has an $N$ = 105. This analysis creates crosstabulations and calculates chi-square statistics for **gender** by **ethnic**.

## 8.3 Step by Step

### 8.3.1 Crosstabulation and Chi-Square Tests of Independence

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | ✳ ⊞ → ✳ All Programs ▷ → ✳ IBM SPSS Statistics → ✳ IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | ✳ → ✳ → ✳ Σ | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| | Step 2 |
|---|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* | |

*Screens 1 and 2 (displayed on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **grades.sav** *file for further analyses:*

| In Screen | Do This | Step 3 |
|---|---|---|
| Front1 | ✳ File → Open → ✳ Data [*or* ✳ ] | |
| Front2 | type `grades.sav` → ✳ Open [*or* ✳ ✳ grades.sav] | Data |

Whether first entering SPSS or returning from earlier operations the standard menu of commands across the top is required. As long as it is visible you may perform any analyses. It is not necessary for the data window to be visible.

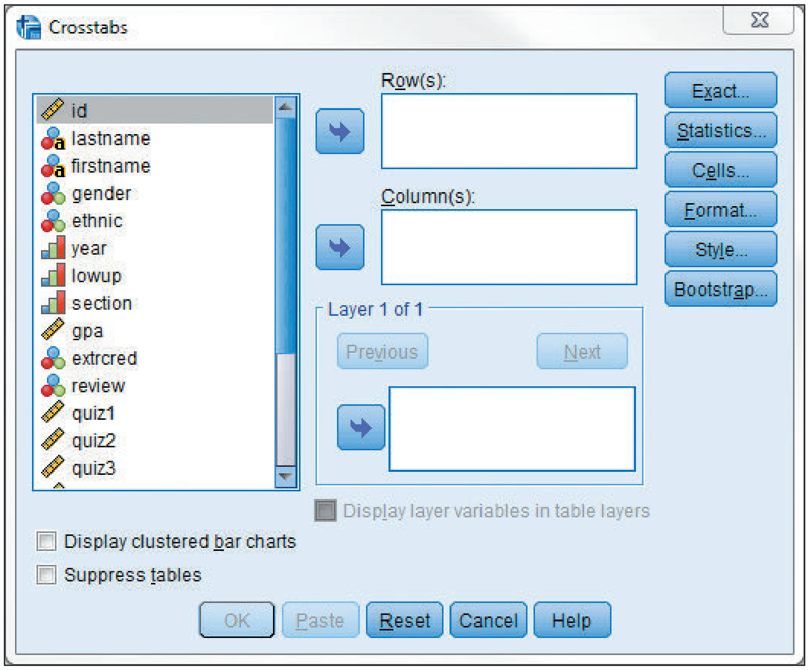
*After completion of Step 3 a screen with the desired menu bar appears. When you click a command (from the menu bar), a series of options will appear (usually) below the selected command. With each new set of options, click the desired item. The sequence to access chi-square statistics begins at any screen with the menu of commands visible:*

| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | ✳ Analyze → ✳ Descriptive Statistics → ✳ Crosstabs | 8.1 |

A new window now appears (Screen 8.1, below) that provides the framework for conducting a crosstabs analysis. The procedure is to click on the desired variable in the list to the left (**gender** in this example) and then click the top right-arrow (➡) to select gender as the row variable. Then click a second variable (**ethnic** in this case) and click the middle right-arrow (to select ethnicity as the column variable). That is all that is required to create a crosstabulation of two variables. This will create a 2 (**gender**) by 5 (**ethnic**) table that contains 10 cells. Each of the 10 cells will indicate the number of subjects in each category—4 Native females, 13 Asian females, and so forth for the other 8 cells in the table.

The lowest box in the window allows for crosstabulation of three or more variables. If, for instance, we wanted to find the gender by ethnicity breakdown for the three sections, you would click the **section** variable in the list of variables and then click the lowest of the three right arrows. This would result in three tables: a gender by ethnicity breakdown for the first section, a gender by ethnicity breakdown for the second section, and a gender by ethnicity breakdown for the third section. The **Previous** and **Next** to the left and right of **Layer 1 of 1** are used if you wanted a gender by ethnicity analysis for more than one variable. For instance, if you wanted this breakdown for both **section** and **year** (year in school), you would click **section**, click the lowest right arrow, click **Next**, click **year**, and then click the lowest right arrow again. This produces three 2 × 5 tables for **section** and four 2 × 5 tables for **year**. The sequence (below) shows how to create a **gender** by **ethnic** by **section** crosstabulation.

## Screen 8.1 The Crosstabs Window



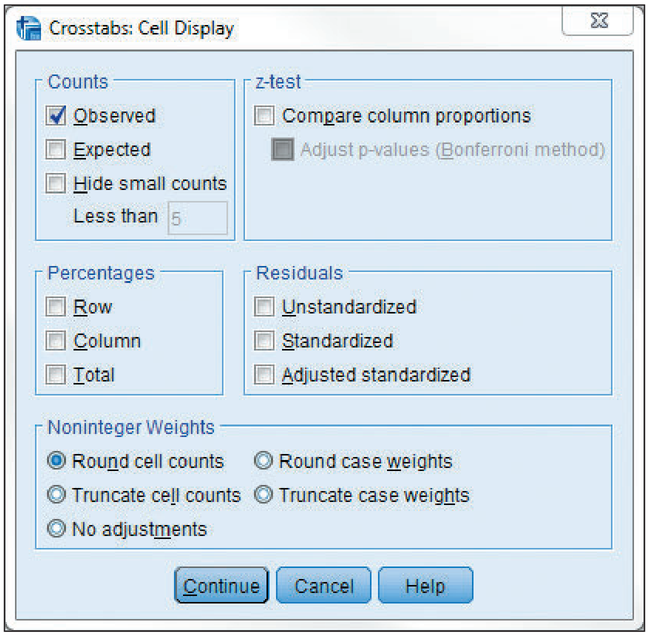
*The following sequence begins from Screen 8.1. To arrive at this screen, perform whichever of Steps 1–4 (page 123) are necessary.*

| In Screen | Do This | Step 5 |
|---|---|---|
| 8.1 | 🖱 **gender** → 🖱 *upper* ➡ → 🖱 **ethnic** → 🖱 *middle* ➡ | |
| | → 🖱 **section** → 🖱 *lowest* ➡ → 🖱 **OK** | Back1 |

For a **gender** by **ethnic** crosstabulation, omit the two clicks that involve the **section** variable.

It is rare for a researcher to want to compute only cell frequencies. In addition to frequencies, it is possible to include within each cell a number of additional options. Those most frequently used are listed below with a brief definition of each. When you press the **Cells** button (Screen 8.1), a new screen appears (Screen 8.2, below) that allows you to select a number of options.

## Screen 8.2 The Crosstabs: Cell Display Window



Crosstabs: Cell Display

Counts
☑ Observed
□ Expected
□ Hide small counts
Less than 5

z-test
□ Compare column proportions
□ Adjust p-values (Bonferroni method)

Percentages
□ Row
□ Column
□ Total

Residuals
□ Unstandardized
□ Standardized
□ Adjusted standardized

Noninteger Weights
● Round cell counts ○ Round case weights
○ Truncate cell counts ○ Truncate case weights
○ No adjustments

The **Observed** count is selected by default. The **Expected** count (more frequently referred to as the *expected value*) is in most cases also desired. Inclusion of other values depends on the preference of the researcher.

| | |
|---|---|
| • **Observed Count:** | The actual number of subjects or cases within each cell. |
| • **Expected Count:** | The expected value for each cell (see page 121) |
| • **Row Percentages:** | The percent of values in each cell for that row. |
| • **Column Percentages:** | The percent of values in each cell for that column. |
| • **Total Percentages:** | The percent of values in each cell for the whole table. |
| • **Unstandardized Residuals:** | Observed value minus expected value. |

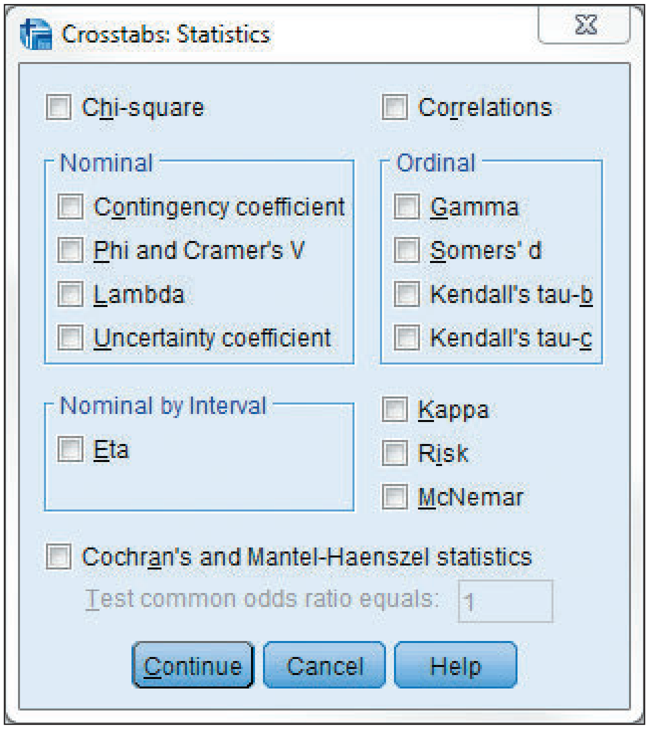
*To create a* **gender** *by* **ethnic** *crosstabulation that includes observed count, expected count, total percentages, and unstandardized residuals within each cell, perform the following sequence of steps. Begin at Screen 8.1; press* **Reset** *if variables remain from a prior analysis.*

| In Screen | Do This | Step 5a |
|---|---|---|
| 8.1 | ✳ **gender** → ✳ *upper* ➡ → ✳ **ethnic** → ✳ *middle* ➡ → ✳ **Cells** | |
| 8.2 | ✳ **Expected** → ✳ **Total** → ✳ **Unstandardized** → ✳ **Continue** | |
| 8.1 | ✳ **OK** | Back1 |

Note: We don't click on **Observed**, because this option is already selected by default.

Thus far we have only created tabulations of numbers within cells. Usually, along with crosstabulation, a chi-square analysis is conducted. This requires a click on the **Statistics** pushbutton (see Screen 8.1). When this button is clicked, a new window opens (Screen 8.3, below). Many different tests of independence or association are listed here. Only **Chi-square**, and **Phi and Cramer's V** (see page 122) and **Correlations** (explained in Chapter 10) will be considered here. As in the **Cells** window, the procedure is to click in the small box to the left of the desired statistic before returning to the previous screen to conduct the analysis. See the *IBM SPSS Statistics 19 Guide to Data Analysis* for a description of the other statistics included in this chart.

Screen 8.3 The Crosstabs: Statistics Window



*In the sequence below, we create a* **gender** *by* **ethnic** *crosstabulation, request* **Observed** ***count***, **Expected count**, *and* **Unstandardized Residuals** *within each cell, and select the* **Chi-square** *and* **Phi and Cramer's V** *as forms of analysis. We don't include correlations because they have no meaning when there is no intrinsic order to the associated variables. The starting point is Screen 8.1. Complete results from this analysis are included in the Output section.*

| In Screen | Do This | Step 5b |
|---|---|---|
| 8.1 | **gender** → *upper* ➡ → **ethnic** → *middle* ➡ → **Cells** | |
| 8.2 | **Expected** → **Unstandardized** → **Continue** | |
| 8.1 | **Statistics** | |
| 8.3 | **Chi-square** → **Phi and Cramer's V** → **Continue** | |
| 8.1 | **OK** | Back1 |

Often we may wish to conduct a crosstabulation and chi-square analysis on a *subset* of a certain variable. For instance, in the **gender** by **ethnic** crosstabulation described earlier, we may wish to delete the "Native" category from the analysis since there are only five of them and earlier analyses have indicated that there is a problem with low count cells. This then creates a 2 (levels of gender) × 4 (levels of ethnicity after excluding the first level) analysis. After you have selected the variables for crosstabulation, have chosen cell values and desired statistics, click on the **Data** command in the Menu Bar at the top of the screen. In the menu of options that opens below, click on **Select Cases**. At this point, a new window will open (Screen 4.8 from Chapter 4). In this window, click the small circle to the right of **If condition is satisfied** (so a black dot fills it), then click the **If** button just below.

A new dialog box again opens (see Screen 4.3, also from Chapter 4) titled **Select Cases: If**. This window provides access to a wide variety of operations described in Chapter 4. For the present we are only concerned with how to select levels 2, 3, 4, and 5 of the **ethnic** variable. First step is to select **ethnic** from the variable list to the left, then click the ➡ to paste the variable into the "active" box, then click the >= (on the small keyboard below the active box), and then click the 2. You have now indicated that you wish to select all levels of **ethnic** greater or equal to 2. Then click **Continue**. Screen 4.8 will reappear; click **OK**, the original screen will appear (Screen 8.1) and then click **OK**. Your analysis will now be completed with only four levels of ethnicity. The step-by-step sequence follows.

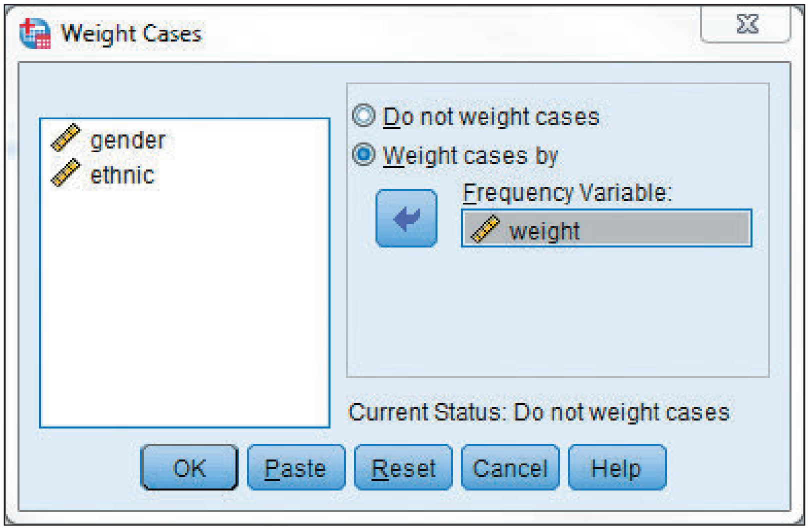
*The starting point for this sequence is Screen 8.1. It may be necessary to press* **Reset**.

| In Screen | Do This | Step 5c |
|---|---|---|
| 8.1 | ✳ **gender** → ✳ *upper* ➡ → ✳ **ethnic** → ✳ *middle* ➡ → ✳ **Cells** | |
| 8.2 | ✳ **Expected** → ✳ **Unstandardized** → ✳ **Continue** | |
| 8.1 | ✳ **Statistics** | |
| 8.3 | ✳ **Chi-square** → ✳ **Phi and Cramer's V** → ✳ **Continue** | |
| 8.1 | ✳ **Data** *(at the top of the screen in the menu bar)* → ✳ **Select Cases** | |
| 4.8 | ✳ *the* **O** *to the left of* **If condition is satisfied** → ✳ **If** *bar just beneath* | |
| 4.3 | ✳ **ethnic** ✳ ➡ → ✳ >= → ✳ 2 → ✳ **Continue** | |
| 4.8 | ✳ **OK** | |
| 8.1 | ✳ **OK** | Back1 |

## 8.4 Weight Cases Procedure: Simplified Data Setup

All of the chi-square analyses assume that the data is set up as described in Chapter 4: Each person has one line, and every line is a case. But sometimes this is inconvenient: If you have thousands of people in your study, and you already know how many people are in each cell, you can no doubt make better use of your time than entering the same information in thousands of rows. In that case, you can use the **Weight Cases** procedure with a simplified data file.

Screen 8.4 The Weight Cases Window



gender
ethnic

○ Do not weight cases
● Weight cases by
Frequency Variable: weight

Current Status: Do not weight cases

| | Gender | Ethnic | Weight |
|---|---|---|---|
| 1 | Female | Native | 4 |
| 2 | Female | Asian | 13 |
| 3 | Female | Black | 14 |
| 4 | Female | White | 26 |
| 5 | Female | Hispanic | 7 |
| 6 | Male | Native | 1 |
| 7 | Male | Asian | 7 |
| 8 | Male | Black | 10 |
| 9 | Male | White | 19 |
| 10 | Male | Hispanic | 4 |

To do this, you first have to set up your categorical variables (the same way you would in a regular data file). For the example in this chapter, you would set up a variable to code for ethnicity and a variable to code for gender. Then you make an additional variable—you can name it whatever you like, but we'd suggest something like "**Weight**" or "**Case-Count**"—that includes how many cases have that particular combination of categorical variables. For example, in the example used in this chapter, there are 4 Native American females, 13 Asian females, 14 Black females, and so on (see the output below on page 129 for all of the counts). To enter this data, you could set the data file up like this (left):

So instead of one row per case, there is one row per combination of categorical variables. The next step is then to tell SPSS that you want it to pretend that each row is actually there many times—as many times as the weighting variables (the variable **weight** in the chart) indicate. Here is the step-by-step sequence that will accomplish that

| In Screen | Do This | Step 5d |
|---|---|---|
| 8.4 | ✳ **Data** → ✳ **Weight Cases** → ✳ **Weight cases by** → ✳ **weight** → ✳ ↪ → ✳ **OK** | |

Then you are ready to run the Crosstabs analysis as if you had actually entered every category multiple times. If, for instance, you completed the steps shown in Step 5a, despite the tiny file (shown above) the results would be identical to the results if you completed the sequences in Step 5a with the far larger **Grades.sav** file.

Upon completion of Step 5, 5a, 5b, or 5c, the output screen will appear (Screen 1, inside back cover). All results from the just-completed analysis are included in the Output Navigator. Make use of the scroll bar arrows (▲▼▶◀) to view the results. Even when viewing output, the standard menu of commands is still listed across the top of the window. Further analyses may be conducted without returning to the data screen.

# 8.5 Printing Results

Results of the analysis (or analyses) that have just been conducted require a window that displays the standard commands (**File Edit Data Transform Analyze . . .**) across the top. A typical print procedure is shown below beginning with the standard output screen (Screen 1, inside back cover).

*To print results, from the Output screen perform the following sequence of steps:*

| In Screen | Do This | Step 6 |
|---|---|---|
| Back1 | *Select desired output or edit (see pages 16–22)* → ✳ **File** → ✳ **Print** | |
| 2.9 | *Consider and select desired print options offered, then* → ✳ **OK** | |

*To exit you may begin from any screen that shows the* **File** *command at the top.*

| In Screen | Do This | Step 7 |
|---|---|---|
| Menu | ✳ **File** → ✳ **Exit** | 2.1 |

Note: After clicking **Exit**, there will frequently be small windows that appear asking if you wish to save or change anything. Simply click each appropriate response.

# 8.6 Output

## 8.6.1 Crosstabulation and Chi-Square ($\chi^2$) Analyses

What follows is partial output from sequence Step 5b, page 126.

| | | | Ethnicity | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Native | Asian | Black | White | Hispanic | Total |
| Gender | Female | Count | 4 | 13 | 14 | 26 | 7 | 64 |
| | | Expected Count | 3.0 | 12.2 | 14.6 | 27.4 | 6.7 | 64.0 |
| | | Residual | 1.0 | .8 | –.6 | –1.4 | .3 | |
| | Male | Count | 1 | 7 | 10 | 19 | 4 | 41 |
| | | Expected Count | 2.0 | 7.8 | 9.4 | 17.6 | 4.3 | 41.0 |
| | | Residual | –1.0 | –.8 | .6 | 1.4 | –.3 | |
| Total | | Count | 5 | 20 | 24 | 45 | 11 | 105 |
| | | Expected Count | 5.0 | 20.0 | 24.0 | 45.0 | 11.0 | 105.0 |

| | Value | df | Significance |
|---|---|---|---|
| Pearson Chi-Square | 1.193[a] | 4 | .879 |
| Likelihood Ratio | 1.268 | 4 | .867 |
| Linear-by-Linear Association | .453 | 1 | .501 |
| Phi | .107 | -- | .879 |
| Cramér's V | .107 | -- | .879 |
| N of valid cases | 105 | | |

[a] 3 cells (30%) have expected count less than 5. The minimum expected count is 1.95.

The first step in interpreting a crosstabulation or chi-square analysis is to observe the actual values and the expected values within each cell. Preliminary observation indicates that observed values and expected values are quite similar. The greatest discrepancy is for Whites (females: actual count = 26, expected = 27.4; and males: actual count = 19, expected = 17.6). Note also that the residual value (the number beneath the other two) is simply the observed value minus the expected value. Even without looking at the chi-square statistics you would anticipate that observed values and expected values do not differ significantly (that is, gender and ethnicity in this sample are independent of each other). The results support this observation with a low chi-square value (1.193) and a significance greater than .8 (.879). Notice that measures of association are also small and do not approach significance. As suggested in the Step by Step section, low-count cells are a problem; 3 of 10 have an expected value less than 5. The usual response would be to redo the analysis after deleting the "Native" category. To further assist understanding, definitions of output terms follow:

| Term | Definition/Description |
|---|---|
| **Count** | The top number in each of the 10 cells (4, 13, 14, . . .), indicates the number of subjects in each category. |
| **Expected Count** | The second number in each of the 10 cells (3.0, 12.2, 14.6, . . .), indicates the number that would appear there if the two variables were perfectly independent of each other. |
| **Residual** | The observed value minus the expected value. |
| **Row Total** | The total number of subjects for each row (64 females, 41 males). |
| **Column Total** | The total number of subjects in each category for each column (5 Natives, 20 Asians, 24 Blacks, 45 Whites, 11 Hispanics). |
| **Chi Square: Pearson and Likelihood Ratio** | Two different methods for computing chi-square statistics. With a large *N*, these two values will be close to equal. The formula for the Pearson chi-square is<br>$\chi^2 = \Sigma[(f_0 - f_e)^2/f_e]$ |
| **Value** | For **PEARSON** and **MAXIMUM LIKELIHOOD** methods, as the test-statistic value gets larger the likelihood that the two variables are *not* independent (e.g., *are* dependent) also increases. The values close to 1 (1.193, 1.268) suggest that gender balance is not dependent on which ethnic groups are involved. |
| **Degrees of Freedom (df)** | Degrees of freedom is the number of levels in the first variable minus 1 (2 – 1 = 1) times the number of levels in the second variable minus 1 (5 – 1 = 4); 1 × 4 = 4. |
| **Significance** | The likelihood that these results could happen by chance. The large *p* value here indicates that observed values do *not* differ significantly from expected values. |
| **Linear-by-Linear Association** | This statistic tests whether the two variables correlate with each other. This measure is often meaningless because there is no logical or numeric relation to the order of the variables. For instance, there is no logical order (from a low value to a high value) of ethnicity. Therefore, the correlation between gender and ethnicity is meaningless. If, however, the second variable were income, ordered from low to high, a valid correlation results. |
| **Minimum Expected Count** | The minimum expected count is for the first cell of the second row (male Natives). The expected value there is rounded off to the nearest 10th (2.0). The value accurate to two decimals is 1.95. |
| **Phi** | A measure of the strength of association or effect size between two categorical variables. A value of .107 represents a very weak association between gender and ethnicity. The equation:<br>$\Phi = \sqrt{\chi^2/N}$ |
| **Cells with Expected Count < 5** | Three of the ten cells have an expected frequency less than 5. If you have many low-count cells (more than 25% is one accepted criteria), the overall chi-square value is less likely to be valid. |
| **Cramér's V** | A measure of the strength of association or effect size between two categorical variables. It differs from phi in that Cramér's V varies strictly between 0 and 1, while in certain cases phi may be greater than 1. The equation follows (Note: *k* is the smaller of the number of rows and columns):<br>$V = \sqrt{\chi^2/[N(k - 1)]}$ |
| **Approximate Significance** | This is the same as the significance for the Pearson chi-square. The high value (.879) indicates very weak association. |

# Exercises

Answers to selected exercises are downloadable at **www.spss-step-by-step.net.**

For each of the chi-square analyses computed below:

1. Circle the observed (actual) values.
2. Box the expected values.
3. Put an * next to the unstandardized residuals.
4. Underline the significance value that shows whether observed and expected values differ significantly.
5. Make a statement about independence of the variables involved.
6. State the nature of the relationship. #5 identifies whether there is a relationship, now you need to indicate what that relationship is. Example: Men tend to help more with goal-disruptive problems, whereas women tend to help more with relational problems.
7. Is there a significant linear association?
8. Does linear association make sense for these variables?
9. Is there a problem with low-count cells?
10. If there is a problem, what would you do about it?

1. File: **grades.sav**. Variables: **gender** by **ethnic**. Select: **observed count**, **expected count**, **unstandarized residuals**. Compute: **Chi-square**, **Phi and Cramer's V**. Edit to fit on one page, print out, and then perform the 10 operations listed above.
2. File: **grades.sav**. Variables: **gender** by **ethnic**. Prior to analysis, complete the procedure shown in Step 5c (page 127) to eliminate the "Native" category (low-count cells). Select: **observed count**, **expected count**, **unstandarized residuals**. Compute: **Chi-square**, **Phi and Cramer's V**. Edit to fit on one page, print out, and then perform the 10 operations listed to the left.
3. File: **helping3.sav**. Variables: **gender** by **problem**. Select: **observed count**, **expected count**, **unstandarized residuals**. Compute: **Chi-square**, **Phi and Cramer's V**. Edit to fit on one page, print out, and then perform the 10 operations listed to the left.
4. File: **helping3.sav**. Variables: **school** by **occupat**. Prior to analysis, select cases: "**school** > 2 & **occupat** < 6". Select: **observed count**, **expected count**, **unstandarized residuals**. Compute: **Chi-square**, **Phi and Cramer's V**. Edit to fit on one page, print out, and then perform the 10 operations listed to the left.
5. File: **helping3.sav**. Variables: **marital** by **problem**. Prior to analysis, eliminate the "DTS" category (marital < 3). Select: **observed count**, **expected count**, **unstandarized residuals**. Compute: **Chi-square**, **Phi and Cramer's V**. Edit to fit on one page, print out, and then perform the 10 operations listed to the left.

# Chapter 9
# The Means Procedure

WHILE THE **Crosstabs** procedure allows you to identify the frequency of certain types of categorical data (Chapter 8), the **Means** command allows you to explore certain characteristics of continuous variables within those categories. By way of comparison, a crosstabulation of **ethnic** by **gender** would indicate that there were 13 White females, 22 White males, 8 Hispanic females, 6 Hispanic males, and so forth. The **Means** command allows you to view certain characteristics of continuous variables (such as total points, GPAs, percents) by groups. Thus if you computed **total** (number of points) for **ethnic** by **gender**, you would find that there were 13 White females who scored an average (mean) of 113.12 points, 22 White males who scored a mean of 115.34 points, 8 Hispanic females who scored a mean of 116.79, 6 Hispanic males with a mean of 113.45, and so forth. This information is, of course, presented in tabular format for ease of reading and interpretation. The utility of the **Means** command for data such as our sample file is several-fold. For a class with more than one section, we might like to see mean scores for each section, or to compare the scores of males with females, or the performance of upper- with lower-division students.

The **Means** command is one of SPSS's simplest procedures. For the selected groups it will list the mean for each group, the standard deviation, and the number of subjects for each category. There is an additional **Options** subcommand with which you may conduct a one-way analysis of variance (ANOVA) based on the means and standard deviations you have just produced. We will include that option in this chapter but will save a detailed explanation of analysis of variance for the **One-Way ANOVA** and **General Linear Models** chapters (Chapters 12–14).

We again make use of the **grades.sav** file ($N = 105$) and the variables **total**, **percent**, **gpa**, **section**, **lowup**, and **gender** to illustrate the **Means** procedure.

## 9.1 Step by Step

### 9.1.1 Describing Subpopulation Differences

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | → All Programs ▷ → IBM SPSS Statistics → IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | → → | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| | Step 2 |
|---|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* | |

*Screens 1 and 2 (displayed on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **grades.sav** *file for further analyses:*

| In Screen | Do This | Step 3 |
|---|---|---|
| Front1 | 🖱 File → Open → 🖱 Data [*or* 🖱 📂] | |
| Front2 | type `grades.sav` → 🖱 Open [*or* 🖱 🖱 grades.sav] | Data |

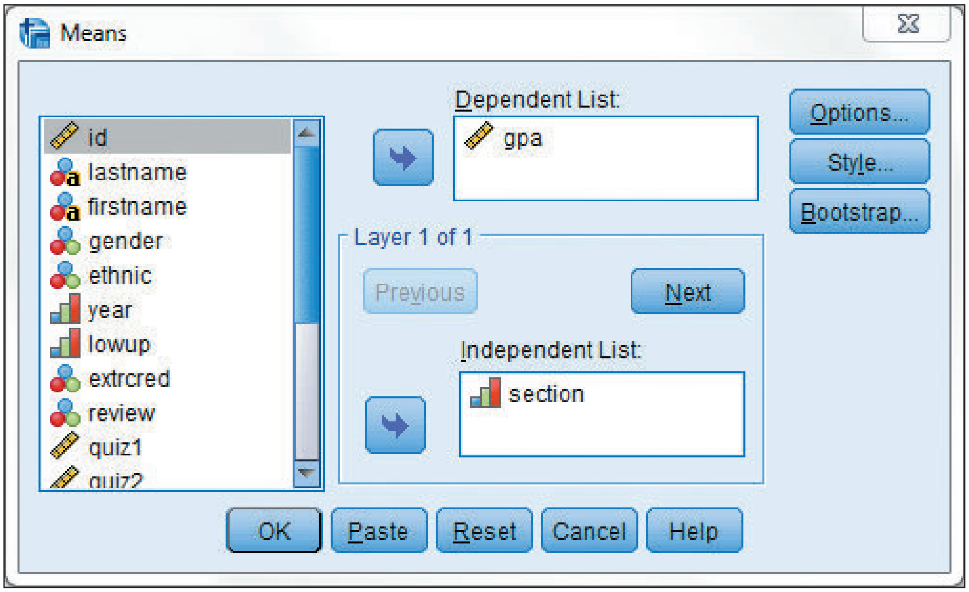
Whether first entering SPSS or returning from earlier operations the standard menu of commands across the top is required. As long as it is visible you may perform any analyses. It is not necessary for the data window to be visible.

*After completion of Step 3 a screen with the desired menu bar appears. When you click a command (from the menu bar), a series of options will appear (usually) below the selected command. With each new set of options, click the desired item. The sequence to access* **Means** *begins at any screen with the menu of commands visible:*

| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | 🖱 Analyze → Compare Means → 🖱 Means | 9.1 |

At this point a new window opens (Screen 9.1, below) that deals with designating for which variables you wish to compare means. At the top of the window is the **Dependent List** box. This is where the *continuous* variables that you wish to analyze will be placed. For instance, you may compare mean scores for previous GPA, for final exam scores, for total points, or for percentage of total points. You may list several variables in this box; SPSS will calculate a separate set of means for each variable.

**Screen 9.1** The Means Window



Dependent List: gpa

Layer 1 of 1

Independent List: section

The lower **Independent List** box is where you identify the categorical variables, such as gender, section, grade, or ethnicity. If you include only one variable (such as **gender**) in the lower **Independent List** box, and a single variable in the upper **Dependent List** box (say, **total** points), the **Means** command will indicate the average (or mean) number of total points earned by women and the mean number of total points earned by men. The *N* and standard deviations are also included by default.

More frequently the researcher will desire more than just two categories. A more common operation would be a **gender** by **section** analysis that would give mean total scores for males and females in each of the three sections, or a **gender** by **year** analysis that would yield mean **total** scores for males and females for each year in college. To specify more than one categorical variable, make use of the **Previous** and **Next** to the left and right of **Layer 1 of 1** in the middle of the screen. As you observe the sequences of steps shown below, how these options are used will become clear. Each sequence that follows begins from Screen 9.1.

*To determine the mean number of total points* (**total**) *in each* **section**, *perform the following sequence of steps. Screen 9.1 is the starting point.*

| In Screen | Do This | Step 5 |
|---|---|---|
| 9.1 | ✱ **total** → ✱ *upper* ➥ → ✱ **section** → ✱ *lower* ➥ → ✱ **OK** | Back1 |

*If you wish to include an additional categorical variable in the analysis* (**lowup**—*lower or upper division student*) *in addition to* **section**, *the appropriate sequence of steps follows*:

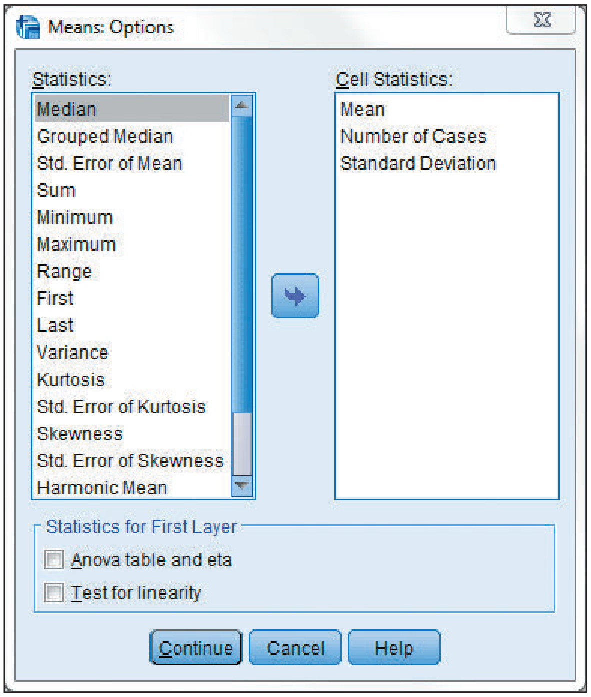
| In Screen | Do This | Step 5a |
|---|---|---|
| 9.1 | ✱ **total** → ✱ *upper* ➥ → ✱ **section** → ✱ *lower* ➥ → | |
| | ✱ **Next** → ✱ **lowup** → ✱ *lower* ➥ → ✱ **OK** | Back1 |

*You may list more than one dependent variable. SPSS will then produce as many columns in the table as there are dependent variables. In the procedure that follows, we compute means and standard deviations for* **gpa**, **total**, *and* **percent** *(three dependent variables) for six categories (three levels of* **section** *by two levels of* **gender**) *in the analysis.*

| In Screen | Do This | Step 5b |
|---|---|---|
| 9.1 | ✱ **gpa** → ✱ *upper* ➥ → ✱ **total** → ✱ *upper* ➥ | |
| | → ✱ **percent** → ✱ *upper* ➥ → ✱ **section** → ✱ *lower* ➥ | |
| | → ✱ **Next** → ✱ **gender** → ✱ *lower* ➥ → ✱ **OK** | Back1 |

If you wish to conduct a one-way analysis of variance or to include additional output within each cell, a click on the **Options** button will open a new screen (Screen 9.2, following page).

Screen 9.2 The Means: Options Window



Statistics:

- Median
- Grouped Median
- Std. Error of Mean
- Sum
- Minimum
- Maximum
- Range
- First
- Last
- Variance
- Kurtosis
- Std. Error of Kurtosis
- Skewness
- Std. Error of Skewness
- Harmonic Mean

Cell Statistics:

- Mean
- Number of Cases
- Standard Deviation

Statistics for First Layer

□ Anova table and eta

□ Test for linearity

This screen allows the researcher to include additional output options for each analysis. For instance, in addition to the default (**Mean**, **Standard Deviation**, and **Number of Cases**), you may, by clicking over items from the box to the left, also include a number of other options. Using the **Means** command, it is possible to conduct a simple one-way analysis of variance (ANOVA). We will save detailed description of this procedure until Chapter 12, but will show how to access that option here. Under the **Statistics for the First Layer** box, click the **Anova table and eta** option. If you conduct this analysis with a "first layer" grouping variable of **section** and a dependent variable of **total**, then an analysis will be conducted that compares the mean total points for each of the three sections.

*The following sequence of steps will produce means on* **total** *points for the six categories of* **gender** *by* **section**. *It will also conduct a one-way analysis of variance for the first-layer variable* (**section**) *on* **total** *points. We begin at Screen 9.1. Do whichever of Steps 1–4 (pages 132–133) are necessary to arrive at this screen. You may need to click the* **Reset** *button.*

| In Screen | Do This | Step 5c |
|---|---|---|
| 9.1 | ✳ **total** → ✳ *upper* ➡ → ✳ **section** → ✳ *lower* ➡ → ✳ **Next** | |
| | → ✳ **gender** → ✳ *lower* ➡ → ✳ **Options** | |
| 9.2 | ✳ **Anova table and eta** → ✳ **Continue** | |
| 9.1 | ✳ **OK** | **Back1** |

Upon completion of Step 5, 5a, 5b, or 5c, the output screen will appear (Screen 1, inside back cover). All results from the just-completed analysis are included in the Output Navigator. Make use of the scroll bar arrows (▲▼▶◀) to view the results. Even when viewing output, the standard menu of commands is still listed across the top of the window. Further analyses may be conducted without returning to the data screen.

## 9.2 Printing Results

Results of the analysis (or analyses) that have just been conducted require a window that displays the standard commands (**File Edit Data Transform Analyze . . .**) across the top. A typical print procedure is shown below beginning with the standard output screen (Screen 1, inside back cover).

*To print results, from the Output screen perform the following sequence of steps:*

| In Screen | Do This | Step 6 |
|---|---|---|
| Back1 | *Select desired output or edit (see pages 16–22)* → ✱ **File** → ✱ **Print** | |
| 2.9 | *Consider and select desired print options offered, then* → ✱ **OK** | |

*To exit you may begin from any screen that shows the* **File** *command at the top.*

| In Screen | Do This | Step 7 |
|---|---|---|
| Menu | ✱ **File** → ✱ **Exit** | 2.1 |

Note: After clicking **Exit**, there will frequently be small windows that appear asking if you wish to save or change anything. Simply click each appropriate response.

## 9.3 Output

### 9.3.1 Describing Subpopulation Differences

What follows is complete output from sequence Step 5c, page 135.

**Means and One-Way ANOVA Results**

**Total**

| Section | Gender | Mean | N | Std. Deviation |
|---|---|---|---|---|
| 1 | Female | 103.95 | 20 | 18.135 |
| | Male | 106.85 | 13 | 13.005 |
| | Total | 105.09 | 33 | 16.148 |
| 2 | Female | 100.00 | 26 | 12.306 |
| | Male | 98.46 | 13 | 11.822 |
| | Total | 99.49 | 39 | 12.013 |
| 3 | Female | 102.83 | 18 | 10.678 |
| | Male | 90.73 | 15 | 21.235 |
| | Total | 97.33 | 33 | 17.184 |
| Total | Female | 102.03 | 64 | 13.896 |
| | Male | 98.29 | 41 | 17.196 |
| | Total | 100.57 | 105 | 15.299 |

### ANOVA Table

| | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| TOTAL * SECTION | Between Groups (Combined) | 1065.910 | 2 | 532.955 | 2.335 | .102 |
| | Within Groups | 23277.804 | 102 | 228.214 | | |
| | Total | 24343.714 | 104 | | | |

### Measures of Association

| | Eta | Eta Squared |
|---|---|---|
| TOTAL * SECTION | .209 | .044 |

Note that the first portion of the output (table, previous page) is simply the means and frequencies for the entire group and for each of the selected categories. SPSS lists the mean, standard deviation, and number of cases for the first of the two categorical variables (**section**) in the upper portion of the chart, and similar information for gender in the lower portion of the chart. Definitions of terms in the ANOVA output are listed below. With means of 105.1, 99.5, and 97.3, the test statistics yield a *p* value of .102. This finding suggests that the sections may have a trend toward a significant influence on the total scores. Pairwise comparisons are not possible with this very simple ANOVA procedure, but visual inspection reveals that the greatest difference is between Section 1 (*M* = 105.1) and Section 3 (*M* = 97.3). See Chapter 12 for a more complete description of one-way ANOVAs. We conclude with definitions of output terms and the exercises.

| Term | Definition/Description |
|---|---|
| **Within-Group Sum of Squares** | The sum of the squared deviations between the mean for each group and the observed values of each subject within that group. |
| **Between-Group Sum of Squares** | The sum of squared deviations between the grand mean and each group mean, weighted (multiplied) by the number of subjects in each group. |
| **Between-Groups Degrees of Freedom** | Number of groups minus one (3 – 1 = 2). |
| **Within-Groups Degrees of Freedom** | Number of subjects minus number of groups (105 – 3 = 102). |
| **Mean Square** | Sum of squares divided by degrees of freedom. |
| ***F* ratio** | Between-groups mean square divided by within-groups mean square. |
| **Significance** | The probability of the observed values happening by chance. This probability answers the question, “Are you fairly certain that the difference tested by the *F* statistic is real instead of random?” In this case, the *p* value indicated here (*p* = .10) indicates that a marginally significant difference between means may exist among at least one of the three pairings of the three sections. We aren’t fairly certain, but it’s probably worth exploring further. |
| **Eta** | This effect size measures the strength of the relation between the two variables, and answers the question “How big is the difference tested by the *F* statistic?” |
| **Eta Squared** | The proportion of the variance in the dependent variable accounted for by the independent variable. For instance, an eta squared of .044 indicates that 4.4% of the variance in the **total** scores is due to membership in one of the three **sections**. Whether you want to use eta or eta squared to answer the question, “How big is the difference tested by the *F* statistic?” will depend on what those in your particular field use; mathematically either is fine as you can convert from one to the other by squaring or taking the square root. Eta squared is easier to compare with other analyses, but eta gives a bigger number so some people prefer it for that reason. |

# Exercises

Answers to selected exercises are downloadable at **www.spss-step-by-step.net**.

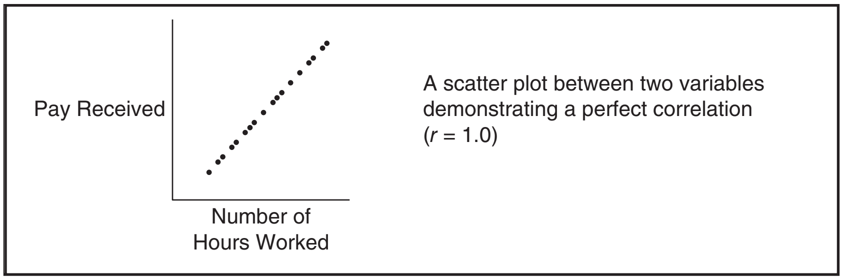
1. Using the **grades.sav** file use the Means procedure to explore the influence of **ethnic** and **section** on **total**. Print output, fit on one page, and, in general terms, describe what the value in each cell means.
2. Using the **grades.sav** file use the Means procedure to explore the influence of **year** and **section** on **final**. Print output, fit on one page, and, in general terms, describe what the value in each cell means.
3. Using the **divorce.sav** file use the Means procedure to explore the influence of gender (**sex**) and marital status (**status**) on **spiritua** (spirituality—high score is spiritual). Print output and, in general terms, describe what the value in each cell means.

# Chapter 10
# Bivariate Correlation

CORRELATIONS MAY be computed by making use of the SPSS command **Correlate**. Correlations are designated by the lowercase letter $r$, and range in value from −1 to +1. A correlation is often called a *bivariate correlation* to designate a simple correlation between two variables, as opposed to relationships among more than two variables, as frequently observed in multiple regression analyses or structural equation modeling. A correlation is also frequently called the Pearson product-moment correlation or the Pearson $r$. Karl S. Pearson is credited with the formula from which these correlations are computed. Although the Pearson $r$ is predicated on the assumption that the two variables involved are approximately normally distributed, the formula often performs well even when assumptions of normality are violated or when one of the variables is discrete. Ideally, when variables are not normally distributed, the Spearman correlation (a value based on the rank order of values) is more appropriate. Both Pearson and Spearman correlations are available using the **Correlate** command. There are other formulas from which correlations are derived that reflect characteristics of different types of data, but a discussion of these goes beyond the scope of this book. See the *IBM SPSS Statistics Guide to Data Analysis* for additional information.

## 10.1 What is a Correlation?

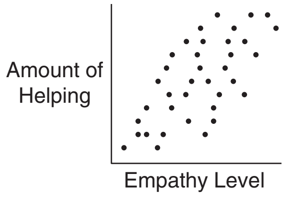
**Perfect positive ($r$ = 1) correlation**: A correlation of +1 designates a perfect, positive correlation. *Perfect* indicates that one variable is precisely predictable from the other variable. *Positive* means that as one variable increases in value, the other variable also increases in value (or conversely, as one variable decreases, the other variable also decreases).



A scatter plot between two variables demonstrating a perfect correlation ($r$ = 1.0)

Perfect correlations are essentially never found in the social sciences and exist only in mathematical formulas and direct physical or numerical relations. An example would be the relationship between the number of hours worked and the amount of pay received. As one number increases, so does the other. Given one of the values, it is possible to precisely determine the other value.

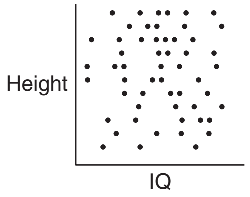
**Positive (0 < $r$ < 1) correlation**: A positive (but not perfect) correlation indicates that as the value of one variable increases, the value of the other variable also *tends* to increase. The closer the correlation value is to 1, the stronger is that tendency, and the closer the correlation value is to 0, the weaker is that tendency.



Example of a positive correlation $(0 < r < 1)$

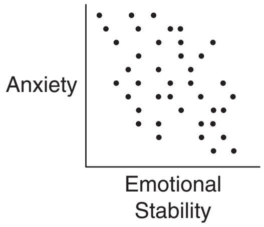
An example of a strong positive correlation is the relation between height and weight in adult humans ($r = .83$). Tall people are usually heavier than short people. An example of a weak positive correlation is the relation between a measure of empathic tendency and amount of help given to a needy person ($r = .12$). Persons with higher empathic tendency scores give more help than persons who score lower, but the relationship is weak.

**No ($r$ = 0) correlation**: A correlation of 0 indicates no relation between the two variables. For example, we would not expect IQ and height in inches to be correlated.



Example of a zero correlation $(r = 0)$

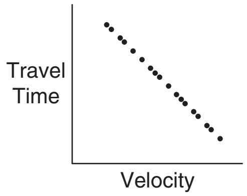
**Negative (–1 < $r$ < 0) correlation**: A negative (but not perfect) correlation indicates a relation in which as one variable *increases* the other variable has a tendency to *decrease*. The closer the correlation value is to $-1$, the stronger is that tendency. The closer the correlation value is to 0, the weaker.



Example of a negative correlation $(-1 < r < 0)$

An example of a strong negative correlation is the relation between anxiety and emotional stability ($r = -.73$). Persons who score higher in anxiety tend to score lower in emotional stability. Persons who score lower in anxiety tend to score higher in emotional stability. A weak negative correlation is demonstrated in the relation between a person's anger toward a friend suffering a problem and the quality of help given to that friend ($r = -.13$). If a person's anger is *less* the quality of help given is *more*, but the relationship is weak.

**Perfect negative ($r$ = –1) correlation**: Once again, perfect correlations (positive or negative) exist only in mathematical formulas and direct physical or numerical



Example of a perfect negative correlation $(r = -1.0)$

relations. An example of a perfect negative correlation is based on the formula **distance = rate × time**. When driving from point A to point B, if you drive *twice* as fast, you will take *half* as long.

## 10.2 Additional Considerations

### 10.2.1 Linear versus Curvilinear

It is important to understand that the **Correlate** command measures only *linear* relationships. There are many relations that are not linear. For instance, nervousness before a major exam: Too much or too little nervousness generally hurts performance while a moderate amount of nervousness typically aids performance. The relation on a scatter plot would look like an inverted U, but computing a Pearson correlation would yield no relation or a weak relation. The chapters on simple regression and multiple regression analysis (Chapters 15 and 16) will consider curvilinear relationships in some detail. It is often a good idea to create a scatter plot of your data before computing correlations, to see if the relationship between two variables is linear. If it is linear, the scatter plot will more or less resemble a straight line. While a scatter plot can aid in detecting linear or curvilinear relationships, it is often true that significant correlations may exist even though they cannot be detected by visual analysis alone.

### 10.2.2 Significance and Effect Size

There are three questions that you, as a researcher, must answer about the $r$ statistic. The first question is, "Are you fairly certain that the strength of relationship tested by this statistic is real instead of random?" As with most other statistical procedures, a significance or probability (or $p$ value) is computed to determine the likelihood that a particular correlation could occur by chance. A significance less than .05 ($p < .05$) means that there is less than a 5% probability this relationship occurred by chance. SPSS has two different significance measures, one-tailed significance and two-tailed significance. To determine which to use, the rule of thumb generally followed is to use two-tailed significance when you compute a table of correlations in which you have little idea as to the *direction* of the correlations. If, however, you have prior expectations about the direction of correlations (positive or negative), then the statistic for one-tailed significance is generally used.

The second question that you need to ask about the $r$ statistic is, "How big is it?" For all inferential statistics you have to think about the effect size to answer this question. For most statistics, there is an additional effect size measure. With correlation, you are lucky: $r$ itself is both the test statistic *and* the measure of effect size. When $r$ is close to 0, the size of the relationship is small, and when $r$ is close to 1 or −1, the size of the relationship is large. The third question that you need to ask about the $r$ statistic is, "Is it important?" To answer that question, you need to consider whether you are fairly certain that the effect is not due to chance, how big the effect is, and how much you care (for theoretical or practical reasons). A small $r$ may not be important if one of your variables is length of beard but may be very important if one of your variables measures the cure for cancer.

### 10.2.3 Causality

Correlation does not necessarily indicate causation. Sometimes causation is clear. If height and weight are correlated, it is clear that additional height causes additional weight. Gaining weight is not known to increase one's height. Also the relationship between gender and empathy shows that women tend to be more empathic than men. If a man becomes more empathic this is unlikely to change his gender. Once again, the *direction* of causality is clear: gender influences empathy, not the other way around.

There are other settings where direction of causality is likely but open to question. For instance self-efficacy (the belief of one's ability to help) is strongly correlated with actual helping. It would generally be thought that belief of ability will influence how much one helps, but one could argue that one who helps more may increase their self-efficacy as a result of their actions. The former answer seems more likely but both may be partially valid.

Thirdly, sometimes it is difficult to have any idea of which causes which. Emotional stability and anxiety are strongly related (more emotionally stable people are less anxious). Does greater emotional stability result in less anxiety, or does greater anxiety result in lower emotional stability? The answer, of course, is yes. They both influence each other.

Finally there is the third variable issue. It is reliably shown that ice cream sales and homicides in New York City are positively correlated. Does eating ice cream cause one to become homicidal? Does committing murders give one a craving for ice cream? The answer is neither. Both ice cream sales and murders are correlated with heat. When the weather is hot more murders occur and more ice cream is sold. The same issue is at play with the reliable finding that across many cities the number of churches is positively correlated with the number of bars. No, it's not that church-going drives one to drink, nor is it that heavy drinking gives one an urge to attend church. There is again the third variable: population. Larger cities have more bars *and* churches while smaller cities have fewer of both.

### 10.2.4 Partial Correlation

We mention this issue because partial correlation is included as an option within the context of the **Correlate** command. We mention it only briefly here because it is covered in some detail in Chapter 14 in the discussion about covariance. Please refer to that chapter for a more detailed description of partial correlation. Partial correlation is the process of finding the correlation between two variables *after* the influence of other variables has been controlled for. If, for instance, we computed a correlation between GPA and total points earned in a class, we could include year as a covariate. We would anticipate that fourth-year students would generally do better than first-year students. By computing the partial correlation, that "partials out" the influence of year, we mathematically eliminate the influence of years of schooling on the correlation between total points and GPA. With the partial correlation option, you may include more than one variable as a covariate if there is reason to do so.

The file we use to illustrate the **Correlate** command is our example introduced in the first chapter. The file is called **grades.sav** and has an $N = 105$. This analysis computes correlations between five variables in the file: **gender**, previous GPA (**gpa**), the first and fifth quizzes (**quiz1**, **quiz5**), and the final exam (**final**).

## 10.3 Step by Step

### 10.3.1 Describing Subpopulation Differences

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | → All Programs ▷ → IBM SPSS Statistics → IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | ✳ 🚀 → ✳ ■ → ✳ Σ | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| | | Step 2 |
|---|---|---|
| | *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* | |

*Screens 1 and 2 (displayed on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **grades.sav** *file for further analyses:*

| In Screen | Do This | Step 3 |
|---|---|---|
| Front1 | ✳ **File** → **Open** → ✳ **Data** [*or* ✳ 📂] | |
| Front2 | **type** `grades.sav` → ✳ **Open** [*or* ✳ ✳ **grades.sav**] | Data |

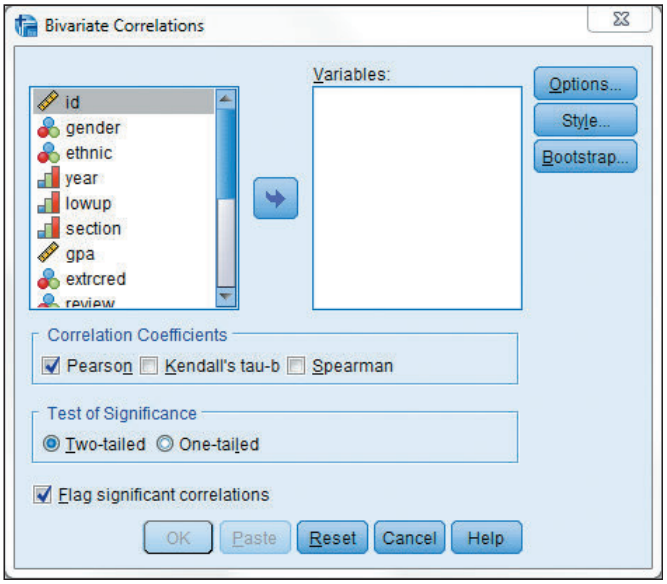
Whether first entering SPSS or returning from earlier operations the standard menu of commands across the top is required. As long as it is visible you may perform any analyses. It is not necessary for the data window to be visible.

*After completion of Step 3 a screen with the desired menu bar appears. When you click a command (from the menu bar), a series of options will appear (usually) below the selected command. With each new set of options, click the desired item. The sequence to access correlations begins at any screen with the menu of commands visible:*

| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | ✳ **Analyze** → **Correlate** → ✳ **Bivariate** | 10.1 |

After clicking **Bivariate**, a new window opens (Screen 10.1, below) that specifies a number of options available with the correlation procedure. First, the box to the left

## Screen 10.1 The Bivariate Correlations Window



Correlation Coefficients
☑ Pearson ☐ Kendall's tau-b ☐ Spearman

Test of Significance
● Two-tailed ○ One-tailed

☑ Flag significant correlations

lists all the *numeric* variables in the file (note the absence of **firstname**, **lastname**, and **grade** all *non*numeric). Moving variables from the list to the **Variables** box is similar to the procedures used in previous chapters. Click the desired variable in the list, click ➔, and that variable is pasted into the **Variables** box. This process is repeated for each desired variable. Also, if there are a number of consecutive variables in the variables list, you may click the first one and then press **shift** key and click the last one to select them all. Then a single click of ➔ will paste all highlighted variables into the active box.

In the next box labeled **Correlation Coefficients**, the **Pearson** *r* is selected by default. If your data are *not* normally distributed, then select **Spearman**. You may select both options and see how the values compare.

Under **Test of Significance**, **Two-tailed** is selected by default. Click on **One-tailed** if you have clear knowledge of the *direction* (positive or negative) of your correlations.

**Flag significant correlations** is selected by default and places an asterisk (*) or double asterisk (**) next to correlations that attain a particular level of significance (usually .05 and .01). Whether or not significant values are flagged, the correlation, the significance accurate to three decimals, and the number of subjects involved in each correlation will be included.

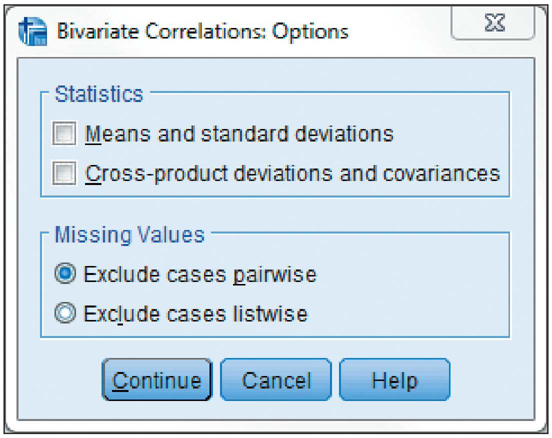
For analyses demonstrated in this chapter we will stick with the **Pearson** correlation, the **Two-tailed** test of significance, and also keep **Flag significant correlations**. If you wish other options, simply click the desired procedure to select or deselect before clicking the final **OK**. For sequences that follow, the starting point is always Screen 10.1. If necessary, perform whichever of Steps 1–4 (pages 141–143) are required to arrive at that screen.

*To produce a correlation matrix of* **gender**, **gpa**, **quiz1**, **quiz5**, *and* **final**, *perform the following sequence of steps*:

| In Screen | Do This | Step 5 |
|---|---|---|
| 10.1 | ✳ ✳ **gender** → ✳ ✳ **gpa** → ✳ ✳ **quiz1** → ✳ ✳ **quiz5** → | |
| | ✳ **final** → ✳ **OK** | Back1 |

Additional procedures are available if you click the **Options** button in the upper right corner of Screen 10.1. This window (Screen 10.2, below) allows you to select additional statistics to be printed and to deal with missing values in two different ways. **Means and standard deviations** may be included by clicking the appropriate option, as may **Cross-product deviations and covariances**.

**Screen 10.2** The Bivariate Correlations: Options Window



Bivariate Correlations: Options

Statistics

□ Means and standard deviations

□ Cross-product deviations and covariances

Missing Values

● Exclude cases pairwise

○ Exclude cases listwise

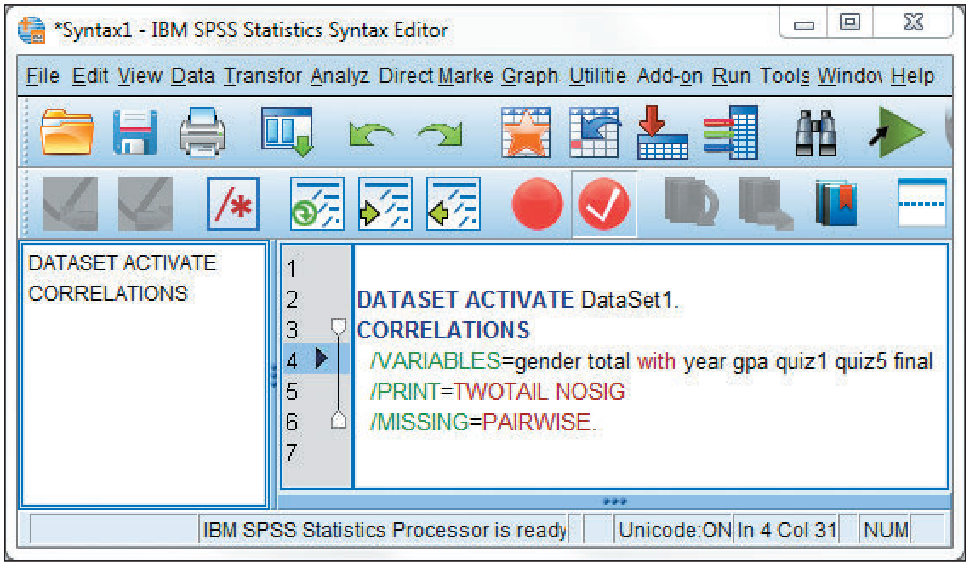
To **Exclude cases pairwise** means that for a particular correlation in the matrix, if a subject has one or two missing values for that comparison, then that subject's influence will not be included in that particular correlation. Thus correlations within a matrix may have different numbers of subjects determining each correlation. To **Exclude cases listwise** means that if a subject has *any* missing values, all data from that subject will be eliminated from any analyses. Missing values is a thorny problem in data analysis and should be dealt with before you get to the analysis stage. See Chapter 4 for a more complete discussion of this issue.

*The following procedure, in addition to producing a correlation matrix similar to that created in sequence Step 5, will print means and standard deviations in one table. Cross-product deviations and covariances will be included in the correlation matrix*:

| In Screen | Do This | Step 5a |
|---|---|---|
| 10.1 | ✳ ✳ **gender** → ✳ ✳ **gpa** → ✳ ✳ **quiz1** → ✳ ✳ **quiz5** → ✳ ✳ **final** → ✳ **Options** | |
| 10.2 | ✳ **Means and standard deviations** → ✳ **Cross-product deviations and covariances** → ✳ **Continue** | |
| 10.1 | ✳ **OK** | Back1 |

What we have illustrated thus far is the creation of a correlation matrix in which there are equal number of rows and columns. Often a researcher wishes to create correlations between one set of variables and another set of variables. For instance she may have created a 12 × 12 correlation matrix but wishes to compute correlations between 2 new variables and the original 12. The windows format does not allow this option and it is necessary to create a "command file," something familiar to users of the PC or mainframe versions of SPSS. If you attempt anything more complex than the sequence shown below, you will probably need to consult a book that explains SPSS syntax: See the Reference section, page 374.

**Screen 10.3** The SPSS Syntax Editor Window
[with the command file from Step 5b (in the following page) included]


*Syntax1 - IBM SPSS Statistics Syntax Editor
File Edit View Data Transfor Analyz Direct Marke Graph Utilitie Add-on Run Tools Windo Help
DATASET ACTIVATE
CORRELATIONS
IBM SPSS Statistics Processor is ready
Unicode:ON
In 4 Col 31
NUM

```
DATASET ACTIVATE DataSet1.
CORRELATIONS
  /VARIABLES=gender total with year gpa quiz1 quiz5 final
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

In previous editions of this book we had you type in your own command lines, but due to some changes in SPSS syntax structure, we find it simpler (and safer) to begin with the windows procedure and then switch to syntax. Recall that the goal is to create a 2 × 5 matrix of correlations between **gender** and **total** (the rows) and *with* **year**, **gpa**, **quiz1**, **quiz5**, *and* **final** (in the columns). Begin by creating what looks like a 7 × 7 matrix with variables in the order shown above. Then click the **Paste** button and a syntax screen (Screen 10.3, previous page) will appear with text very similar to what you see in the window of Screen 10.3. The only difference is that we have typed in the word "with" between **total** and **year**, allowing a single space before and after the word "with." All you need to do then is click the ▶ button (the arrow (➚) on Screen 10.3 identifies its location) and the desired output appears.

*You may begin this sequence from any screen that shows the standard menu of commands across the top of the screen. Below we create a two by five matrix of Pearson correlations comparing* **gender** *and* **total** *with* **year**, **gpa**, **quiz1**, **quiz5**, *and* **final**.

| In Screen | Do This | Step 5b |
|---|---|---|
| 10.1 | 🖱 🖱 **gender** → 🖱 🖱 **total** → 🖱 🖱 **year** → 🖱 🖱 **gpa** → 🖱 🖱 **quiz1** → 🖱 🖱 **quiz5** → 🖱 🖱 **final** → 🖱 **Paste** | |
| 10.3 | **type** `with` *in the* **/VARIABLES** *line between* "**total**" *and* "**year**" *leaving a single space before and behind the* "with" → 🖱 ▶ | Back1 |

When you use this format, simply replace the variables shown here by the variables you desire. You may have as many variables as you wish both before and after the "with."

Upon completion of Step 5, 5a, or 5b, the output screen will appear (Screen 1, inside back cover). All results from the just-completed analysis are included in the Output Navigator. Even when viewing output, the standard menu of commands is still listed across the top of the window. Further analyses may be conducted without returning to the data screen.

## 10.4 Printing Results

Results of the analysis (or analyses) that have just been conducted require a window that displays the standard commands (**File Edit Data Transform Analyze . . .**) across the top. A typical print procedure is shown below beginning with the standard output screen (Screen 1, inside back cover).

*To print results, from the Output screen perform the following sequence of steps:*

| In Screen | Do This | Step 6 |
|---|---|---|
| Back1 | *Select desired output or edit (see pages 16–22)* → 🖱 **File** → 🖱 **Print** | |
| 2.9 | *Consider and select desired print options offered, then* → 🖱 **OK** | |

*To exit you may begin from any screen that shows the* **File** *command at the top.*

| In Screen | Do This | Step 7 |
|---|---|---|
| Menu | 🖱 **File** → 🖱 **Exit** | 2.1 |

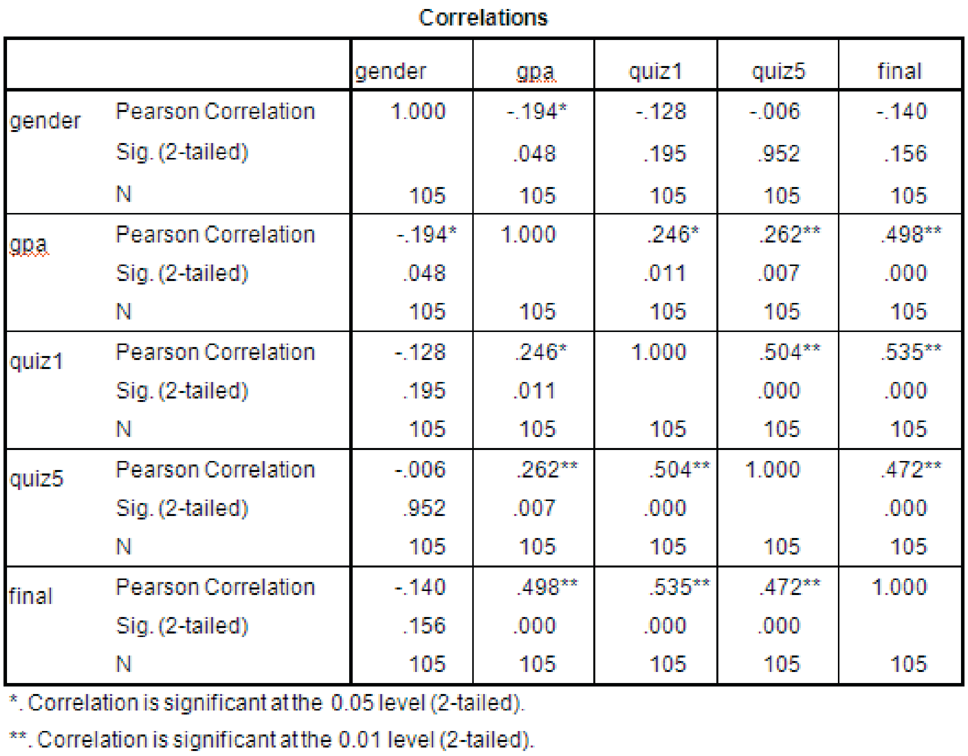
Note: After clicking **Exit**, there will frequently be small windows that appear asking if you wish to save or change anything. Simply click each appropriate response.

# 10.5 Output

## 10.5.1 Correlations

This output is from sequence Step 5 (page 144) with two-tailed significance selected and significant correlations flagged.

Notice first of all the structure of the output. The upper portion of each cell identifies the correlations between variables accurate to three decimals. The middle portion indicates the significance of each corresponding correlation. The lower portion records the number of subjects involved in each correlation. Only if there are missing values is it possible that the number of subjects involved in one correlation may differ from the number of subjects involved in others. The notes below the table identify the meaning of the asterisks and indicate whether the significance levels are one-tailed or two-tailed.



Correlations

| | | gender | gpa | quiz1 | quiz5 | final |
|---|---|---|---|---|---|---|
| gender | Pearson Correlation | 1.000 | -.194* | -.128 | -.006 | -.140 |
| | Sig. (2-tailed) | | .048 | .195 | .952 | .156 |
| | N | 105 | 105 | 105 | 105 | 105 |
| gpa | Pearson Correlation | -.194* | 1.000 | .246* | .262** | .498** |
| | Sig. (2-tailed) | .048 | | .011 | .007 | .000 |
| | N | 105 | 105 | 105 | 105 | 105 |
| quiz1 | Pearson Correlation | -.128 | .246* | 1.000 | .504** | .535** |
| | Sig. (2-tailed) | .195 | .011 | | .000 | .000 |
| | N | 105 | 105 | 105 | 105 | 105 |
| quiz5 | Pearson Correlation | -.006 | .262** | .504** | 1.000 | .472** |
| | Sig. (2-tailed) | .952 | .007 | .000 | | .000 |
| | N | 105 | 105 | 105 | 105 | 105 |
| final | Pearson Correlation | -.140 | .498** | .535** | .472** | 1.000 |
| | Sig. (2-tailed) | .156 | .000 | .000 | .000 | |
| | N | 105 | 105 | 105 | 105 | 105 |

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

The diagonal of 1.000s (some versions delete the ".000") shows that a variable is perfectly correlated with itself. Since the computation of correlations is identical regardless of which variable comes first, the half of the table above the diagonal of 1.000s has identical values to the half of the table below the diagonal. Note the moderately strong positive relationship between **final** and **quiz5** means we can be fairly certain is not due to chance ($r = .472$, $p < .001$). As described in the introduction of this chapter, these values indicate a positive relationship between the score on the fifth quiz and the score on the final. Those who scored higher on the fifth quiz tended to score higher on the final as well.

# Exercises

Answers to selected exercises are downloadable at **www.spss-step-by-step.net.**

1. Using the **grades.sav** file create a correlation matrix of the following variables: **id**, **ethnic**, **gender**, **year**, **section**, **gpa**, **quiz1**, **quiz2**, **quiz3**, **quiz4**, **quiz5**, **final**, **total**; select one-tailed significance; flag significant correlations. Print out results on a single page.
   - Draw a single line through the columns and rows where the correlations are meaningless.
   - Draw a double line through cells where correlations exhibit linear dependency.
   - Circle the one "largest" (greatest absolute value) NEGATIVE correlation (the *p* value will be less than .05) and explain what it means.
   - Box the three largest POSITIVE correlations (each *p* value will be less than .05) and explain what they mean.
   - Create a scatterplot of **gpa** by **total** and include the regression line (see Chapter 5, page 98 for instructions).
2. Using the **divorce.sav** file create a correlation matrix of the following variables: **sex**, **age**, **sep**, **mar**, **status**, **ethnic**, **school**, **income**, **avoicop**, **iq**, **close**, **locus**, **asq**, **socsupp**, **spiritua**, **trauma**, **lsatisy**; select one-tailed significance; flag significant correlations. Print results on a single page. Note: Use **Data Files** descriptions (page 364) for meaning of variables.
   - Draw a single line through the columns and rows where the correlations are meaningless.
   - Draw a double line through the correlations where there is linear dependency
   - Circle the three "largest" (greatest absolute value) NEGATIVE correlations (each *p* value will be less than .05) and explain what they mean.
   - Box the three largest POSITIVE correlations (each *p* value will be less than .05) and explain what they mean.
   - Create a scatterplot of **close** by **lsatisy** and include the regression line (see Chapter 5, page 98 for instructions).
   - Create a scatterplot of **avoicop** by **trauma** and include the regression line.

# Chapter 11
# The *t* Test Procedure

A *t* TEST is a procedure used for comparing sample means to see if there is sufficient evidence to infer that the means of the corresponding population distributions also differ. More specifically, for an independent-samples *t* test, a sample is taken from two populations. The two samples are measured on some variable of interest. A *t* test will determine if the means of the two sample distributions differ significantly from each other. *t* tests may be used to explore issues such as: Does treatment A yield a higher rate of recovery than treatment B? Does one advertising technique produce higher sales than another technique? Do men or women score higher on a measure of empathic tendency? Does one training method produce faster times on the track than another training method? The key word is *two*: *t* tests always compare *two* different means or values.

In its chapter on *t* tests, the *IBM SPSS Statistics Guide to Data Analysis* spends several pages talking about null hypotheses, populations, random samples, normal distributions, and a variety of research concerns. All its comments are germane and of considerable importance for conducting meaningful research. However, discussion of these issues goes beyond the scope of this book. The topic of this chapter is *t* tests: what they do, how to access them in SPSS, and how to interpret the results. We direct you to the manual if you wish to review some of those additional concerns.

## 11.1 Independent-Samples *t* Tests

SPSS provides three different types of *t* tests. The first type, the independent-samples *t* test, compares the means of two different samples. The two samples share some variable of interest in common, but there is no overlap between membership of the two groups. Examples include the difference between males and females on an exam score, the difference on number of pull-ups by Americans and Europeans, the difference on achievement test scores for a sample of low-SES first graders and a sample of high-SES first graders, or the difference of life-satisfaction scores between those who are married and those who are unmarried. Note again that there is no overlap of membership between the two groups.

## 11.2 Paired-Samples *t* Tests

The second type of *t* test, the paired-samples *t* test, is usually based on groups of individuals who experience both conditions of the variables of interest. Examples include students' scores on the first quiz versus the same students' scores on the second quiz; subjects' depression scores after treatment A as compared to the same subjects' scores after treatment B; a set of students' scores on the SAT compared to the same students' scores on the GRE several years later; elementary school students' achievement test percentiles after one year at a low-SES school as compared to their achievement test percentiles after one year at a high-SES school. Note here that the same group experiences both levels of the variable.

## 11.3 One-Sample *t* Tests

The third type of test is a one-sample *t* test. It is designed to test whether the mean of a distribution differs significantly from some preset value. An example: Does a course offered to college seniors result in a GRE score greater than or equal to 1200? Did the performance of a particular class differ significantly from the professor's goal of an 82% average? During the previous season, the mean time for the cross country athletes' best seasonal performances was 18 minutes. The coach set a goal of 17 minutes for the current season. Did the athletes' times differ significantly from the 17-minute goal set by the coach? In this procedure, the sample mean is compared to a single fixed value.

## 11.4 Significance and Effect Size

There are three questions that every researcher needs to ask about the difference between groups (or conditions, or the sample and population means). The first question is, "Can I be certain that the difference between groups (or between conditions, or between the sample mean and population mean) is not due to random chance?" The significance value (*p* value) answers this question. But in the case of a *t* test, you need to know whether to use a one-tailed or a two-tailed test of significance. The two-tailed test examines whether the mean of one distribution differs significantly from the mean of the other distribution, regardless of the direction (positive or negative) of the difference. The one-tailed test measures only whether the second distribution differs in a particular direction from the first. For instance, at a weight-loss clinic, the concern is only about the amount of weight *loss*. Any amount of weight gain is of no interest. Likewise, an advertising campaign concerns itself only with sales *increases*.

Usually the context of the research will make clear which type of test is appropriate. The only computational difference between the two is that the *p*-value of one is twice as much as the *p*-value of the other. If SPSS output produces a two-tailed significance value (this is the default), simply divide that number by two to give you the probability for a one-tailed test.

The second question that you need to ask about the difference tested with the *t* statistic is, "How big is the difference?" For a *t* test, the usual measure of effect size is the Cohen's *d* statistic, which is a measure of how many standard deviations apart the means are. SPSS does not calculate the *d*, but it is easy to calculate:

$$d = \frac{\text{Mean}_{\text{Group 1}} - \text{Mean}_{\text{Group 2}}}{\text{Standard Deviation}}$$

Note that, depending on which version of the *t* test you do, you may need to calculate the overall standard deviation of the two groups or conditions in your data.

The third question that you need to ask about the difference tested is, "Is this difference important?" For that, you need to look at the *p* value (if it's not significant, we can't be certain that it's real, much less important), the *d* effect size (bigger is more likely to be important), and you need to think about whether you think it's important or not. Is $d = .3$ (of a standard deviation) important? If that represents how much better treatment A improves hypertension than treatment B, for sure!

For demonstration of these commands we once more make use of the **grades.sav** file with $N = 105$. Variables of interest for this chapter include **gender**, **total** (total points in the class), **year** (first, second, third or fourth year in college), the quizzes **quiz1** to **quiz5**, and **percent** (the final class percent).

# 11.5 Step by Step

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | ✳ → ✳ All Programs ▷ → ✳ IBM SPSS Statistics → ✳ IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | ✳ → ✳ → ✳ | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| Step 2 |
|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* |

*Screens 1 and 2 (displayed on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **grades.sav** *file for further analyses:*

| In Screen | Do This | Step 3 |
|---|---|---|
| Front1 | ✳ **File** → **Open** → ✳ **Data** [*or* ✳ ] | |
| Front2 | **type** `grades.sav` → ✳ **Open** [*or* ✳ ✳ **grades.sav**] | Data |

Whether first entering SPSS or returning from earlier operations the standard menu of commands across the top is required. As long as it is visible you may perform any analyses. It is not necessary for the data window to be visible.

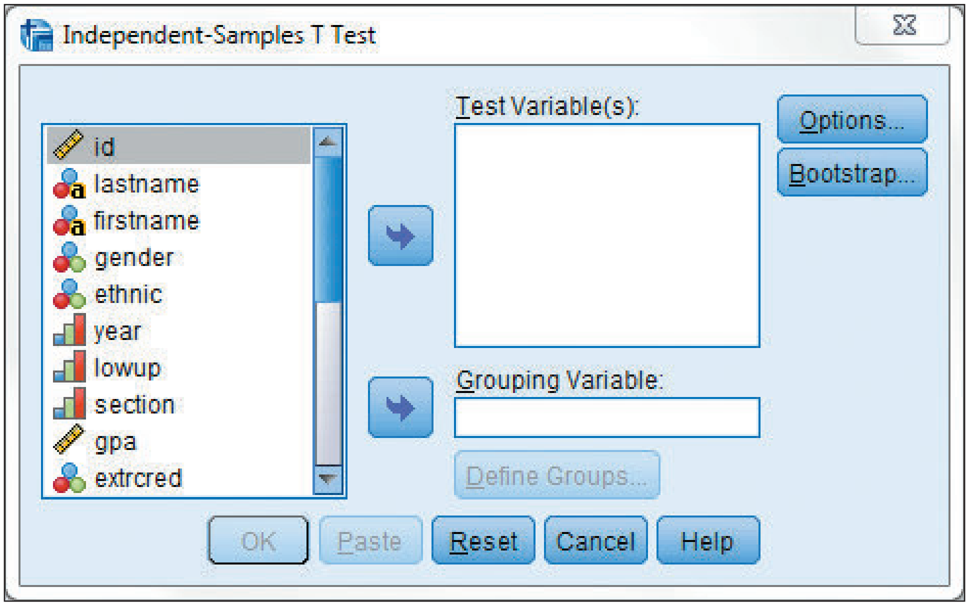
## 11.5.1 Independent-Samples *t* Tests

*From any screen that shows the menu of commands, select the following three options:*

| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | ✳ **Analyze** → **Compare Means** → ✳ **Independent-Samples T Test** | 11.1 |

At this point a new window opens (Screen 11.1, on the following page) that allows you to conduct an independent-samples *t* test. Note the structure of this screen. To the left is the list of variables; to the right is a box to indicate the **Test Variable(s)**. Test variables are the *continuous* variables (such as total points, final grade, etc.) for which we wish to make comparisons between two independent groups. A single variable or any number of variables may be included in this box. Below is the **Grouping Variable** box where the single variable that identifies the two groups will be indicated. This variable is usually discrete, that is, there are exactly two levels of the variable (such as gender or a pass/fail grade). It is possible, however, to use a variable with more than two levels (such as ethnicity—five levels, or grade—five levels) by indicating how you wish to divide the variable into exactly two groups. For instance, for grade, you might

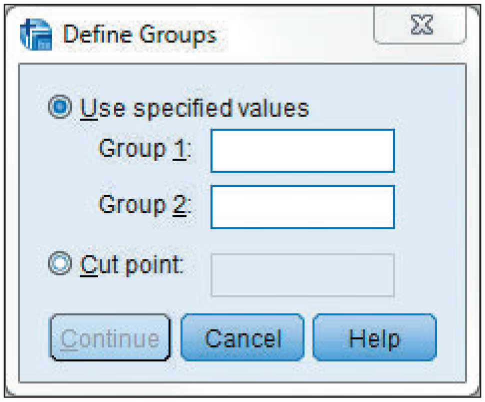
**Screen 11.1** The Independent-Samples *t* Test Window



compare As and Bs (as one group) with Cs, Ds, and Fs (as the other group). A continuous variable could even be included here if you indicate the number at which to divide subjects into exactly two groups.

Once you have identified the grouping variable, you next click on the **Define Groups** button. Even for a variable that has exactly two levels, it is necessary to identify the two levels of the grouping variable. At this point a new window opens (Screen 11.2, below) that, next to **Group 1**, allows you to designate the single number that identifies the first level of the variable (e.g., female = 1) and then, next to **Group 2**, the second level of the variable (e.g., male = 2). The **Cut point** option allows you to select a single dividing point for a variable that has more than two levels.

**Screen 11.2** The Define Groups Window



*To test whether there is a difference between males and females on* **total** *points earned:*

| In Screen | Do This | Step 5 |
|---|---|---|
| 11.1 | ✱ **total** → ✱ *upper* ➡ → ✱ **gender** → ✱ *lower* ➡ → ✱ **Define Groups** | |
| 11.2 | **press** **TAB** → **type** 1 → **press** **TAB** → **type** 2 → ✱ **Continue** | |
| 11.1 | ✱ **OK** | **Back1** |

*The* **year** *variable has four levels and the use of the* **Cut point** *option is necessary to divide this variable into exactly two groups. The number selected (3 in this case) divides the group into the selected value and larger (3 and 4) for the first group, and all values smaller than the selected value (1 and 2) for the second group. The starting point for this procedure is Screen 11.1. Do whichever of Steps 1–4 (page 151) are necessary to arrive at this screen.*

| In Screen | Do This | Step 5a |
|---|---|---|
| 11.1 | **total** → *upper* ➡ → **year** → *lower* ➡ → **Define Groups** | |
| 11.2 | **Cut point** → press **TAB** → type 3 → **Continue** | |
| 11.1 | **OK** | Back1 |

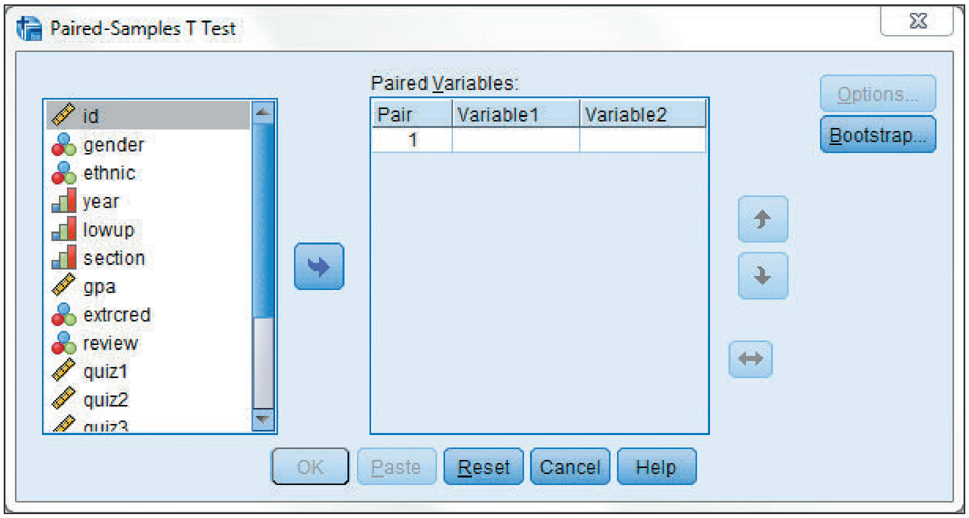
## 11.5.2 Paired-Samples *t* Tests

*From any screen that shows the menu of commands, select the following three options:*

| In Screen | Do This | Step 4a |
|---|---|---|
| Menu | **Analyze** → **Compare Means** → **Paired-Samples T Test** | 11.3 |

The procedure for paired-samples *t* tests is actually simpler than the procedure for independent samples. Only one window is involved and you do not need to designate levels of a particular variable. Upon clicking the **Paired-Samples T Test** option, a new screen appears (Screen 11.3, below). To the left is the now familiar list of variable names, and, since you will be comparing all subjects on two different variables (**quiz1** and **quiz2** in the first example) you need to designate both variables before clicking the ➡ in the middle of the screen. You may select as many pairs of variables as you would like to paste into the **Paired Variables** box before conducting the analysis. There are, however, no automatic functions or click and drag options if you wish to conduct many comparisons. You must paste them, a pair at a time, into the **Paired Variables** box.

**Screen 11.3** The Paired-Samples T Test Window



*To compare the distribution of scores on* **quiz1** *with scores on* **quiz2**, *the following sequence of steps is required.*

| In Screen | Do This | Step 5b |
|---|---|---|
| 11.3 | ✳ **quiz1** ⟶ *hold down* **Ctrl** *key and* ✳ **quiz2** ⟶ ✳ ➡ ⟶ ✳ **OK** | Back1 |

*If you wish to compute several t tests in the same setting, you will paste all desired pairs of variables into the* **Paired Variables** *box. In the sequence that follows, the scores on* **quiz1** *are compared with the scores on each of the other four quizzes (***quiz2** *to* **quiz5***). Screen 11.3 is the starting point for this sequence. Perform whichever of Steps 1–4 (page 151) to arrive at this screen. A click of* **Reset** *may be necessary to clear former variables.*

| In Screen | Do This | Step 5c |
|---|---|---|
| 11.3 | ✳ **quiz1** ⟶ *hold down* **Ctrl** *key and* ✳ **quiz2** ⟶ ✳ ➡ ⟶ | |
| | ✳ **quiz1** ⟶ *hold down* **Ctrl** *key and* ✳ **quiz3** ⟶ ✳ ➡ ⟶ | |
| | ✳ **quiz1** ⟶ *hold down* **Ctrl** *key and* ✳ **quiz4** ⟶ ✳ ➡ ⟶ | |
| | ✳ **quiz1** ⟶ *hold down* **Ctrl** *key and* ✳ **quiz5** ⟶ ✳ ➡ ⟶ ☝ **OK** | Back1 |

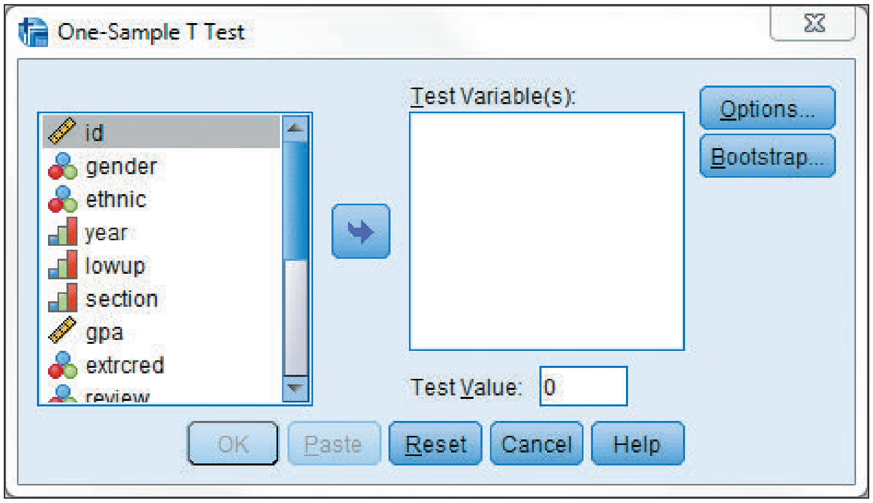
## 11.5.3 One-Sample *t* Tests

*From any screen that shows the menu of commands, select the following three options.*

| In Screen | Do This | Step 4b |
|---|---|---|
| Menu | ✳ **Analyze** ⟶ **Compare Means** ⟶ ✳ **One-Sample T Test** | 11.4 |

It is often desirable to compare the mean of a distribution with some objective standard. With the **grades.sav** file, the instructor may have taught the class a number of times and has determined what he feels is an acceptable mean value for a successful class. If the desired value for final **percent** is 85, he may wish to compare the present class against that standard. Does this class differ significantly from what he considers to be an acceptable class performance?

**Screen 11.4** The One-Sample T Test Window



Test Value: 0

Upon clicking the **One-Sample T Test** option, the screen that allows one-sample tests to be conducted appears (Screen 11.4, previous page). This simple procedure requires pasting variable(s) from the list of variables on the left into the **Test Variable(s)** box, typing the desired value into the box next to the **Test Value** label and then clicking the **OK**. For each variable selected this procedure will compare it to the designated value. Be sure that if you select several variables you want to compare them all to the *same* number. Otherwise, conduct separate analyses.

*To determine if the* **percent** *values for the entire class differed significantly from 85, conduct the following sequence of steps.*

| In Screen | Do This | Step 5d |
|---|---|---|
| 11.4 | ✳ **percent** → ✳ ➡ → **press TAB** → **type** 85 → ✳ **OK** | Back1 |

Upon completion of Step 5, 5a, 5b, 5c, or 5d, the output screen will appear (Screen 1, inside back cover). All results from the just-completed analysis are included in the Output Navigator. Make use of the scroll bar arrows (▲▼▶◀) to view the results. Even when viewing output, the standard menu of commands is still listed across the top of the window. Further analyses may be conducted without returning to the data screen.

## 11.6 Printing Results

Results of the analysis (or analyses) that have just been conducted require a window that displays the standard commands (**File Edit Data Transform Analyze . . .**) across the top. A typical print procedure is shown below beginning with the standard output screen (Screen 1, inside back cover).

*To print results, from the Output screen perform the following sequence of steps*:

| In Screen | Do This | Step 6 |
|---|---|---|
| Back1 | *Select desired output or edit (see pages 16–22)* → ✳ **File** → ✳ **Print** | |
| 2.9 | *Consider and select desired print options offered, then* → ✳ **OK** | |

*To exit you may begin from any screen that shows the* **File** *command at the top.*

| In Screen | Do This | Step 7 |
|---|---|---|
| Menu | ✳ **File** → ✳ **Exit** | 2.1 |

Note: After clicking **Exit**, there will frequently be small windows that appear asking if you wish to save or change anything. Simply click each appropriate response.

## 11.7 Output

In this section we present the three types of *t* tests separately, each under its own heading. A brief description of findings follows each output section; definition of terms (for all three sections) finishes the chapter. Output format for all three sets of results is slightly different (more space conserving) from that provided by SPSS.

### 11.7.1 Independent-Samples *t* Test:

What follows is complete output from sequence Step 5 on page 152.

| Dependent Variable | Gender | N | Mean | Std. Deviation | S.E. of Mean |
|---|---|---|---|---|---|
| total | Female | 64 | 102.03 | 13.896 | 1.737 |
| | Male | 41 | 98.29 | 17.196 | 2.686 |

**Independent-Samples *t* test**

| | Levene's Test for Equality of Variances | | | | | | | 95% Confidence Interval of the Difference | |
|---|---|---|---|---|---|---|---|---|---|
| Variances | F | Sig. | *t*-Values | df | Sig. (2-tailed) | Mean Difference | St. Error Difference | | |
| Equal | 2.019 | .158 | 1.224 | 103 | .224 | 3.739 | 3.053 | –2.317 | 9.794 |
| Unequal | | | 1.169 | 72.421 | .246 | 3.739 | 3.198 | –2.637 | 10.114 |

This independent-samples *t* test analysis indicates that the 64 females had a mean of 102.03 total points in the class, the 41 males had a mean of 98.29 total points in the class, and the means did not differ significantly at the $p < .05$ level (note: $p = .224$). Levene's test for Equality of Variances indicates variances for males and females do not differ significantly from each other (note: $p = .158$). This result allows you to use the slightly more powerful equal-variance *t* test. If Levene's test did show significant differences, then it would be necessary to use the unequal variance test. Definitions of other terms are listed on page 157.

### 11.7.2 Paired-Samples *t* Test:

Below is partial output (only the first comparison is included) from sequence Step 5b on page 154.

**Paired-Samples Statistics**

| Variables | N | Correlation | 2-tailed sig. | Mean | Std. Deviation | S.E. of Mean |
|---|---|---|---|---|---|---|
| quiz1 | 105 | .673 | .000 | 7.47 | 2.481 | .242 |
| quiz2 | | | | 7.98 | 1.623 | .158 |

**Paired-Samples Test**

| Paired Differences | | | | | | | |
|---|---|---|---|---|---|---|---|
| Mean | Std. Deviation | S.E. of Mean | 95% Conf. Interval | | *t*-value | df | 2-tailed sig. |
| –.514 | 1.835 | .179 | –.869 | –.159 | –2.872 | 104 | .005 |

This paired-samples *t* test analysis indicates that for the 105 subjects, the mean score on the second quiz ($M = 7.98$) was significantly greater at the $p < .01$ level (note: $p = .005$) than the mean score on the first quiz ($M = 7.47$). These results also indicate that a significant correlation exists between these two variables ($r = .673$, $p < .001$) indicating that those who score high on one of the quizzes tend to score high on the other.

### 11.7.3 One-Sample *t* Test:

What follows is complete output from sequence Step 5d on page 155.

**One-Sample Statistics**

| Variable | N | Mean | Std. Deviation | Std. Error of Mean |
|---|---|---|---|---|
| Percent | 105 | 80.381 | 12.1767 | 1.1883 |

One-Sample Test

| | Test Value = 85 | | | | | |
|---|---|---|---|---|---|---|
| | t | Df | Sig. (2-tailed) | Mean difference | 95% Confidence Interval | |
| Percent | −3.887 | 104 | .000 | −4.6190 | −6.976 | −2.263 |

This one-sample *t* test analysis indicates that the mean percent for this class ($N = 105$, $M = 80.38$) was significantly lower at the $p < .001$ level than the instructor's goal (Test Value) of 85%. The Mean Difference is simply the observed mean (80.38) minus the Test Value (85.0).

## 11.7.4 Definitions of Terms

| Term | Definition/Description |
|---|---|
| **STD. ERROR** | The standard deviation divided by the square root of *N*. This is a measure of stability or sampling error of the sample means. |
| **F-VALUE** | This value is used to determine if the variances of the two distributions differ significantly from each other. This is called a test of heteroschedasticity, a wonderful word with which to impress your friends. |
| **P =** (for Levene's test) | If variances do not differ significantly, then the *equal-variance estimate* may be used instead of the *unequal-variance estimate*. The *p*-value here, .158, indicates that the two variances do not differ significantly; so the statistically more powerful equal-variance estimate may be used. |
| **t-VALUES** | Based on either the equal-variance estimate equation or the unequal-variance estimate equation. Conceptually, both formulas compare the within-group deviations from the mean with the between-group deviations from the mean. The slightly larger (absolute values) equal-variance estimate may be used here because variances do not differ significantly. The actual *t* value is the difference between means divided by the standard error. |
| **df** (degrees of freedom) | For the equal-variance estimate, number of subjects minus number of groups (105 – 2 = 103). The fractional degrees of freedom (72.42) for the unequal-variance estimate is a formula-derived value. For the paired-samples and one-sample tests, the value is number of subjects minus 1 (105 – 1 = 104). |
| **2-TAIL SIG** | (associated with the *t* values) The probability that the difference in means could happen by chance. |
| **MEAN DIFFERENCE** | The difference between the two means. |
| **STD. DEVIATION** | This is the standard deviation of the difference, and it is used to calculate the *t* value for the paired *t* test. For each subject in a paired *t* test, there is a difference between the two quizzes (sometimes 0, of course). This particular statistic is the standard deviation of this distribution of mean-difference scores. |
| **CORRELATION** | Measures to what extent one variable varies systematically with another variable. The statistic presented here is the Pearson product-moment correlation, designated with an *r*. See Chapter 10 for a description of correlations. |
| **2-TAIL SIG** (of the correlation) | The probability that such a pattern could happen by chance. In the paired-samples test, the $r = .67$ and $p < .001$ indicate a substantial and significant correlation between quiz1 and quiz2. |
| **95% CI (CONFIDENCE INTERVAL)** | With *t* tests, the confidence interval deals with the difference-between-means value. If a large number of samples were drawn from a population, 95% of the mean differences will fall between the lower and upper values indicated. |

# Exercises

Answers to selected exercises are downloadable at **www.spss-step-by-step.net.**

For questions 1–7, perform the following operations:

a) Print out results

b) Circle the two mean values that are being compared.

c) Circle the appropriate significance value (be sure to consider equal or unequal variance).

d) For statistically significant results ($p$ < .05) write up each finding in standard APA format.

1. Using the **grades.sav** file, compare men with women (**gender**) for **quiz1**, **quiz2**, **quiz3**, **quiz4**, **quiz5**, **final**, and **total**.
2. Using the **grades.sav** file, determine whether the following pairings produce significant differences: **quiz1** with **quiz2**, **quiz1** with **quiz3**, **quiz1** with **quiz4**, and **quiz1** with **quiz5**.
3. Using the **grades.sav** file, compare the GPA variable (**gpa**) with the mean GPA of the university of 2.89.
4. Using the **divorce.sav** file, compare men with women (**sex**) for **lsatisfy**, **trauma**, **age**, **school**, **cogcope**, **behcope**, **avoicop**, **iq**, **close**, **locus**, **asq**, **socsupp**, and **spiritua**.
5. Using the **helping3.sav** file, compare men with women (**gender**) for **age**, **school**, **income**, **hclose**, **hcontrot**, **sympathi**, **angert**, **hcopet**, **hseveret**, **empathyt**, **effict**, **thelplnz**, **tqualitz**, and **tothelp**. See the **Data Files** section (page 364) for meaning of each variable.
6. Using the **helping3.sav** file, determine whether the following pairings produce significant differences: **sympathi** with **angert**, **sympathi** with **empathyt**, **empahelp** with **insthelp**, **empahelp** with **infhelp**, and **insthelp** with **infhelp**.
7. Using the **helping3.sav** file, compare the age variable (**age**) with the mean age for North Americans (33.0).
8. In an experiment, 10 participants were given a test of mental performance in stressful situations. Their scores were 2, 2, 4, 1, 4, 3, 0, 2, 7, and 5. Ten other participants were given the same test after they had been trained in stress-reducing techniques. Their scores were 4, 4, 6, 0, 6, 5, 2, 3, 6, and 4. Do the appropriate *t* test to determine if the group that had been trained had different mental performance scores than the group that had not been trained in stress reduction techniques. What do these results mean?
9. In a similar experiment, 10 participants who were given a test of mental performance in stressful situations at the start of the study were then trained in stress reduction techniques and were finally given the same test again at the end of the study. In an amazing coincidence, the participants received the same scores as the participants in question 8: The first two people in the study received a score of 2 on the pretest and a score of 4 on the posttest; the third person received a score of 4 on the pretest and 6 on the posttest; and so on. Do the appropriate *t* test to determine if there was a significant difference between the pretest and posttest scores. What do these results mean? How was this similar and how was this different than the results in question 1? Why?
10. You happen to know that the population mean for the test of mental performance in stressful situations is exactly three. Do a *t* test to determine whether the posttest scores in question 9 above (the same numbers as the training group scores in question 8) is significantly different than three. What do these results mean? How was this similar and how was this different than the results in question 9? Why?

# Chapter 12
# The One-Way ANOVA Procedure

ONE-WAY analysis of variance is obtained through the SPSS **One-Way ANOVA** command. While a one-way analysis of variance could also be accomplished using the **General Linear Model** command (Chapters 13 and 14), the **One-Way ANOVA** command has certain options not available in **General Linear Model**, such as planned comparisons of different groups or composites of groups, and is generally more straight forward.

## 12.1 Introduction to One-Way Analysis of Variance

Analysis of variance is a procedure used for comparing sample means to see if there is sufficient evidence to infer that the means of the corresponding population distributions also differ. One-way analysis of variance is most easily explained by contrasting it with *t* tests (Chapter 11). Although *t* tests compare only two distributions, analysis of variance is able to compare many. If, for instance, a sample of students takes a 10-point quiz and we wish to see whether women or men scored higher on this quiz, a *t test* would be appropriate. There is a distribution of women's scores and a distribution of men's scores, and a t test will tell if the means of these two distributions differ significantly from each other. If, however, we wished to see if any of five different ethnic groups' scores differed significantly from each other on the same quiz, it would require one-way analysis of variance to accomplish this. If we were to run such a test, **One-Way ANOVA** could tell us if there are significant differences within any of the comparisons of the five groups in our sample. Further tests (such as the Scheffé test, described in this chapter) are necessary to determine between *which* groups significant differences occur.

The previous paragraph briefly describes analysis of variance. What does the "one-way" part mean? Using the **One-Way ANOVA** command, you may have exactly *one* dependent variable (always continuous) and exactly *one* independent variable (always categorical). The independent variable illustrated above (ethnicity) is one variable, but it has several levels. In our example it has five: Native, Asian, Black, White, and Hispanic. **General Linear Model** (next two chapters) may also have a maximum of *one* dependent variable, but it may have two or more independent variables. In MANOVA, multivariate analysis of variance (Chapters 23 and 24), there may be multiple dependent variables *and* multiple independent variables.

The explanation that follows gives a conceptual feel for what one-way analysis of variance is attempting to accomplish. The mean (average) quiz scores for each of the ethnic groups are compared with each other: Natives with Asians, Natives with Blacks, Natives with Whites, Natives with Hispanics, Asians with Blacks, Asians with Whites, Asians with Hispanics, Blacks with Whites, Blacks with Hispanics, and Whites with Hispanics. **One-Way ANOVA** will generate a significance value indicating whether there are significant differences within the comparisons being made. This significance value answers the first question that every researcher must answer when using a one-way ANOVA: Can I be confident that the differences between all of the groups in my study is not due to random chance or error? Note that the *p* value (significance) does not indicate *where* the difference is or *what* the differences are, but a Scheffé test can identify which groups differ significantly from each other. Be aware that there are other tests besides Scheffé that are able to identify pairwise differences. Along with the Scheffé procedure, Tukey's honestly significant difference (HSD), least-significant difference (LSD), and Bonferroni are also popular tests of bivariate comparisons.

If the groups are significantly different, that does not indicate whether the difference between the groups are large or small. For that, an effect size measure is necessary. For ANOVA, the common effect size measures are eta ($\eta$) and eta squared ($\eta^2$). The bad news is, SPSS does not calculate these values. The good news is, they are very easy to calculate from the values given in the output:

$$\eta^2 = \frac{\text{Between Groups Sum of Squares}}{\text{Total Sum of Squares}}$$

Eta squared is a measure of the percentage of variance in the dependent variable accounted for by the independent variable and is similar to $R^2$ in linear regression (see Chapter 16). If you prefer, you may choose to report eta instead of eta squared (yes, just take the square root). Eta can be read roughly the same as a correlation ($r$).

Of course, if the $F$ value is significant (so that you know that the groups are probably really different than each other and any differences are unlikely to be due to chance), you then have to determine if the eta or eta-squared value is important. For that, you need to think about the theoretical or practical importance of the effect size.

The file we use to illustrate **One-Way ANOVA** is our familiar example. The file is called **grades.sav** and has an $N = 105$. This analysis contrasts scores on **quiz4** (the dependent variable) with five levels of **ethnic** (the independent variable)—Native, Asian, Black, White, and Hispanic.

# 12.2 Step by Step

## 12.2.1 One-Way Analysis of Variance

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | ✳ ⊞ → ✳ All Programs ▷ → ✳ IBM SPSS Statistics → ✳ Σ IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | ✳ 🚀 → ✳ ■ → ✳ Σα | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| | Step 2 |
|---|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* | |

*Screens 1 and 2 (displayed on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **grades.sav** *file for further analyses:*

| In Screen | Do This | Step 3 |
|---|---|---|
| Front1 | ✳ File → Open → ✳ Data [or ✳ 📂] | |
| Front2 | type `grades.sav` → ✳ Open [or ✳ ✳ grades.sav] | Data |

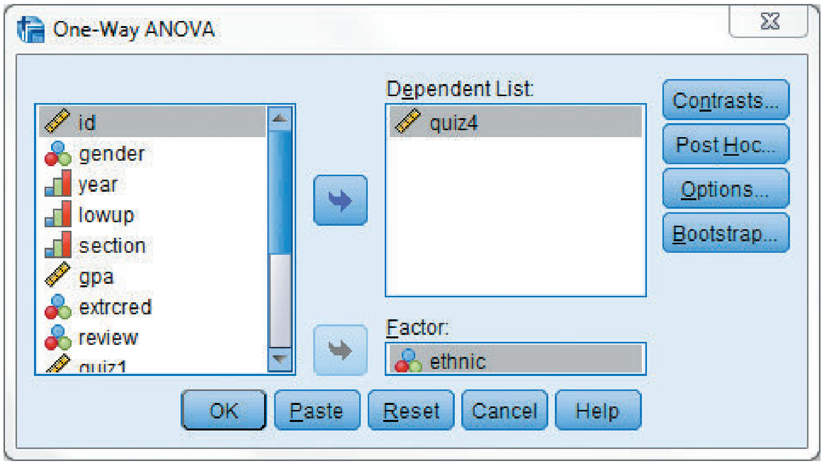
Whether first entering SPSS or returning from earlier operations the standard menu of commands across the top is required. As long as it is visible you may perform any analyses. It is not necessary for the data window to be visible.

*After completion of Step 3 a screen with the desired menu bar appears. When you click a command (from the menu bar), a series of options will appear (usually) below the selected command. With each new set of options, click the desired item. The sequence to access one-way analysis of variance begins at any screen with the menu of commands visible:*

| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | ✳ Analyze → Compare Means → ✳ One-Way ANOVA | 12.1 |

The initial screen that appears after clicking on **One-Way ANOVA** gives a clear idea of the structure of the command (Screen 12.1, below). Often all three of the options (**Options, Post Hoc**, and **Contrasts**) will be employed in conducting a one-way analysis of variance. First note the familiar list of variables to the left. Then note the large box near the top of the screen (titled **Dependent List**). In this box will go a single continuous variable (**quiz4** in this example), or several continuous variables. SPSS will print out separate analyses of variance results for each dependent variable included. Beneath is the **Factor** box. Here a single categorical variable will be placed (**ethnic** in our example). This analysis will compare the quiz scores for each of four ethnic groups. There are five ethnic groups listed under the **ethnic** variable, but because there were only five subjects in the Native category, we will include only the other four groups. See pages 73–75 to refresh your memory on selecting cases.

**Screen 12.1** The One-Way ANOVA window



Dependent List: quiz4

Factor: ethnic

In a retrogressive move the magnitude of the demise of the rumble seat, the death of the drive-in theater, SPSS 22 has eliminated the **Define Range** option. Now, instead of the simple two steps to select a range of values that represent a subset of a variable, you must go through the nine-step **Select Cases** procedure. As mentioned earlier, it would be desirable to eliminate the Native category from the analysis due to the presence of only five subjects. A sequence Step 4a is thus provided to remind you of that procedure (covered in Chapter 4). For the three Step-5 options that follow, if you wish to conduct the analysis for only four (rather than five) levels of ethnicity, then Step 4a must precede them.

*To select only the four levels of ethnicity described in the previous paragraph, perform the following sequence of steps. Screen 12.1 is an acceptable starting point.*

| In Screen | Do This | Step 4a |
|---|---|---|
| 12.1 | ✱ **Data** → ✱ **Select Cases** *(Note: this is in the menu bar across the top)* | |
| 4.8 | ✱ **If condition is satisfied** → ✱ **If** | |
| 4.3 | ✱ **ethnic** → ✱ ➡ → ✱ >= → **type** 2 → ✱ **Continue** | |
| 4.8 | ✱ **OK** | 12.1 |

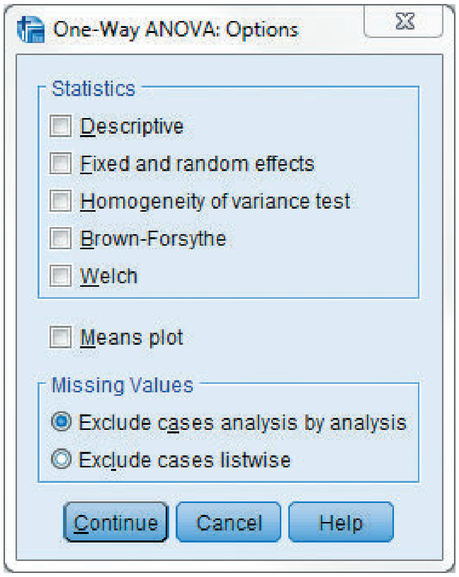
*To conduct a one-way analysis of variance to see if any of four ethnic groups differ on their* **quiz4** *scores, perform the following sequence*:

| In Screen | Do This | Step 5 |
|---|---|---|
| 12.1 | ✱ **quiz4** → ✱ *upper* ➡ → ✱ **ethnic** → ✱ *lower* ➡ → ✱ **OK** | Back1 |

Two additional options now present themselves as desirable. Although the prior analysis (Step 5, above) will tell you if a significant difference exists among the comparisons being made, it tells you little else. You don't know the mean values for each group, you have no information about the psychometric validity of your variables, and it is not possible to tell which groups differ from each other. The first omission is solved by clicking the **Options** button. Screen 12.2 opens allowing two important options. The **Descriptives** option provides means for each level, standard deviations, standard errors, 95% confidence limits, minimums, and maximums. The **Homogeneity-of-variance** selection also provides important psychometric information about the suitability of your variables for analysis. The **Means plot**, if selected, will produce a line graph that displays the mean of each category (each ethnic group in this case) in a plot-graph display.

A second important issue considers pairwise comparisons—that is, comparisons of each possible pair of levels of the categorical variable. For instance in our ethnic comparisons, we are interested in whether one group scores significantly higher than another: scores for Whites versus scores for Asians, Blacks versus Hispanics, and so forth. The title of the window (Screen 12.3, following page) is **Post Hoc Multiple Comparisons**. *"Post Hoc"*

**Screen 12.2** The One-Way ANOVA: Options window



One-Way ANOVA: Options

Statistics

□ Descriptive

□ Fixed and random effects

□ Homogeneity of variance test

□ Brown-Forsythe

□ Welch

□ Means plot

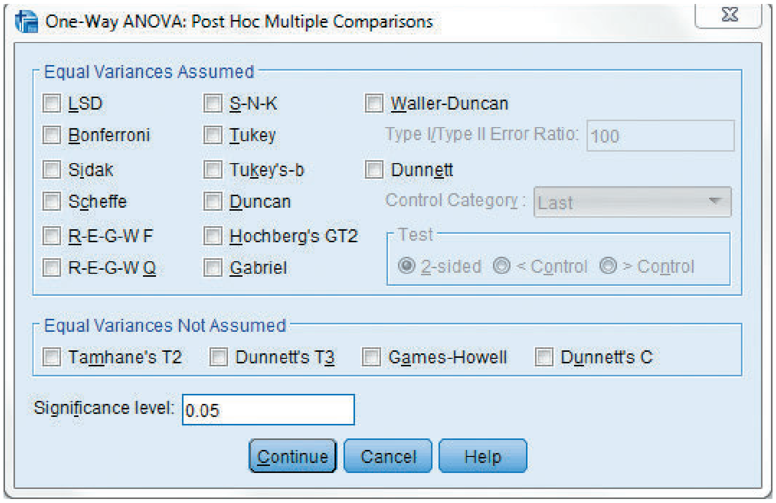
Missing Values

● Exclude cases analysis by analysis

○ Exclude cases listwise

means after the fact. “Multiple comparisons“ means that all possible pairs of factors are compared. There are 14 options if equal variances for levels of the variable are assumed, and an additional 4 if equal variances are not assumed. The number of test options is beyond bewildering. We have never met anyone who knew them all, or even wanted to know them all. The **LSD** (least significant difference) is the most liberal of the tests (that is, you are most likely to show significant differences in comparisons) because it is simply a series of *t* tests. **Scheffe** and **Bonferroni** are probably the most conservative of the set. **Tukey** (HSD—honestly significant difference) is another popular option.

**Screen 12.3** The One-Way ANOVA: Post Hoc Multiple Comparisons window



One-Way ANOVA: Post Hoc Multiple Comparisons

Equal Variances Assumed

| | | |
|---|---|---|
| □ LSD | □ S-N-K | □ Waller-Duncan |
| □ Bonferroni | □ Tukey | Type I/Type II Error Ratio: 100 |
| □ Sidak | □ Tukey's-b | □ Dunnett |
| □ Scheffe | □ Duncan | Control Category: Last |
| □ R-E-G-W F | □ Hochberg's GT2 | Test |
| □ R-E-G-W Q | □ Gabriel | ● 2-sided ○ < Control ○ > Control |

Equal Variances Not Assumed

□ Tamhane's T2 □ Dunnett's T3 □ Games-Howell □ Dunnett's C

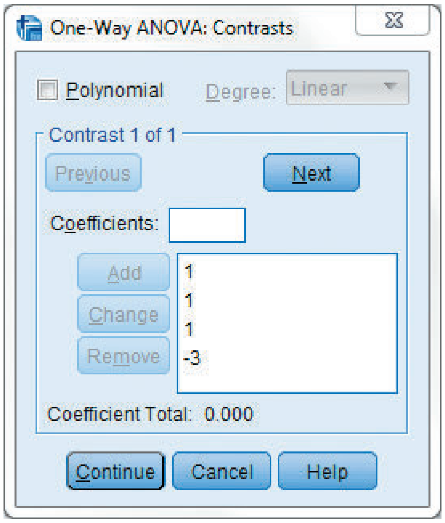
Significance level: 0.05

*The procedure that follows will conduct a one-way ANOVA, request descriptive statistics, measures of heteroschedasticity (differences between variances), and select the least significant difference* (**LSD**) *method of testing for pairwise comparisons. As before, the categorical variable will be* **ethnic**, *and the dependent variable,* **quiz4**.

| In Screen | Do This | Step 5a |
|---|---|---|
| 12.1 | **quiz4** → *upper* ➡ → **ethnic** → *lower* ➡ → **Options** | |
| 12.2 | **Descriptive** → **Homogeneity of variance test** → **Continue** | |
| 12.1 | **Post Hoc** | |
| 12.3 | **LSD** *(least significant difference)* → **Continue** | |
| 12.1 | **OK** | Back1 |

The final option on the initial screen involves **Contrasts**. Upon clicking this button, Screen 12.4 appears (top of next page). This procedure allows you to compare one level of the categorical variable with another (e.g., Asians with Hispanics), one level of the variable with the composite of others (e.g., Whites with non-Whites), or one composite with another composite (e.g., a White/Hispanic group with a Black/Asian group). Once the groups are identified, **Contrasts** computes a *t* test between the two groups. In this procedure, levels of a variable are coded by their value label. Present coding is Asians = 2, Blacks = 3, Whites = 4, and Hispanics = 5.

**Screen 12.4** The One-Way ANOVA: Contrasts Window



In the Coefficients boxes, you need to insert numbers that contrast positive numbers in one group with negative numbers in another group. It is always required that these coefficient numbers sum to zero. For instance, a Hispanic-Asian comparison could be coded (−1 0 0 1), Whites with non-Whites contrast (1 1 −3 1), and levels 2 and 3 contrasted with levels 4 and 5 (−1 −1 1 1). Note that each of these sequences sums to zero. The comparison shown in the screen at left contrasts Hispanics (the fifth group) with non-Hispanics (groups 2, 3, and 4).

The procedure includes typing the number into the **Coefficients** box that represents the first level of the categorical variable, clicking the **Add**, then typing the number that represents the second level of the variable and clicking the **Add** button and so on until all levels of the variable have been designated. Then, if you wish another contrast, click the **Next** button to the right of **Contrast 1 of 1** and repeat the procedure. The sequence that follows not only includes all the selections in Step 5a but also adds contrasts of Whites with Asians and Hispanics with non-Hispanics on **quiz4** scores.

*The contrasts selected below compare scores on the fourth quiz of (1) Whites and Asians with Hispanics and Blacks and (2) non-Hispanics with Hispanics. The starting point: Screen 12.1.*

| In Screen | Do This | Step 5b |
|---|---|---|
| 12.1 | ✳ **quiz4** → ✳ *upper* ➡ → ✳ **ethnic** → ✳ *lower* ➡ → ✳ **Options** | |
| 12.2 | ✳ **Descriptive** → ✳ **Homogeneity of variance test** → ✳ **Continue** | |
| 12.1 | ✳ **Post Hoc** | |
| 12.3 | ✳ **LSD** *(least significant difference)* → ✳ **Continue** | |
| 12.1 | ✳ **Contrasts** | |
| 12.4 | **press** **TAB** → **type** 1 → ✳ **Add** → | |
| | **press** **SHIFT-TAB** → **type** −1 → ✳ **Add** → | |
| | **press** **SHIFT-TAB** → **type** 1 → ✳ **Add** → | |
| | **press** **SHIFT-TAB** → **type** −1 → ✳ **Add** → ✳ **Next** | |
| | ✳ *in box to the right of* **Coefficients** → **type** 1 → ✳ **Add** → | |
| | **press** **SHIFT-TAB** → **type** 1 → ✳ **Add** → | |
| | **press** **SHIFT-TAB** → **type** 1 → ✳ **Add** → | |
| | **press** **SHIFT-TAB** → **type** −3 → ✳ **Add** → ✳ **Continue** | |
| 12.1 | ✳ **OK** | **Back1** |

Upon completion of Step 5, 5a, or 5b, the output screen will appear (Screen 1, inside back cover). All results from the just-completed analysis are included in the Output Navigator. Make use of the scroll bar arrows (▲▼▶◀) to view the results.

Even when viewing output, the standard menu of commands is still listed across the top of the window. Further analyses may be conducted without returning to the data screen.

## 12.3 Printing Results

Results of the analysis (or analyses) that have just been conducted require a window that displays the standard commands (**File Edit Data Transform Analyze...**) across the top. A typical print procedure is shown below beginning with the standard output screen (Screen 1, inside back cover).

*To print results, from the Output screen perform the following sequence of steps*:

| In Screen | Do This | Step 6 |
|---|---|---|
| Back1 | *Select desired output or edit (see pages 16–22)* → ✱ **File** → ✱ **Print** | |
| 2.9 | *Consider and select desired print options offered, then* → ✱ **OK** | |

*To exit you may begin from any screen that shows the* **File** *command at the top.*

| In Screen | Do This | Step 7 |
|---|---|---|
| Menu | ✱ **File** → ✱ **Exit** | 2.1 |

Note: After clicking **Exit**, there will frequently be small windows that appear asking if you wish to save or change anything. Simply click each appropriate response.

## 12.4 Output

### 12.4.1 One-Way Analysis of Variance

This is somewhat condensed output from sequence Step 5b from page 164. (Don't forget Step 4a on page 162.)

**Descriptives**

| quiz 4 | | | Std. | Std. | 95% CI | | |
|---|---|---|---|---|---|---|---|
| **Ethnic** | **N** | **Mean** | **Deviation** | **Error** | **for Mean** | **Minimum** | **Maximum** |
| Asian | 20 | 8.35 | 1.53 | .34 | 7.63 to 9.07 | 6 | 10 |
| Black | 24 | 7.75 | 2.13 | .44 | 6.85 to 8.65 | 4 | 10 |
| White | 45 | 8.04 | 2.26 | .34 | 7.73 to 8.72 | 2 | 10 |
| Hispanic | 11 | 6.27 | 3.32 | 1.00 | 4.04 to 8.50 | 2 | 10 |
| Total | 100 | 7.84 | 2.29 | .23 | 7.39 to 8.29 | 2 | 10 |

| Term | Definition/Description |
|---|---|
| **N** | Number of subjects in each level of ethnic. |
| **MEAN** | Average score for each group. |
| **STANDARD DEVIATION** | The standard measure of variability around the mean. |
| **STANDARD ERROR** | Standard deviation divided by square root of *N*. |
| **95% CI (CONFIDENCE INTERVAL) FOR MEAN** | Given a large number of samples drawn from a population, 95% of the means for these samples will fall between the lower and upper values. These values are based on the *t* distribution and are approximately equal to the mean ± 2 × the standard error. |
| **MINIMUM/MAXIMUM** | Smallest and largest observed values for that group. |

**Test of Homogeneity of Variance**

| | Levene Statistic | df1 between groups | df2 within groups | Significance |
|---|---|---|---|---|
| quiz4 | 5.517 | 3 | 96 | .002 |

Levene's test for homogeneity of variance with a significance value of .002 indicates that variances for **quiz4** scores for each of the ethnic groups do indeed differ significantly. Note that these values vary between a narrow variance for Asians of $1.53^2$ (= 2.34) to a much wider variance for Hispanics of $3.32^2$ (= 11.02). Most researchers, upon seeing this result, would check the distributions for measures of normality (skewness and kurtosis), and if they found nothing unusual, would probably ignore these results and accept the ANOVA analysis as valid. These measures of homogeneity of variance act more as a warning than as a disqualifier. However, in the contrast coefficient matrices (below), you will use the slightly less powerful estimate based on assumption of unequal variances.

**ANOVA**

| quiz 4 | Sum of Squares | df | Mean Square | F | Sig |
|---|---|---|---|---|---|
| Between groups | 34.297 | 3 | 11.432 | 2.272 | .085 |
| Within groups | 483.143 | 96 | 5.033 | | |
| Total | 517.440 | 99 | | | |

The key interpretive element of interest in the original ANOVA table is that, based on a $p$ = .085, a marginally significant difference (or differences) exists within comparisons of **quiz4** scores among the four ethnic groups. However, with just marginal significance sometimes we find significant pairwise post hoc differences. Definitions of terms in the ANOVA table follow.

| Term | Definition/Description |
|---|---|
| **WITHIN-GROUPS SUM OF SQUARES** | The sum of squared deviations between the mean for each group and the observed values of each subject within that group. |
| **BETWEEN-GROUPS SUM OF SQUARES** | The sum of squared deviations between the grand mean and each group mean weighted (multiplied) by the number of subjects in each group. |
| **BETWEEN-GROUPS DF** | Number of groups minus one. |
| **WITHIN-GROUPS DF** | Number of subjects minus number of groups minus one. |
| **MEAN SQUARE** | Sum of squares divided by degrees of freedom. |
| **F RATIO** | Between-groups mean square divided by within-groups mean square. |
| **SIGNIFICANCE** | The probability of the observed value happening by chance. The result here indicates that there is/are marginally significant difference(s) between means of the four groups as noted by a probability value of .085. |

**Contrast Coefficients**

| | ETHNIC | | | |
|---|---|---|---|---|
| Contrast | Asian | Black | White | Hispanic |
| 1 | 1 | –1 | 1 | –1 |
| 2 | 1 | 1 | 1 | –3 |

**Contrast Tests**

| | | Contrast | Value of Contrast | Std. Error | t | df | Significance 2-tailed |
|---|---|---|---|---|---|---|---|
| quiz4 | Assume equal variances | 1 | 2.37 | 1.015 | 2.336 | 96 | .022 |
| | | 2 | 5.33 | 2.166 | 2.459 | 96 | .016 |
| | Do not assume equal variances | 1 | 2.37 | 1.192 | 1.989 | 19.631 | .061 |
| | | 2 | 5.33 | 3.072 | 1.734 | 10.949 | .111 |

The first chart simply restates the contrasts typed in earlier. Two types of *t* comparisons are made: Under the equal-variance estimates, both contrasts are significant: between Asians/Whites and Hispanics/Blacks ($p$ = .022); between Hispanics and non-Hispanics ($p$ = .016). For the unequal-variance estimate, however, neither contrast achieves significance. Since variances *do* differ significantly, we should accept the unequal-variance estimate as valid, thus, no significant differences. Terms in this portion of the analysis are defined below.

| Term | Definition/Description |
|---|---|
| **CONTRAST** | Identifies which contrast (from the chart above) is being considered. |
| **VALUE OF CONTRAST** | A weighted number used in computation. |
| **STANDARD ERROR** | Standard deviation divided by square root of *N*. |
| **T-VALUES** | For either the equal- or unequal-variance estimate, *t* is determined by the **VALUE** divided by the standard error. |
| **DF (DEGREES OF FREEDOM)** | Number of subjects minus number of groups for the equal-variance estimate. It is a little-known formula that computes the fractional degrees of freedom value for the unequal-variance estimate. |
| **SIGNIFICANCE (2-TAILED)** | The likelihood that these values would happen by chance. The results indicate that, for scores on **quiz4**, for the unequal-variance estimates, neither contrast achieves significance. |

**Post Hoc Tests: Multiple Comparisons (condensed)**

| Ethnic (I) vs. | Ethnic (J) | Mean (I – J) Difference | Std. Error | Significance | 95% Confidence Interval |
|---|---|---|---|---|---|
| Asian | Black | .600 | .679 | .379 | –.75 to 1.95 |
| | White | .306 | .603 | .613 | –.89 to 1.50 |
| | Hispanic | 2.077* | .842 | .015 | .41 to 3.75 |
| Black | White | –.294 | .567 | .605 | –1.42 to .83 |
| | Hispanic | 1.477 | .817 | .074 | –.14 to 3.10 |
| White | Hispanic | 1.772* | .755 | .021 | .27 to 3.27 |

The mean value (average score for **quiz4**) for each of the four groups is listed in the previous chart. The asterisks (*) indicate there are two pairs of groups whose means differ significantly (at the $p < .05$ level) from each other: According to these fictional data, Asians ($M = 8.35$) and Whites ($M = 8.04$) both scored significantly higher on **quiz4** than did Hispanics ($M = 6.27$). Note the associated significance values of .015 and .021. The fact that the overall ANOVA results showed only marginal significance ($p$ = .085) and that pairwise comparisons yield two differences that are statistically significant is because the overall ANOVA compares all values simultaneously (thus weakening statistical power) while the LSD procedure is just a series of independent *t* tests (increasing experimentwise error—the chance that some of the significant differences are only due to chance).

# Exercises

Answers to selected exercises are downloadable at **www.spss-step-by-step.net.**

Perform one-way ANOVAs with the specifications listed below. If there are significant findings write them up in APA format (or in the professional format associated with your discipline). Examples of correct APA format are shown on the website. Further, notice that the final five problems make use of the **helping3.sav** data file. This data set (and all data files used in this book) is also available for download at the website listed above. For meaning and specification of each variable, make use of **Data Files** section of this book beginning on page 362.

1. File: **grades.sav**; dependent variable: **quiz4**; factor: **ethnic** (2,5) (note: this means use levels 2, 3, 4, 5, and exclude level 1); use **LSD** procedure for post hoc comparisons, compute two planned comparisons. This problem asks you to reproduce the output on pages 165–167. Note that you will need to perform a select-cases procedure (see page 73) to delete the "**1 = Native**" category.
2. File: **helping3.sav**; dependent variable: **tothelp**; factor: **ethnic** (1,4); use **LSD** procedure for post hoc comparisons, compute two planned comparisons.
3. File: **helping3.sav**; dependent variable: **tothelp**; factor: **problem** (1,4); use **LSD** procedure for post hoc comparisons, compute two planned comparisons.
4. File: **helping3.sav**; dependent variable: **angert**; factor: **occupat** (1,6); use **LSD** procedure for post hoc comparisons, compute two planned comparisons.
5. File: **helping3.sav**; dependent variable: **sympathi**; factor: **occupat** (1,6); use **LSD** procedure for post hoc comparisons, compute two planned comparisons.
6. File: **helping3.sav**; dependent variable: **effict**; factor: **ethnic** (1,4); use **LSD** procedure for post hoc comparisons, compute two planned comparisons.

# Chapter 13

# General Linear Model: Two-Way ANOVA

THIS CHAPTER describes a simple two-way (or two factor) analysis of variance (ANOVA). A two-way ANOVA is a procedure that designates a single dependent variable (always continuous) and utilizes exactly two independent variables (always categorical) to gain an understanding of how the independent variables influence the dependent variable. This operation requires the use of the **General Linear Model** → **Univariate** commands because the **One-Way ANOVA** command (Chapter 12) is capable of conducting only one-way analyses (that is, analyses with one independent variable).

The three ANOVA chapters in this book (Chapters 12, 13, and 14) should be read sequentially. A thorough grasp of one-way ANOVA (Chapter 12) is necessary before two-way ANOVA (this chapter) can be understood. This paves the way for Chapter 14 that considers three-way ANOVA and the influence of covariates. The present chapter is (intentionally) the simplest of the three; it includes a simple two-factor hierarchical design and nothing more. Yet understanding this chapter will provide the foundation for the considerably more complex Chapter 14. ANOVA is an involved process, and even in Chapter 14 we do not consider all the options that SPSS provides. Please refer to the *IBM SPSS Statistics 19 Guide to Data Analysis* for a more complete description of additional choices.

One "problem" when conducting ANOVA is that it is easy to get a statistical package such as SPSS to conduct a two-way, three-way, or even eight-way analysis of variance. The Windows format makes it as simple as clicking the right buttons; the computer does all the arithmetic, and you end up with reams of output of which you understand very little. The ease of calculating analysis of variance on a computer has a tendency to mask the fact that a successful study requires many hours of careful planning. Also, although a one-way ANOVA is straightforward and simple to interpret, a two-way ANOVA requires some training and frequently involves a thorough examination of tables and charts before interpretation is clear. Understanding a three-way ANOVA usually requires an experienced researcher, and comprehending a four-way ANOVA extends beyond the abilities of most humans.

The present chapter focuses on the fundamentals of how to conduct a hierarchical analysis of variance with a single dependent variable (**total**) and two independent variables or factors (**gender** and **section**), and then explains how to graph the cell means and interpret the results.

## 13.1 Statistical Power

One statistical concept needs to be introduced here, as it is not available for the previous procedures but is often useful for ANOVA: **Power**. Power is simply the probability of finding a significant difference in a sample, given a difference (between groups) of a particular size and a specific sample size. Often power is calculated before conducting a study. If you know the expected magnitude of a difference between groups, you can (for example) calculate how large a sample you need to be 80% sure of finding a significant difference between groups.

SPSS does not calculate power before you start a study, but it can calculate **observed power**. Observed power is the likelihood of finding a significant difference between groups in any particular sample with the sample size in your study, assuming that the differences between groups you find in your sample is the same size as the differences between groups in the population. (Feel free to read that sentence again.) This can be helpful in understanding what it means when a difference between groups does not reach statistical significance. For example, if you do not find a significant difference between groups and the observed power is .25, then there would only be a 25% chance of finding a significant difference with your sample size. You need to consider whether the observed difference between groups is big or small, depending on other research findings, pilot studies, or what is theoretically interesting.

## 13.2 Two-Way Analysis of Variance

As described in the last chapter, analysis of variance attempts to find significant differences between groups (or populations) by comparing the means of those groups on some variable of interest. To assist in understanding two-way ANOVA, we'll briefly summarize a one-way analysis. In the one-way analysis of variance described in the previous chapter, we sought to discover if four ethnic groups differed from each other on their performances for **quiz4**. The *one-way* part of the term indicates that there is exactly one independent variable (**ethnic** in the Chapter 12 example), and the ANOVA part (as opposed to MANOVA) indicates that there is exactly one dependent variable as well (**quiz4**). For conducting a one-way analysis of variance, the **One-Way ANOVA** command is often preferable to the **General Linear Model** command. While **General Linear Model** is able to conduct two-way, three-way, or higher-order analyses, the **One-Way ANOVA** command is simpler.

The sample we will use to demonstrate two-way analysis of variance will again be the **grades.sav** file. The **total** variable (total points in the class) will serve as the dependent variable. The independent variables will be **gender** and **section**. We will be attempting to discover if **gender**, or **section**, or a **gender** by **section** interaction has an effect on performance in the class. The typical research questions of interest are:

1. Is there a *main effect for* **gender**? That is, do females and males differ significantly in number of points earned, and which group is higher than the other?
2. Is there a *main effect for* **section**? Do students in the three sections differ significantly in points scored in the class, and which section scored higher than which?
3. Is there a **gender** *by* **section** *interaction*? Is the influence of the two variables idiosyncratic, such that in this case, gender has one effect in a particular section but a different effect in a different section? For example: Perhaps females score significantly higher in section 1 than males, but males score significantly higher than females in section 3. Interactions are often tricky to explain, and producing a plot displaying the cell means (shown in the output section) often helps to clarify.

SPSS can print cell means for all combinations of variables, and compute *F* values and associated significance values to indicate how confident you can be that the differences you are examining are not due to random chance. These values will indicate if there are significant main effects and/or if there are significant interactions between variables. It is quite possible to have significant main effects but not to have a significant interaction. The reverse is also possible. The output section of this chapter clarifies some of these issues.

Of course in addition to asking if you are confident that the effect you are examining is real, you will also probably want to ask how big the effect is. For this, there are two different effect sizes that you may want to use:

- Eta squared is a measure of the percentage of variance in the dependent variable accounted for by the independent variable, and is similar to $R^2$ in linear regression. SPSS does not calculate eta squared, but it is easy to calculate. For every $F$ and $p$ value, you can calculate eta squared with simple division:

$$\eta^2 = \frac{\text{Between Groups Sum of Squares}}{\text{Total Sum of Squares}}$$

- SPSS can calculate partial eta squared, which is the percentage of the variance in the dependent variable accounted for by the independent variable when removing the effect of all of the other independent variables.

Whether you choose to use eta squared or partial eta squared will depend both on what other people in your discipline use, and on what kind of comparisons you wish to make. In general, eta squared is more useful for comparing effect sizes within a study, and partial eta squared is more useful for comparing effect sizes across studies (especially if not all the studies are examining the same variables).

## 13.3 Step by Step

### 13.3.1 Two-Way Analysis of Variance

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | ✳ ⊞ → ✳ All Programs ▷ → ✳ IBM SPSS Statistics → ✳ IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | ✳ 🚀 → ✳ ■ → ✳ Σα | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| Step 2 |
|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* |

*Screens 1 and 2 (displayed on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **grades.sav** *file for further analyses:*

| In Screen | Do This | Step 3 |
|---|---|---|
| Front1 | ✳ File → Open → ✳ Data [*or* ✳ 📂] | |
| Front2 | type `grades.sav` → ✳ Open [*or* ✳ ✳ **grades.sav**] | Data |

Whether first entering SPSS or returning from earlier operations the standard menu of commands across the top is required. As long as it is visible you may perform any analyses. It is not necessary for the data window to be visible.

*After completion of Step 3 a screen with the desired menu bar appears. The sequence to access two-way analysis of variance begins at any screen with the menu of commands visible.*

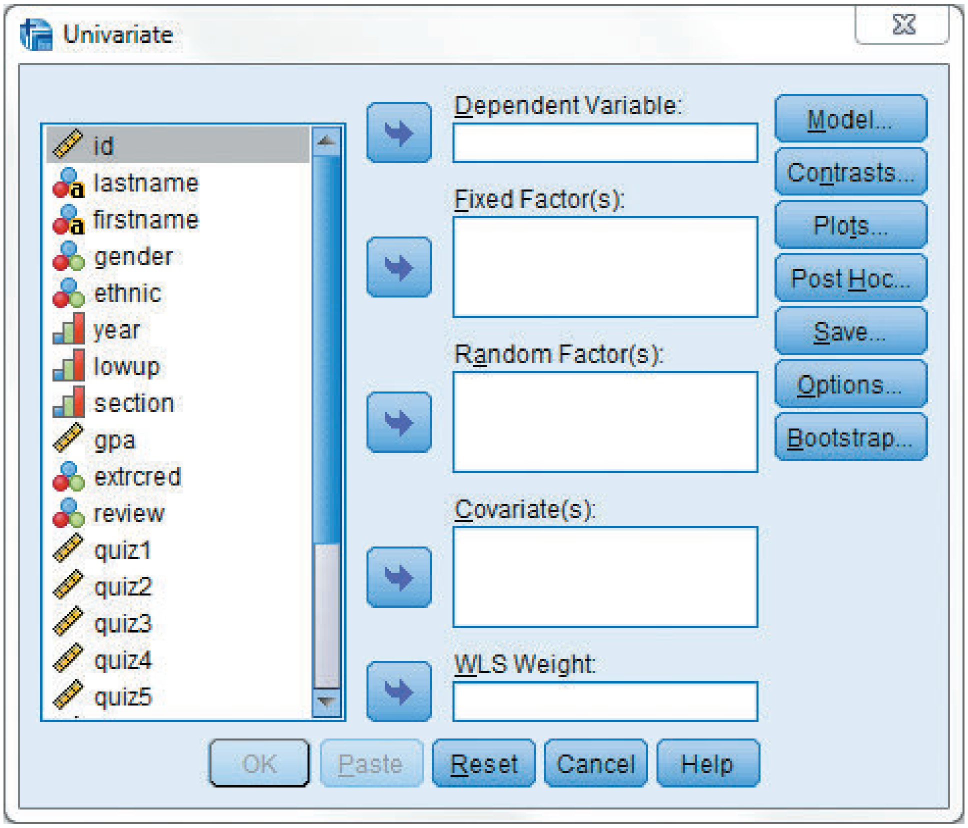
| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | **Analyze** → **General Linear Model** → **Univariate** | 13.1 |

The first window that opens upon clicking **Univariate** (Screen 13.1, below) provides the structure for conducting a two-way analysis of variance. To the left is the usual list of variables. To the right are five empty boxes. We will use the **Dependent Variable** and **Fixed Factor(s)** boxes in this chapter, and discuss the **Covariates** box in Chapter 14. The upper (**Dependent Variable**) box designates the single continuous dependent variable (**total** in this case), and the next (**Fixed Factor(s)**) box indicates independent variables or factors (**gender** and **section** in our example). **Random Factor(s)** and **WLS Weight** options are rarely used, and will not be described. Concerning the buttons to the right, only the final button **Options** will be discussed in this chapter. However, the **Post Hoc** procedure is nearly identical to the option presented in the discussion of one-way ANOVA (see page 160).

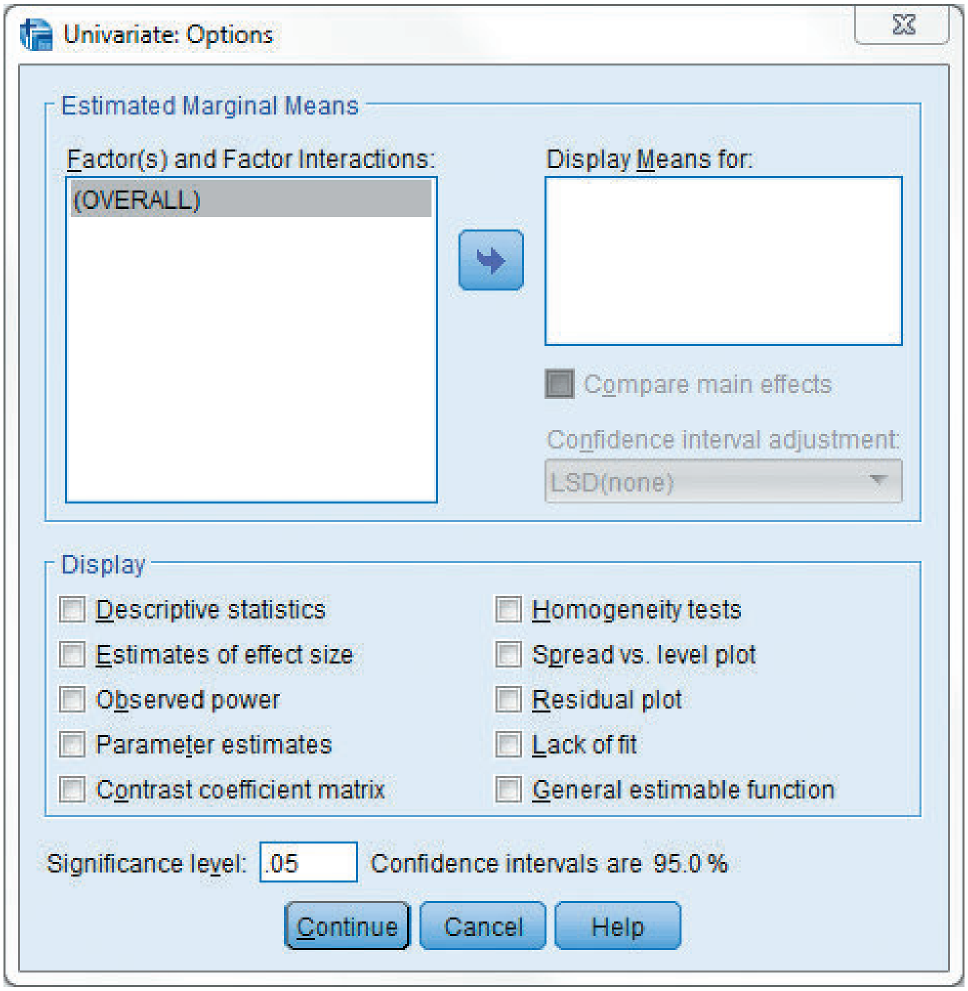
After both the dependent variable and the independent variables are selected, the next step is to click the **Options** button. A new window opens (Screen 13.2, following page) that allows you to request that SPSS **Display** important statistics (such as descriptive statistics and effect sizes) to complete the analysis.

In addition to cell means, **Descriptive statistics** produces standard deviations and sample sizes for each cell. Other frequently used options include, **Estimates of effect size**, which calculates the partial eta squared ($\eta^2$) effect size for each main effect and the interaction (indicating how large each effect is), and **Observed power**, which calculates the power given your effect size and sample size.

Screen 13.1 The Univariate General Linear Model Window



Screen 13.2 The Univariate: Options Window



Univariate: Options

Estimated Marginal Means

Factor(s) and Factor Interactions:
(OVERALL)

Display Means for:

☐ Compare main effects

Confidence interval adjustment:
LSD(none)

Display

☐ Descriptive statistics
☐ Estimates of effect size
☐ Observed power
☐ Parameter estimates
☐ Contrast coefficient matrix
☐ Homogeneity tests
☐ Spread vs. level plot
☐ Residual plot
☐ Lack of fit
☐ General estimable function

Significance level: .05 Confidence intervals are 95.0 %

*The following sequence begins with Screen 13.1. Do whichever of Steps 1–4 (pages* 171 *and* 172*) are necessary to arrive at this screen. Click* **Reset** *if necessary to clear prior variables.*

| In Screen | Do This | Step 5 |
|---|---|---|
| 13.1 | ✱ **total** → ✱ *upper* ➡ → ✱ **gender** → ✱ *second* ➡ → ✱ **section** → | |
| | ✱ *second* ➡ | |
| 13.1 | ✱ **Options** | |
| 13.2 | ✱ **Descriptive statistics** → ✱ **Estimates of effect size** → ✱ **Observed power** | |
| | → ✱ **Continue** | |
| 13.1 | ✱ **OK** | Back1 |

Upon completion of Step 5, the output screen will appear (Screen 1, inside back cover). All results from the just-completed analysis are included in the Output Navigator. Make use of the scroll bar arrows (▲▼▶◀) to view the results. Even when viewing output, the standard menu of commands is still listed across the top of the window. Further analyses may be conducted without returning to the data screen.

## 13.4 Printing Results

Results of the analysis (or analyses) that have just been conducted require a window that displays the standard commands (**File Edit Data Transform Analyze . . .**) across the top. A typical print procedure is shown below beginning with the standard output screen (Screen 1, inside back cover).

*To print results, from the Output screen perform the following sequence of steps*:

| In Screen | Do This | Step 6 |
|---|---|---|
| Back1 | *Select desired output or edit (see pages 16–22)* → ✳ **File** → ✳ **(Print)** | |
| 2.9 | *Consider and select desired print options offered, then* → ✳ **OK** | |

*To exit you may begin from any screen that shows the* **File** *command at the top.*

| In Screen | Do This | Step 7 |
|---|---|---|
| Menu | ✳ **File** → ✳ **Exit** | 2.1 |

Note: After clicking **Exit**, there will frequently be small windows that appear asking if you wish to save or change anything. Simply click each appropriate response.

## 13.5 Output

### 13.5.1 Two-Way Analysis of Variance

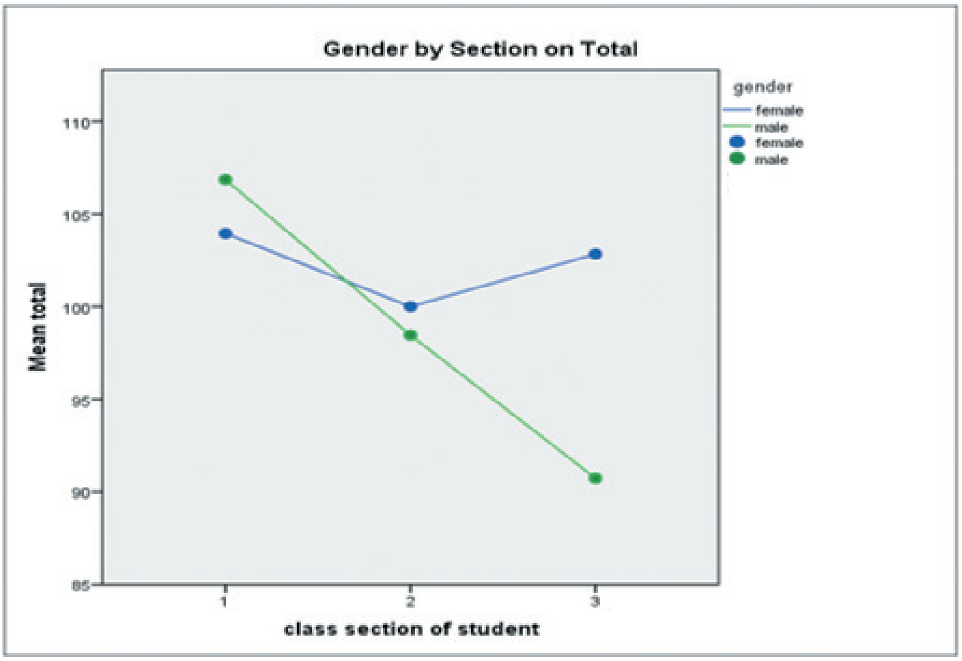
This is slightly condensed output from sequence Step 5 (previous page). We begin with the cell means produced with the **Descriptive statistics** option. A graphical display of the cell means and the ANOVA table (labeled "Tests of Between-Subject Effects") follow this. Commentary and definitions complete the chapter.

**Descriptive Statistics**

Dependent Variable: total

| Gender | Section | Mean | Std. Deviation | N |
|---|---|---|---|---|
| Female | 1 | 103.95 | 18.135 | 20 |
| | 2 | 100.00 | 12.306 | 26 |
| | 3 | 102.83 | 10.678 | 18 |
| | Total | 102.03 | 13.896 | 64 |
| Male | 1 | 106.85 | 13.005 | 13 |
| | 2 | 98.46 | 11.822 | 13 |
| | 3 | 90.73 | 21.235 | 15 |
| | Total | 98.29 | 17.196 | 41 |
| Total | 1 | 105.09 | 16.148 | 33 |
| | 2 | 99.49 | 12.013 | 39 |
| | 3 | 97.33 | 17.184 | 33 |
| | Total | 100.57 | 15.299 | 105 |

**Graph of Gender by Section on Total Interaction**



**ANOVA (Tests of Between-Subjects Effects)**

Dependent Variable: total

| Source | Type III Sum of Squres | df | Mean Square | F | Sig. | Partial Eta Squared | Noncent Parameter | Observed Power[a] |
|---|---|---|---|---|---|---|---|---|
| Corrected Model | 2350.408[b] | 5 | 470.082 | 2.116 | .070 | .097 | 10.580 | .678 |
| Intercept | 996892.692 | 1 | 99892.692 | 4487.382 | .000 | .978 | 4487.382 | 1.000 |
| gender | 316.564 | 1 | 316.564 | 1.425 | .235 | .014 | 1.425 | .219 |
| section | 1262.728 | 2 | 631.364 | 2.842 | .063 | .054 | 5.684 | .546 |
| gender * section | 960.444 | 2 | 480.222 | 2.162 | .121 | .042 | 4.323 | .433 |
| Error | 21993.306 | 99 | 222.155 | | | | | |
| Total | 1086378.000 | 105 | | | | | | |
| Corrected Total | 24343.714 | 104 | | | | | | |

a. Computed using alpha = .05
b. R Squared = .097 (Adjusted R Squared = .051)

Main effect of section

Main effect for gender

Two-way gender by section interaction

The first display (Descriptive Statistics) provides the mean total points for males, for females, for each of the three sections, and for the six cells of the gender by section crosstabulation. The graph above displays the six gender by section means. You can easily plot this graph by hand, but if you want SPSS to produce this graph for you, you may follow the instructions at the end of Chapter 14 (see page 188). (Note that the **Plots** option in the **General Linear Model** → **Univariate** procedure plots *estimated* means based on the model, rather than actual means of the data; this can be a problem in some models.) The tests of Between-Subjects Effects table (in this case, an ANOVA table) provides the answers to the three experimental questions:

1. There is no significant main effect for **gender**: Females ($M = 102.0$) did not score significantly higher than males ($M = 98.3$), $F(1, 99) = 1.425$, $p = .235$.
2. There is a marginally significant main effect for **section**: Results show that those in section 1 ($M = 105.09$) scored higher than students in section 2 ($M = 99.49$) or 3 ($M = 97.33$), $F\ (2, 99) = 2.842$, $p = .063$.

3. There is no significant **gender** by **section** interaction: Despite the lack of statistical significance, we might note that while scores between males and females did not differ much in the first two sections, in the third section females scored much higher ($M = 102.83$) than did males ($M = 90.73$), $F(2, 99) = 2.162$, $p = .12$. These data are displayed on the graph shown on the previous page.

Note that the $F$ values for the main effects are slightly different than in older versions of SPSS. This is because there are several ways of calculating sums of squares, and the default method of calculating these numbers has been changed to better account for designs that don't have the same number of participants in each cell.

A more complete discussion of ANOVA interactions will take place in the following chapter. Definitions of related terms follow:

| Term | Definition/Description |
|---|---|
| **SUM OF SQUARES** | This is the sum of squared deviations from the mean. In a broader sense, the corrected model sum of squares represents the amount of variation in the dependent variable (**total** in this case) explained by each independent variable or interaction, and corrected total sum of squares represents the total amount of variance present in the data. Error sum of squares (sometimes called residual sum of squares) represents the portion of the variance *not* accounted for by the independent variables or their interaction(s). In this example, only about 10% (2,350.408/24,343.714) of the variation in **total** is explained by **gender**, **section**, and the **gender** × **section** interaction. The remaining 90% remains unexplained by these two variables. Total and intercept values relate to the internal model that SPSS uses to calculate two-way ANOVA; it is not necessary to understand them to understand two-way ANOVA. |
| **DEGREES OF FREEDOM (DF)** | **SECTION**: (Levels of **section** − 1) = 3 − 1 = 2.<br>**GENDER**: (Levels of **gender** − 1) = 2 − 1 = 1.<br>**2-WAY INTERACTION:** (Levels of **gender** – 1)(Levels of **section** − 1) = (2 − 1)(3 − 1) = 2.<br>**CORRECTED MODEL**: (DF of Main effects + DF of Interaction) = 3 + 2 = 5.<br>**ERROR**: (*N* – Explained DF − 1) = 105 − 5 − 1 = 99.<br>**CORRECTED TOTAL**: (*N* − 1) = 105 − 1 = 104. |
| **MEAN SQUARE** | Sums of squares divided by degrees of freedom for each category. |
| **F** | The mean square for **gender** divided by the residual mean square; the mean square for **section** divided by the residual mean square; and the mean square for the **gender** × **section** interaction divided by the residual mean square. |
| **SIG** | Likelihood that each result could happen by chance. |
| **PARTIAL ETA SQUARED** | The effect-size measure. This indicates how much of the total variance is explained by each independent variable when removing the effect of the other independent variables and interactions. The corrected model row summarizes the combined effect of all of the main effects and interactions. Note that if you want to report eta squared instead of partial eta squared, you will have to calculate those effect sizes by hand (but it's easy; see page 171 for the equation). |
| **OBSERVED POWER** | The probability of finding a significant effect (at the .05 level) with the sample size of the data being analyzed, assuming that the effect size in the population is the same as the effect size in the current sample. |

# Exercises

The exercises for two-way ANOVAs are combined with the exercises for three-way ANOVAs at the end of Chapter 14.

# Chapter 14

# General Linear Model: Three-Way ANOVA

THE PREVIOUS two chapters—Chapter 12 (one-way ANOVA) and Chapter 13 (two-way ANOVA)—have overviewed the goals of analysis of variance. A three-way analysis of variance asks similar questions, and we are not going to reiterate here what we have already presented. We will address in this chapter those things that are unique to three-way analysis of variance (as contrasted with one-way or two-way ANOVAs), and also describe what a covariate is and how it can be used with the **General Linear Model** command.

This chapter will be organized in the same manner as all the analysis chapters (Chapters 6–27), but the emphasis will be somewhat different. This introductory portion will deal with the experimental questions that are usually asked in a three-way ANOVA and then explain what a covariate is and how it is typically used. The Step by Step section will include two different types of analyses. A three-way ANOVA will be described based on the **grades.sav** file, in which total points earned (**total**) will serve as the dependent variable, and **gender**, **section**, and **lowup** (lower- or upper-division student) will be designated as independent variables. The second analysis will be identical except that it will include one covariate, **gpa**, to "partial out" variance based on each student's previous GPA. The output from an analysis of variance is divided into two major sections: (1) The *descriptive statistics* portion, which lists the mean and frequency for each category created by the ANOVA, and (2) the ANOVA table, which indicates sums of squares, *F*-values, significance values, and other statistical information. The first portion (the descriptive statistics section) of the analysis will be identical whether or not a covariate is used, but the second portion (the ANOVA table) will often be quite different.

The major difference between this chapter and previous chapters is that most of the explanation will occur in the Output section. A three-way or higher-order ANOVA is complex by any standards, and a very thorough and careful explanation of the analysis done here will hopefully assist you in untangling any three-way or higher-order ANOVAs that you might conduct. An almost universally practiced procedure for helping to clarify ANOVA results is to graph the cell means for interactions. We will create and examine such graphs in this chapter (instructions for producing these graphs in SPSS are included at the end of the chapter). In fact, even *non* significant interactions will be graphed and related to the appropriate lines of output so that you can understand visually why there *is* a significant effect (when one occurs) and why there *isn't* (when a significant effect does *not* occur). A graph is not normally scrutinized for nonsignificant effects, but we hope that inclusion of them here will assist you toward a clearer understanding of the entire output.

## 14.1 Three-Way Analysis of Variance

Restating briefly, the dependent variable for this analysis is **total**, with independent variables of **gender**, **section**, and **lowup**. For this analysis (and for any three-way

ANOVA) there will be seven possible experimental questions of interest. The first three questions deal with main effects, the next three questions deal with two-way interactions, and the final experimental question concerns whether there is a significant three-way interaction. The seven questions examined in this example are:

1. Is there a main effect for **gender** (i.e., do males score significantly differently from females on **total** points)?
2. Is there a main effect for **section** (i.e., do mean **total** points earned in one section differ significantly from the other sections)?
3. Is there a main effect for **lowup** (i.e., do upper-division students score significantly differently than lower-division students)?
4. Is there a significant interaction between **gender** and **section** on **total** points?
5. Is there a significant interaction between **gender** and **lowup** on **total** points?
6. Is there a significant interaction between **section** and **lowup** on **total** points?
7. Is there a significant interaction between **gender, section**, and **lowup** on **total**?

SPSS will print cell means for all combinations of variables and compute *F* values and associated significance values. These values will indicate if there are significant main effects and/or if there are significant interactions between variables. As stated in the previous chapter, it is quite possible to have significant main effects but *not* to have significant interactions, just as it is possible to have significant interactions without significant main effects. How to interpret interactions is discussed in some detail in the Output section.

## 14.2 The Influence of Covariates

The purpose of covariates is to partition out the influence of one or more variables before conducting the analysis of interest. A *covariate* could best be defined as a variable that has a substantial correlation with the dependent variable and is included in the experiment to adjust the results for differences existing among subjects before the start of the experiment. For example, a 1987 study explored the influence of personality traits on the competitive success (a standardized score of fastest racing times) of a sample of runners who ranged in ability from slow fitness runners to world-class athletes. In addition to a measure of 16 personality traits, each participant's weekly running mileage, weight/height ratio, and age were acquired. The influence of these physiological measures was well known prior to analysis (runners who run more miles, are skinnier, and who are younger, run faster). In the analysis, all three measures were used as covariates. The purpose was to exclude variance in the dependent variable (fast racing times) that was determined by the weekly mileage, weight/height ratio, and age. This would allow the researcher to see more clearly the influence of the psychological measures without the additional physiological factors included.

In our sample study **gpa** will be used as a covariate. In many different educational settings it has been demonstrated that students who have a higher overall GPA tend to do better in any given class. Thus if we wish to see the influence of **gender**, **section**, and **lowup** (lower- or upper-division student) *independently* of the influence of the student's previous GPA, then **gpa** can be included as a covariate. The inclusion of a covariate or covariates does not influence the cell means in the initial output, but it often has substantial influence (usually reduction) on the sum of squares, but can sometimes increase the *F* values (along with a corresponding decrease of *p*-values) because the error is reduced. Note that an ANOVA with a covariate is often abbreviated as ANCOVA. These issues will be further clarified in the Output section.

# 14.3 Step by Step

## 14.3.1 Three-Way Analysis of Variance

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | ✳ → ✳ All Programs ▷ → ✳ IBM SPSS Statistics → ✳ IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | ✳ → ✳ → ✳ | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| Step 2 |
|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* |

*Screens 1 and 2 (displayed on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **grades.sav** *file for further analyses:*

| In Screen | Do This | | Step 3 |
|---|---|---|---|
| Front1 | ✳ File → Open → ✳ Data | [*or* ✳ 📂] | |
| Front2 | type `grades.sav` → ✳ Open | [*or* ✳ ✳ grades.sav] | Data |

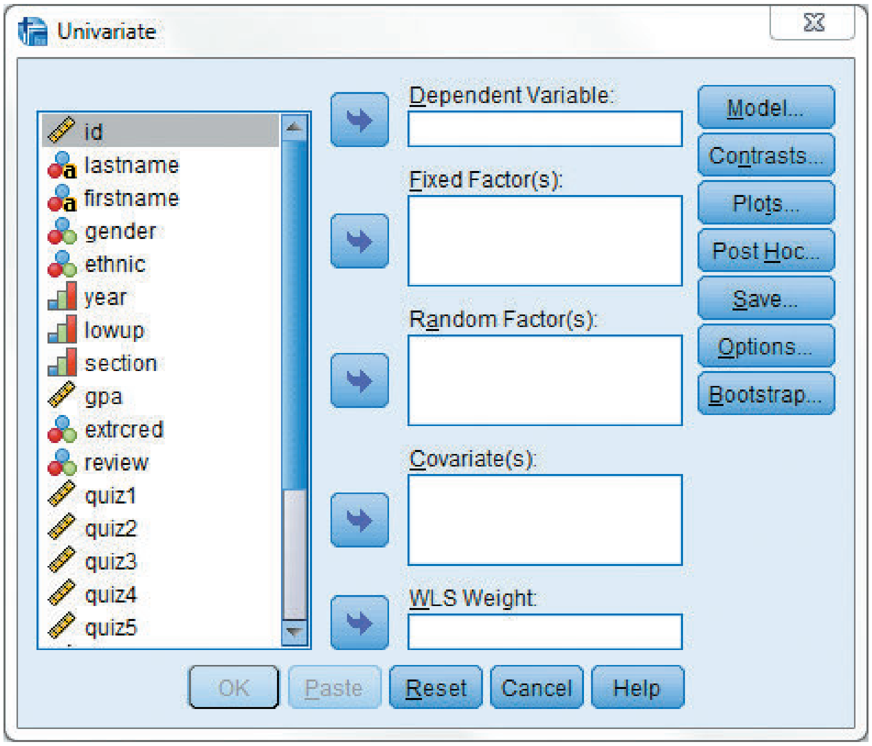
Whether first entering SPSS or returning from earlier operations the standard menu of commands across the top is required. As long as it is visible you may perform any analyses. It is not necessary for the data window to be visible.

*After completion of Step 3 a screen with the desired menu bar appears. The sequence to access three-way analysis of variance begins at any screen with the menu of commands visible.*

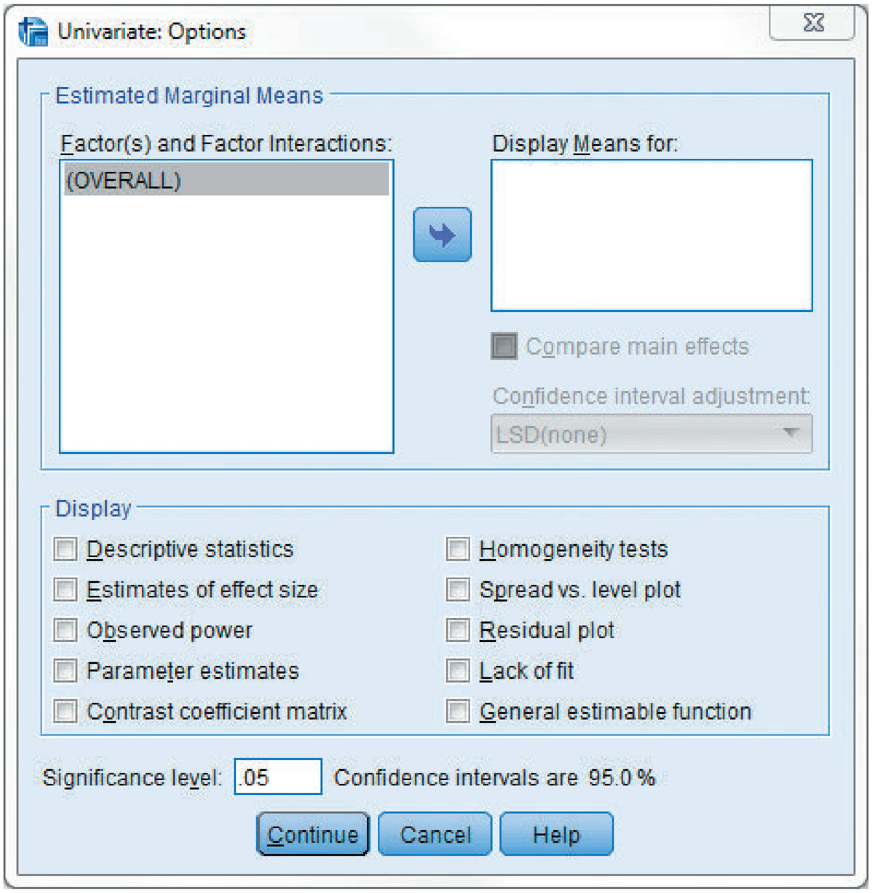
| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | ✳ Analyze → General Linear Model → ✳ Univariate | 14.1 |

First, a word concerning the four choices that appear upon clicking the **General Linear Model** button. In addition to the **Univariate** choice that has been selected in the previous chapter (and in this one), there are three other options: **Multivariate**, **Repeated Measures**, and **Variance Components**. The **Multivariate** option includes operations more frequently identified by the acronyms MANOVA (multivariate analysis of variance) and MANCOVA (multivariate analysis of covariance) and is covered in Chapter 23. **Repeated Measures** is introduced in Chapter 24 and may be applied in either an ANOVA or a MANOVA context. The **Variance Components** option allows for greater flexibility, but its complexity takes it beyond the scope of this book.

Screen 14.1 The Univariate ANOVA Window



Screen 14.2 The Univariate: Options Window



Significance level: .05 Confidence intervals are 95.0 %

The first screen that appears upon clicking **Univariate** (Screen 14.1, to the left) provides the structure for conducting a three-way analysis of variance. To the left is the usual list of variables. To the right of the variables are five boxes, three of which will interest us. The upper (**Dependent Variable**) box is used to designate the single continuous dependent variable, the variable **total** in this example. The next (**Fixed Factor(s)**) box is used to specify the categorical independent variables or factors (exactly three in this study), and a lower (**Covariate(s)**) box designates covariates, described in the introduction of this chapter. After selecting the dependent variable, **total**, and pasting it into the top box, each independent variable is pasted into the **Fixed Factor(s)** box.

After the dependent variable (**total**) has been designated, all three independent variables (**gender**, **section**, **lowup**) have been selected and the covariate (**gpa**) chosen, the next step is to click the **Options** button. A new window opens (Screen 14.2, at left) that allows you to **Display** important statistics to complete the analysis. **Descriptive statistics** produces means, standard deviations, and sample sizes for each cell. One other frequently used option, **Estimates of effect size**, calculates partial eta squared ($\eta^2$) effect size for each main effect and the interactions, indicating how large each effect is.

One technique that is often used to aid in the interpretation of main effects, the use of *post hoc* tests, will be discussed in Chapter 23, page 301. Although that chapter deals with a somewhat different procedure (**Multivariate** analysis of variance), the **Post Hoc** dialog box is identical to one used for univariate post hoc tests described in Chapter 12.

*The following sequence begins with Screen 14.1. Do whichever of Steps 1–4 (page 179) are necessary to arrive at this screen. Click the* **Reset** *button if necessary.*

| In Screen | Do This | Step 5 |
|---|---|---|
| 14.1 | ✳ **total** → ✳ *upper* ➡ → ✳ **gender** → ✳ *second* ➡ → | |
| | ✳ **section** → ✳ *second* ➡ → ✳ **lowup** → ✳ *second* ➡ | |
| 14.1 | ✳ **gpa** → ✳ *fourth "Covariate(s)"* ➡ → ✳ **Options** | |
| 14.2 | ✳ **Descriptive statistics** → ✳ **Estimates of effect size** → ✳ **Continue** | |
| 14.1 | ✳ **OK** | Back1 |

*Note: If you wish to conduct the analysis without the covariate, simply omit the line selecting **gpa**.

Upon completion of Step 5, the output screen will appear (inside back cover, Screen 3). All results from the just-completed analysis are included in the Output Navigator. Make use of the scroll bar arrows (▲▼▶◀) to view the results. Even when viewing output, the standard menu of commands is still listed across the top of the window. Further analyses may be conducted without returning to the data screen.

## 14.4 Printing Results

Results of the analysis (or analyses) that have just been conducted require a window that displays the standard commands (**File Edit Data Transform Analyze. . .**) across the top. A typical print procedure is shown below beginning with the standard output screen (Screen 1, inside back cover).

*To print results, from the Output screen perform the following sequence of steps:*

| In Screen | Do This | Step 6 |
|---|---|---|
| Back1 | *Select desired output or edit (see pages 16–22)* → ✳ **File** → ✳ **Print** | |
| 2.9 | *Consider and select desired print options offered, then* → ✳ **OK** | |

*To exit you may begin from any screen that shows the* **File** *command at the top.*

| In Screen | Do This | Step 7 |
|---|---|---|
| Menu | ✳ **File** → ✳ **Exit** | 2.1 |

Note: After clicking **Exit**, there will frequently be small windows that appear asking if you wish to save or change anything. Simply click each appropriate response.

## 14.5 Output

### 14.5.1 Three-Way Analysis of Variance and Analysis of Covariance

The format in this Output section will differ from most other output sections in this book. The output will be divided into two separate presentations. In the first of these, we show the results of a three-way analysis of variance that does *not* include a covariate. For this first analysis, instead of reproducing the output generated by SPSS we will integrate the descriptive statistics portion of the output (the first segment) with the ANOVA table portion (which follows). After each of the nine sections of the cell-means portion, the ANOVA line that relates to that portion will follow immediately. (For the real SPSS output, you'll need to hunt around to find the appropriate statistics.) For the five tables that involve interactions, a graph of the output will be included to help clarify the relation among the table, the graph, and the ANOVA output. These graphs may be produced by hand, or SPSS can produce these graphs for you; instructions for producing these graphs are included at the end of the chapter (pages 188–189). Explanation or clarification will follow after each table/graph/line presentation, rather than at the end of the Output section.

For the second presentation, the output from a three-way analysis of variance *that includes a covariate* will be shown. The tables of descriptive statistics and associated graphs (the first presentation) are identical whether or not a covariate is included; so these will not be produced a second time. What we will show is the complete ANOVA table output in similar format as that produced by SPSS. To demonstrate how mean square, *F*-, and *p*-values differ when a covariate is included, the ANOVA results *without* the covariate will *also* be included in the same table. The output that does *not* include the covariate will be shown *italicized* so as not to interfere in the interpretation of the original output. Explanation will then follow.

*The following output is produced from sequence Step 5 on page* 181 *(this initial portion are results without the covariate). Instructions for producing the graphs are on pages* 188–189. *For each result the cell means portion of the output is presented first followed by the associated ANOVA line.*

## 14.5.2 Total Population

| gender | section | lowup | Mean | Std. Deviation | N |
|---|---|---|---|---|---|
| Total | Total | Total | 100.57 | 15.299 | 105 |

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|
| Corrected Total | 24343.714 | 104 | | | | |

This cell identifies the overall mean for the variable **total** for the entire sample ($N = 105$). The sum of squares is the total of squared deviations from the grand mean for all subjects. The degrees of freedom is $N - 1$ ($105 - 1 = 104$). There are no mean square, *F*, *p*, or *partial eta squared* statistics generated for a single value.

## 14.5.3 Main Effect for GENDER

| gender | section | lowup | Mean | Std. Deviation | N |
|---|---|---|---|---|---|
| Female | Total | Total | 102.03 | 13.896 | 64 |
| Male | | | 98.29 | 17.196 | 41 |

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|
| gender | 454.062 | 1 | 454.062 | 2.166 | .144 | .023 |

The table indicates that 64 females scored an average of 102.03 **total** points while 41 males scored an average of 98.29. A visual display (e.g., graph) is rarely needed for main effects. It is clear that females scored higher than males. The number of levels of the variable minus one determines the degrees of freedom for main effects. Thus degrees of freedom for **gender** is 2 − 1 = 1; for **section** is 3 − 1 = 2; and for **lowup** is 2 − 1 = 1. The *F*-value is determined by dividing the mean square for the variable of interest (**gender** in this case) by the residual mean square. The *F* = 2.166 and *p* = .144 verify that there is no significant main effect for gender; that is, the scores for females and males do not differ significantly. The partial eta squared value indicates the proportion of the variance that is accounted for by this variable when removing the effects of the other variables and interactions; in this case, 2.3% of the variance in **total** is uniquely accounted for by **gender**.

## 14.5.4 Main Effect for SECTION

| Gender | Section | Lowup | Mean | Std. Deviation | N |
|---|---|---|---|---|---|
| Total | 1 | Total | 105.09 | 16.148 | 33 |
| | 2 | | 99.49 | 12.013 | 39 |
| | 3 | | 97.33 | 17.184 | 33 |

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|
| section | 1181.397 | 2 | 590.698 | 2.818 | .065 | .057 |

This table documents that 33 students from the first section scored a mean of 105.09 points, 39 students from the second section scored a mean of 99.49 points, and 33 students from the third section scored a mean of 97.33 points. An *F*-value of 2.82 (*p* = .065) indicates that there is only a marginally significant main effect for **section**—not enough to rely upon, but perhaps worthy of future study. Visual inspection verifies that the difference in scores between the first and the third sections is greatest.

## 14.5.5 Main Effect for LOWUP

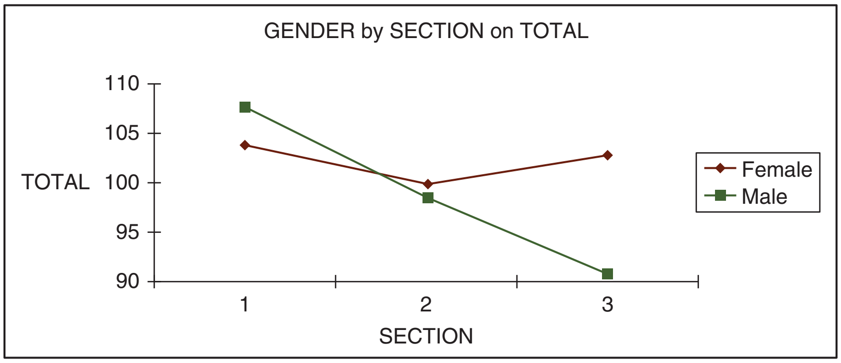
| Gender | Section | Lowup | Mean | Std. Deviation | N |
|---|---|---|---|---|---|
| Total | Total | lower division | 99.55 | 14.66 | 22 |
| | | upper division | 100.84 | 15.54 | 83 |

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|
| LOWUP | 14.507 | 1 | 14.507 | .069 | .793 | .001 |

This result indicates that 22 lower-division students scored a mean **total** of 99.55, while 83 upper-division students scored a mean **total** of 100.84. A graph, once again, is unnecessary. An *F*-value of .069 and *p* of .793 (along with a partial eta squared of .001, indicating that only 0.1% of the variance of **total** is accounted for by **lowup** when removing the effects of gender and section and of the interactions) indicate no significant differences for scores of upper-division and lower-division students.

## 14.5.6 Two-Way Interaction, GENDER by SECTION

| gender | section | lowup | Mean | Std. Deviation | N |
|---|---|---|---|---|---|
| Female | 1 | Total | 103.95 | 18.135 | 20 |
| | 2 | | 100.00 | 12.306 | 26 |
| | 3 | | 102.83 | 10.678 | 18 |
| Male | 1 | | 106.85 | 13.005 | 13 |
| | 2 | | 98.46 | 11.822 | 13 |
| | 3 | | 90.73 | 21.235 | 15 |



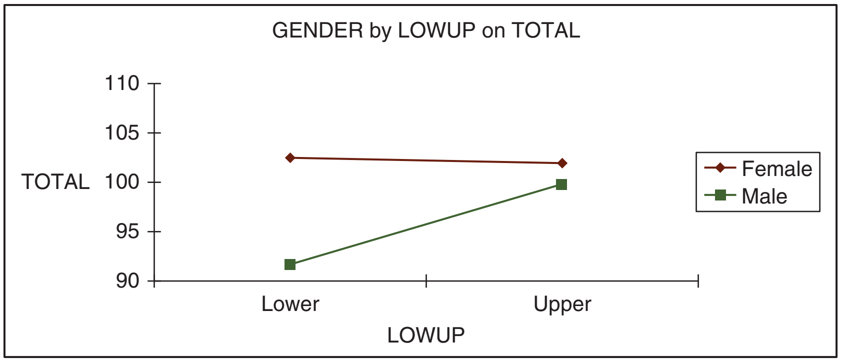
| Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|
| gender * section | 118.155 | 2 | 59.078 | .282 | .755 | .006 |

These results indicate no significant **gender** by **section** interaction ($F = .282$, $p = .755$). The top table identifies the mean **total** score and the number of subjects within each of six categories. One way to visually identify an interaction is to see whether two or more lines of the graph are parallel. If the lines are parallel, this indicates that there is *no* interaction. There is no *significant* interaction when the lines do not differ significantly from parallel. There is a significant interaction when the two lines *do* differ significantly from parallel. Be careful! Interactions cannot be determined by viewing a graph alone. The vertical and horizontal scales of a graph can be manipulated to indicate a greater effect than actually exists. The related output from the ANOVA table will indicate whether or not there is a significant effect. In the graph above, clearly the "action" is happening in the third section—but there is not enough "action" to be sure it is not just random (thus, it is not significant). These data could also be presented with the **gender** variable along the horizontal axis and the **section** variable coded to the right. The decision of which variable to place on which axis usually comes down to what relationship between the variables you want to highlight; if it's not clear, you can graph both and see which tells a better story.

## 14.5.7 Two-Way Interaction, GENDER by LOWUP

| Gender | Section | Lowup | Mean | Std. Deviation | N |
|---|---|---|---|---|---|
| Female | Total | lower division | 102.50 | 14.482 | 16 |
| | | upper division | 101.87 | 13.850 | 48 |
| Male | | lower division | 91.67 | 13.095 | 6 |
| | | upper division | 99.43 | 17.709 | 35 |

| Source | Type III Sum of Squares | Df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|
| gender * lowup | 269.812 | 1 | 269.812 | 1.287 | .260 | .014 |

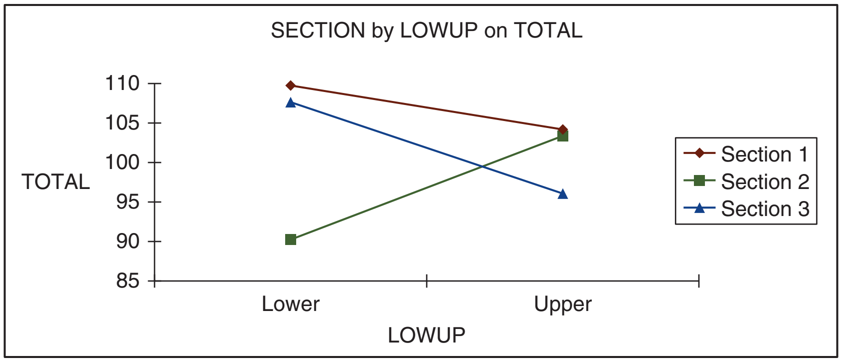


An *F*-value of 1.29 (*p* = .26) indicates no significant interaction between **gender** and **lowup**. Note that the two lines *appear* to deviate substantially from parallel, and yet there's no significance. Two factors influence this: (1) Values on the vertical axis vary only from 90 to 110, but **total** points vary from 48 to 125. We display only a narrow segment of the actual range, causing an exaggerated illusion of deviation from parallel. (2) The significance of a main effect or interaction is substantially influenced by the sample size. In this case, two of the cells have only 6 and 16 subjects, decreasing the *power* of the analysis. Lower power means that a greater difference is required to show significance than if the sample size was larger.

## 14.5.8 Two-Way Interaction, SECTION by LOWUP

| Gender | Section | Lowup | Mean | Std. Deviation | N |
|---|---|---|---|---|---|
| Total | 1 | lower division | 109.86 | 9.512 | 7 |
| | | upper division | 103.81 | 17.436 | 26 |
| | 2 | lower division | 90.09 | 13.126 | 11 |
| | | upper division | 103.18 | 9.444 | 28 |
| | 3 | lower division | 107.50 | 9.469 | 4 |
| | | upper division | 95.93 | 17.637 | 29 |

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|
| section * lowup | 1571.905 | 2 | 785.952 | 3.749 | .027 | .075 |

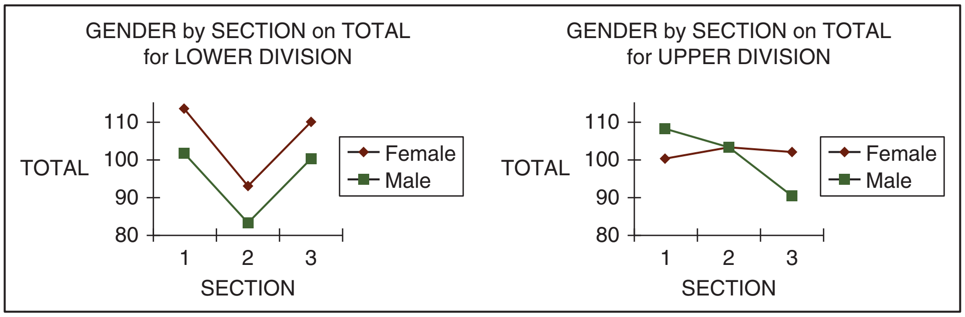


Both the ANOVA line (note the *F* = 3.75, *p* = .027) and the graph indicate a significant **section** by **lowup** interaction. Although the first and third sections' lines are relatively close to parallel, the second section shows a nearly opposite trend. A reasonable interpretation of the interaction might be: Although lower-division students tend to score higher than upper-division students in sections 1 and 3, the reverse is true in section 2. As indicated by the partial eta squared value, the **section** by **lowup** interaction accounts for 7.5% of the variance in **total** when removing the effects of the other main and interaction effects.

## 14.5.9 Three-Way Interaction, GENDER by SECTION by LOWUP

Three-way and higher-order interactions are often quite difficult to interpret. The researcher must have a strong conceptual grasp of the nature of the data set and what constitutes meaningful relationships. For the sample listed here, an *F* = .409 and *p* = .665 (table at the top of the following page) indicate that the three-way interaction does not even hint at significance. When there is no significant interaction, the researcher would usually not draw graphs. We have included them here to demonstrate one way that you might visually display a three-way interaction. Although one might note the almost identical pattern for females and males in the first chart, a more important thing to notice is the very low cell counts for all six cells of the lower division group (5, 8, 3, 2, 3, and 1, respectively). That is why it is important to examine the *F* values in addition to the means and graphs. The degrees of freedom for a three-way interaction is the number of levels of the first variable minus one, times levels of the second variable minus one, times levels of the third variable minus one [$(2-1) \times (3-1) \times (2-1) = 2$].

| Gender | Section | Lowup | Mean | Std. Deviation | N |
|---|---|---|---|---|---|
| Female | 1 | lower division | 113.20 | 6.458 | 5 |
| | | upper division | 100.867 | 19.842 | 15 |
| | 2 | lower division | 93.000 | 13.743 | 8 |
| | | upper division | 103.111 | 10.566 | 18 |
| | 3 | lower division | 110.000 | 9.849 | 3 |
| | | upper division | 101.400 | 10.555 | 15 |
| Male | 1 | lower division | 101.500 | 13.435 | 2 |
| | | upper division | 107.818 | 13.348 | 11 |
| | 2 | lower division | 82.333 | 8.737 | 3 |
| | | upper division | 103.300 | 7.528 | 10 |
| | 3 | lower division | 100.000 | . | 1 |
| | | upper division | 90.071 | 21.875 | 14 |



| Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|
| gender * section * lowup | 171.572 | 2 | 85.786 | .409 | .66 | .009 |

## 14.6 A Three-Way Anova that Includes a Covariate

*The following output is produced by sequence Step 5 (page* 181*) and includes results both with and without a covariate. Values WITH a covariate are shown in* **boldface***, values WITHOUT a covariate are shown (italicized and in parentheses). Note that SPSS produces more significant digits than are presented in this table; we show values rounded to one digit beyond the decimal point due to space considerations. Commentary follows.*

| Source | Type III Sum of Squares | df | Mean Square | F | Sig of F | Partial Eta$^2$ |
|---|---|---|---|---|---|---|
| **Corrected Model** | **8006.6** *(4846.0)* | **12** *(11)* | **667.2** *(440.5)* | **3.76** *(2.10)* | **.00** *(.03)* | **.33** *(.20)* |
| **Intercept** | **35588.6** *(494712.2)* | **1** *(1)* | **35588.6** *(494712.2)* | **200.4** *(2359.7)* | **.00** *(.00)* | **.69** *(.96)* |
| Covariate: **Gpa** | **3160.6** *(na)* | **1** *(na)* | **3160.6** *(na)* | **17.80** *(na)* | **.00** *(na)* | **.16** *(na)* |
| Main Effects | | | | | | |
| **Gender** | **78.0** *(454.1)* | **1** *(1)* | **78.0** *(454.1)* | **0.44** *(2.17)* | **.51** *(.14)* | **.01** *(.02)* |
| **Section** | **896.7** *(1181.4)* | **2** *(2)* | **448.3** *(590.7)* | **2.53** *(2.82)* | **.09** *(.07)* | **.05** *(.06)* |
| **Lowup** | **53.6** *(14.5)* | **1** *(1)* | **53.6** *(14.5)* | **0.30** *(0.07)* | **.58** *(.79)* | **.00** *(.00)* |
| 2-way Interactions | | | | | | |
| **gender * section** | **38.1** *(118.2)* | **2** *(2)* | **19.1** *(59.1)* | **0.11** *(0.28)* | **.90** *(.76)* | **.00** *(.01)* |
| **gender * lowup** | **79.3** *(269.8)* | **1** *(1)* | **79.3** *(269.8)* | **0.45** *(1.29)* | **.51** *(.26)* | **.01** *(.01)* |
| **section * lowup** | **1510.0** *(1571.9)* | **2** *(2)* | **755.0** *(786.0)* | **4.25** *(3.75)* | **.02** *(.03)* | **.09** *(.08)* |
| 3-way Interaction | | | | | | |
| **gender * section* lowup** | **162.6** *(171.6)* | **2** *(2)* | **81.3** *(85.5)* | **0.46** *(0.41)* | **.63** *(.67)* | **.01** *(.01)* |
| **Error** | **16337.1** *(19497.7)* | **92** *(93)* | **177.6** *(209.7)* | | | |
| **Total** | **1086378.0** *(same)* | **105** *(105)* | | | | |
| **Corrected Total** | **24343.7** *(same)* | **104** *(104)* | | | | |

In this final section of the output, we show the ANOVA output in the form displayed by SPSS. The cell means and graphs (previous section) will be identical whether or not a covariate or covariates are included. What changes is the sum of squares and, correspondingly, the degrees of freedom (in some cases), the mean squares, *F* values, significance values, and partial eta squared. If the covariate has a substantial influence, the analysis will usually demonstrate lower mean square and *F* values for most of the main effects and interactions, and higher significance values.

Some general observations concerning the output (above) are mentioned here, followed by the definition of terms. Note that the covariate **gpa** accounts for a substantial portion of the total variation in the dependent variable **total**. The partial eta squared value indicates the size of the effect; notice that **gpa** accounts for 16.2% of the variance in **total** when removing the effects of the other variables and interactions. The *corrected model* values (top line of data in the chart above) are the sums of the sum of squares and degrees of freedom from all explained sources of variation—covariates, main effects, two-way interactions, and the three-way interaction.

Note also the contrasts of the components of the ANOVA table when comparing analyses with and without a covariate. Since the covariate "consumes" much of the variance, most of the *F* values are lower and corresponding *p* values (for main effects and interactions) are higher when the covariate is included. This is not strictly true, however. Notice that the main effect for **lowup** and the **section × lowup** interaction show higher *F* values and correspondingly lower *p* values when the covariate is included.

| Term | Definition/Description |
|---|---|
| **SUM OF SQUARES** | This is the sum of squared deviations from the mean. In a broader sense, explained sum of squares represents the amount of variation in the dependent variable **(total** in this case) explained by each independent variable or interaction. Error sum of squares represents the portion of the variance *not* accounted for by the covariate(s), independent variables, or their interaction(s). |
| **DEGREES OF FREEDOM** | Covariates: 1 degree of freedom for each covariate.<br>Main effects: Sum of the main effects degrees of freedom for each variable: 1 + 2 + 1 = 4.<br>**gender:** (Levels of **gender** – 1): 2 – 1 = 1.<br>**section:** (Levels of **section** – 1): 3 – 1 = 2.<br>**lowup:** (Levels of **lowup** – 1): 2 – 1 = 1.<br>Two-way interactions: Sum of degrees of freedom for the three interactions: 2 + 1 + 2 = 5.<br>Three-way interactions: Degrees of freedom for the individual three-way interaction (2).<br>Corrected Model: (df covariates + Df main effects + df of Interactions): 1 + 4 + 5 + 2 = 12.<br>Error: (*N* – Explained df – 1): 105 – 12 – 1 = 92.<br>Total: (*N* – 1): 105 – 1 = 104. |

| Term | Definition/Description |
|---|---|
| **MEAN SQUARE** | Sums of squares divided by degrees of freedom for each category. |
| **F** | The mean square for each main effect or interaction divided by the error mean square. |
| **SIG.** | Likelihood that each result could happen by chance. |
| **PARTIAL ETA SQUARED** | The effect-size measure. This indicates how much of the total variance is explained by each independent variable or interaction (when removing the effect of the other variables and interactions). If you wish to calculate the eta squared (instead of the partial eta squared), divide the sum of squares for a main or interaction effect by the corrected total sum of squares. |

### 14.6.1 Graphing Interactions Using SPSS

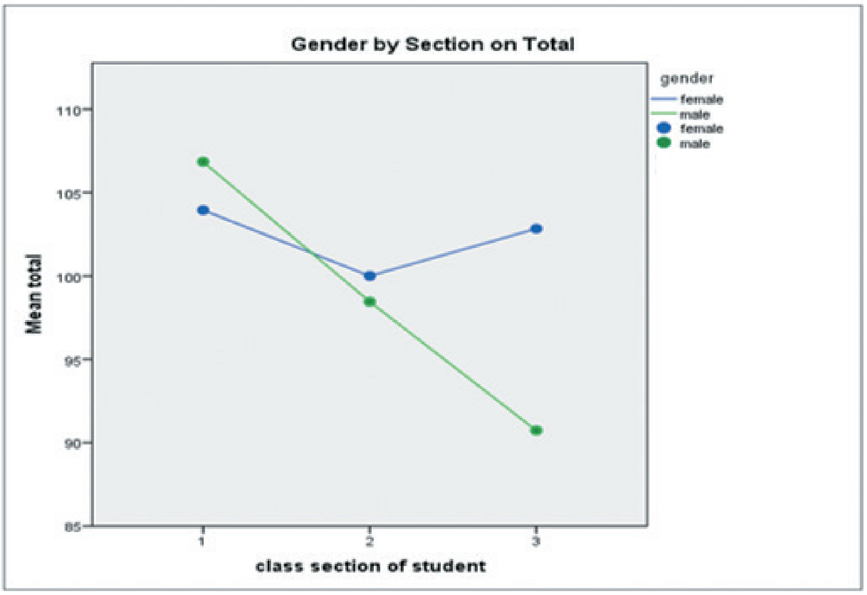
The line graph created here is identical to the first two-way interaction described in this chapter (and Chapter 13). This procedure works for any two-way interaction; if you want to produce three-way interactions, then you can simply produce several graphs for each level of a third independent variable. To select one level of an independent variable, use the **Select Cases** option (described in Chapter 4, pages 73–75).

To produce a graph of a two-way interaction, select the **Graphs** menu, then **Chart Builder**, followed by either the **Bar** option or the **Line** option. (Purists—and APA Style—say that you should use a bar graph for this kind of graph, because the X-axis is categorical instead of a scale. But tradition—and most statistics books—still use line graphs. Extra points for the first to figure out what your instructor likes!) Our examples include both a line chart and a bar chart.

*To achieve a multiple line graph that represents an ANOVA interaction*:

| In Screen | Do This | Step 5a |
|---|---|---|
| Menu | ✳ **Graphs** → ✳ **Chart Builder** → ✳ **Line** | |
| 5.1 | ✳ 📈 → *drag to* **Chart preview** *box (directly above)* | |
| 5.2 | ✳ **total** → *drag to* **Y-Axis?** *box* → ✳ **section** → *drag to* **X-Axis?** *box* → | |
| | ✳ **gender** → *drag to* **Set color** *box* → ✳ **OK** | Back1 |

This procedure produces (after edits) the following graph:

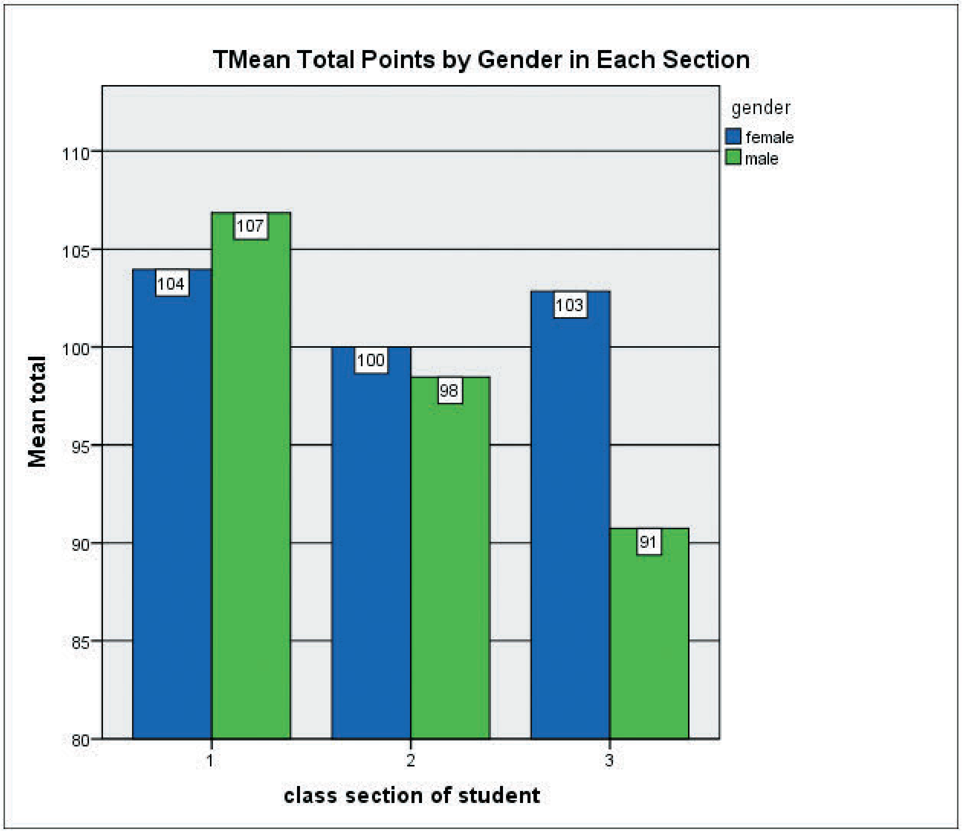


For a number of editing options, see Chapter 5, pages 90–91.

*If you want a bar chart instead of a line graph, do this instead:*

| In Screen | Do This | Step 5 |
|---|---|---|
| Menu | **Graphs** → **Chart Builder** → **Bar** | |
| 5.1 | 📊 → *drag to* **Chart preview** *box (directly above)* | |
| 5.2 | **total** → *drag to* **Y-Axis?** *box* → **section** → *drag to* **X-Axis?** *box* → | |
| | **gender** → *drag to* **Cluster on X: set color** *box* → **OK** | Back1 |

This procedure produces (after some edits) the following graph. Note that the graph means the same thing as the line graph, but may be easier or more difficult to interpret, depending on your goals and what you've practiced before. To access a number of editing options for bar charts, see Chapter 5, page 88–89.



# Exercises

Answers to selected exercises are downloadable at **www.spss-step-by-step.net.**

Notice that data files other than the **grades.sav** file are being used here. Please refer to the **Data Files** section starting on page 364 to acquire all necessary information about these files and the meaning of the variables. As a reminder, all data files are downloadable from the web address shown above.

For the first five problems below, perform the following:

- Print out the cell means portion of the output.
- Print out the ANOVA results (main effects, interactions, and so forth).
- Interpret and write up correctly (APA format) all main effects and interactions.
- Create multiple-line graphs (or clustered bar charts) for all significant interactions.

1. File: **helping3.sav**; dependent variable: **tothelp**; independent variables: **gender**, **problem**.
2. File: **helping3.sav**; dependent variable: **tothelp**; independent variables: **gender**, **income**.
3. File: **helping3.sav**; dependent variable: **hseveret**; independent variables: **ethnic**, **problem**.
4. File: **helping3.sav**; dependent variable: **thelplnz**; independent variables: **gender**, **problem**; covariate: **tqualitz**.
5. File: **helping3.sav**; dependent variable: **thelplnz**; independent variables: **gender**, **income**, **marital**.
6. In an experiment, participants were given a test of mental performance in stressful situations. Some participants were given no stress-reduction training, some were given a short stress-reduction training session, and some were given a long stress-reduction training session. In addition, some participants who were tested had a low level of stress in their lives, and others had a high level of stress in their lives. Perform an ANOVA on these data (listed below). What do the results mean?

| Training: | None | | Short |
|---|---|---|---|
| Life Stress: | High | Low | High |
| Performance Score: | 5 4 2 5 4 | 4 4 6 6 2 | 6 4 5 4 3 |

| Training: | Short | Long | |
|---|---|---|---|
| Life Stress: | Low | High | Low |
| Performance Score: | 7 6 6 5 7 | 5 5 5 3 5 | 7 7 9 9 8 |

7. In an experiment, participants were given a test of mental performance in stressful situations. Some participants were given no stress-reduction training, and some were given a stress-reduction training session. In addition, some participants who were tested had a low level of stress in their lives, and others had a high level of stress in their lives. Finally, some participants were tested after a full night's sleep, and some were tested after an all-night study session on three-way ANOVA. Perform an ANOVA on these data (listed below question 8; ignore the "caffeine" column for now). What do these results mean?
8. In the experiment described in problem 7, data were also collected for caffeine levels. Perform an ANOVA on these data (listed below). What do these results mean? What is similar to and different than the results in question 7?

| Training? | Stress Level | Sleep/ Study | Performance | Caffeine |
|---|---|---|---|---|
| No | Low | Sleep | 8 | 12 |
| No | Low | Sleep | 9 | 13 |
| No | Low | Sleep | 8 | 15 |
| No | Low | Study | 15 | 10 |
| No | Low | Study | 14 | 10 |
| No | Low | Study | 15 | 11 |
| No | High | Sleep | 10 | 14 |
| No | High | Sleep | 11 | 15 |
| No | High | Sleep | 11 | 16 |
| No | High | Study | 18 | 11 |
| No | High | Study | 19 | 10 |
| No | High | Study | 19 | 11 |
| Yes | Low | Sleep | 18 | 11 |
| Yes | Low | Sleep | 17 | 10 |
| Yes | Low | Sleep | 18 | 11 |
| Yes | Low | Study | 10 | 4 |
| Yes | Low | Study | 10 | 4 |
| Yes | Low | Study | 11 | 4 |
| Yes | High | Sleep | 22 | 14 |
| Yes | High | Sleep | 22 | 14 |
| Yes | High | Sleep | 23 | 14 |
| Yes | High | Study | 13 | 5 |
| Yes | High | Study | 13 | 5 |
| Yes | High | Study | 12 | 4 |

# Chapter 15
# Simple Linear Regression

THE **REGRESSION** procedure is designed to perform either simple regression (Chapter 15) or multiple regression (Chapter 16). We split the command into two chapters largely for the sake of clarity. If the reader is unacquainted with *multiple* regression, this chapter, on simple regression, will serve as an introduction. Several things will be covered in the introductory portion of this chapter: (a) the concept of predicted values and the regression equation, (b) the relationship between bivariate correlation and simple regression, (c) the proportion of variance in one variable explained by another, and (d) a test for curvilinear relationships.

Several words of caution are appropriate here: First, a number of thick volumes have been written on regression analysis. We are in no way attempting in a few pages to duplicate those efforts. The introductions of these two chapters are designed primarily to give an overview and a conceptual feel for the regression procedure. Second, in this chapter (and Chapter 16), in addition to describing the standard linear relationships, we explain how to conduct regression that considers curvilinear tendencies in the data. We suggest that those less acquainted with regression should spend the time to thoroughly understand *linear* regression before attempting the much less frequently used tests for curvilinear trends.

## 15.1 Predicted Values and the Regression Equation

There are many times when, given information about one characteristic of a particular phenomenon, we have some idea as to the nature of another characteristic. Consider the height and weight of adult humans. If we know that a person is 7 feet (214 cm) tall, we would suspect (with a fair degree of certainty) that this person probably weighs more than 200 pounds (91 kg). If a person is 4 feet 6 inches (137 cm) tall, we would suspect that such a person would weigh less than 100 pounds (45 kg). There is a wide variety of phenomena in which, given information about one variable, we have some clues about characteristics of another: IQ and academic success, oxygen uptake and ability to run a fast mile, percentage of fast-twitch muscle fibers and speed in a 100-meter race, type of automobile one owns and monetary net worth, average daily caloric intake and body weight, feelings of sympathy toward a needy person and likelihood of helping that person. Throughout a lifetime humans make thousands of such inferences (e.g., he's fat, he must eat a lot). Sometimes our inferences are correct, other times not. Simple regression is designed to help us come to more accurate inferences. It cannot guarantee that our inferences are correct, but it can determine the likelihood or *probability* that our inferences are sound; and given a value for one variable, it can predict the most likely value for the other variable based on available information.

To illustrate regression, we will introduce our example at this time. While it would be possible to use the **grades.sav** file to illustrate simple regression (e.g., the influence of previous GPA on final points), we have chosen to create a new data set that is able to demonstrate both linear regression and curvilinear regression. The new file is called **anxiety.sav** and consists of a data set in which 73 students are measured on their level of pre-exam anxiety on a *none* (1) to *severe* (10) scale, and then measured on a 100-point exam. The hypothesis for a linear relationship might be that those with very low anxiety will do poorly because they don't care much and that those with high anxiety will do better because they are motivated to spend more time in preparation. The dependent (criterion) variable is **exam**, and the independent (predictor) variable is **anxiety**. In other words we are attempting to predict the exam score from the anxiety score. Among other things that regression accomplishes, it is able to create a *regression equation* to calculate a person's *predicted* score on the exam based on his or her anxiety score. The regression equation follows the model of the general equation designed to predict a student's *true* or actual score. The equation for the student's true score follows:

$$\mathbf{exam}_{(\text{true})} = \text{some } \mathbf{constant} + \text{a coefficient} \times \mathbf{anxiety} + \mathbf{residual}$$

That is, the true **exam** score is equal to a **constant** plus some weighted number (coefficient) times the **anxiety** score plus the **residual**. The inclusion of the **residual** (also known as error) term is to acknowledge that *predicted* values in the social sciences are almost never exactly correct and that to acquire a *true* value requires the inclusion of a term that adjusts for the discrepancy between the predicted score and the actual score. This difference is called the *residual*. For instance, the equation based on our data set (with constant and coefficient generated by the regression procedure) follows:

$$\mathbf{exam}_{(\text{true})} = 64.247 + 2.818(\mathbf{anxiety}) + \mathbf{residual}$$

To demonstrate the use of the equation, we will insert the anxiety value for student #24, who scored 6.5 on the anxiety scale.

$$\mathbf{exam}_{(\text{true})} = 64.247 + 2.818(\mathbf{6.5}) + \mathbf{residual}$$

$$\mathbf{exam}_{(\text{true})} = 82.56 + \mathbf{residual}$$

The 82.56 is the student's *predicted* score based on his 6.5 anxiety score. We know that the actual exam score for subject 24 was 94. We can now determine the value of the **residual** (how far off our predicted value was), but we can do this only *after* we know the true value of the dependent variable (**exam** in this case). The residual is simply the true value minus the predicted value (94 – 82.56), or 11.44. The equation with all values inserted looks like this:

$$\begin{array}{ccccc} 94 & = & 82.56 & + & 11.44 \\ \textbf{true score} & = & \textbf{predicted score} & + & \textbf{residual} \end{array}$$

We have included a brief description of the residual term because you will see it so frequently in the study of statistics, but we now turn our attention to the issue of predicted values based on a calculated regression equation. A more extensive discussion of residuals takes place in Chapter 28. The regression equation for the predicted value of **exam** is:

$$\mathbf{exam}_{(\text{predicted})} = 64.247 + 2.818(\mathbf{anxiety})$$

To demonstrate computation, subjects 2, 43, and 72 scored 1.5, 7.0, and 9.0 anxiety points, respectively. Computation of the predicted scores for each of the three follows. Following the predicted value is the actual score achieved by the three subjects (in parentheses), to demonstrate how well (or poorly) the equation was able to predict the true scores:

| | | |
|---|---|---|
| Subject 2: | $\mathbf{exam}_{(\text{predicted})} = 64.247 + 2.818(\mathbf{1.5}) = \mathbf{68.47}$ | (actual score was 52) |
| Subject 43: | $\mathbf{exam}_{(\text{predicted})} = 64.247 + 2.818(\mathbf{7.0}) = \mathbf{83.97}$ | (actual score was 87) |
| Subject 72: | $\mathbf{exam}_{(\text{predicted})} = 64.247 + 2.818(\mathbf{9.0}) = \mathbf{89.61}$ | (actual score was 71) |

We notice that for subject 2, the predicted value was much too high (68.47 vs. 52); for subject 43, the predicted value was quite close to the actual score (83.97 vs. 87); and for subject 72, the predicted value was also much too high (89.61 vs. 71). From this limited observation we might conclude that the ability of our equation to predict values is pretty good for midrange **anxiety** scores, but much poorer at the extremes. Or, we may conclude that there are factors other than a measure of the subject's pre-exam anxiety that influence his or her exam score. The issue of *several* factors influencing a variable of interest is called *multiple* regression and will be addressed in the next chapter.

## 15.2 Simple Regression and the Amount of Variance Explained

We are not left at the mercy of our intuition to determine whether or not a regression equation is able to do a good job of predicting scores. The output generated by the **Regression** command calculates four different values that are of particular interest to the researcher:

1. SPSS generates a score that measures the strength of relationship between the dependent variable (**exam**) and the independent variable (**anxiety**). This score, which can answer the question "How strong is this relationship?" is designated with a capital $R$ and is simply our old friend, the bivariate correlation ($r$). The capital $R$ is used (rather than a lower case $r$) because the **Regression** command is usually used to compute *multiple* correlations (that is, the strength of relationship between several independent variables and a single dependent variable). For a description of correlation, please refer to Chapter 10.
2. $R$ square (or $R^2$) is simply the square of $R$, and is another way to think about effect size or the strength of the relationship between the variables. The $R^2$ value (like the eta squared value in Chapters 12, 13, and 14) is the proportion of variance in one variable accounted for (or explained) by the other variable. For the relationship between **anxiety** and **exam**, SPSS calculated values of $R = .48794$ ($p < .0001$) and $R^2 = .23808$. The $R$ square value indicates that 23.8% of the variance in the **exam** score is accounted for by pretest **anxiety**. The standard disclaimers must be inserted here: With a correlation, be cautious about inferring causation. In this case the *direction* of causation is safe to assume because an exam score cannot influence *pre*-exam anxiety.
3. Along with the computation of $R$, SPSS prints out a probability value ($p$) associated with $R$ to indicate the significance of that association. This answers the question, "How confident can we be that this result is not due to random chance?" Once again, a $p < .05$ is generally interpreted as indicating a statistically significant correlation. If $p > .05$, the strength of association between the two variables is usually not considered statistically significant; whatever relationship between the variables we see, it is likely that the result is due to randomness.
4. SPSS calculates the constant and the coefficient (called $B$ values) for the regression equation. As already noted, the constant and coefficient computed for the regression equation identifying the relationship between **anxiety** and **exam** were 64.247 and 2.818, respectively.

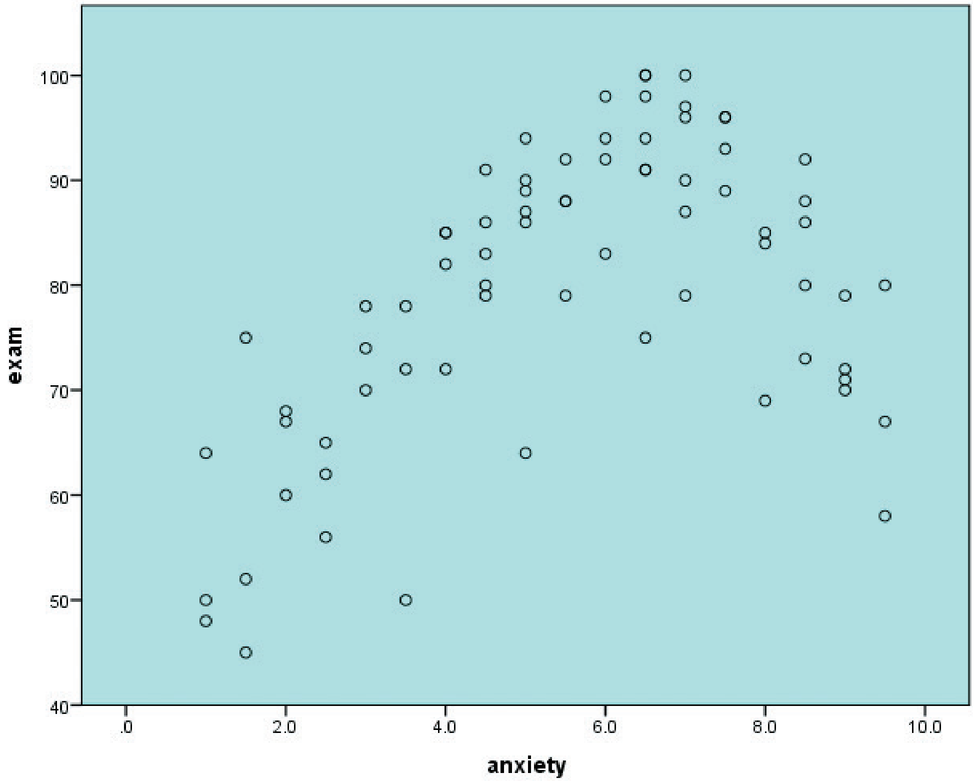
## 15.3 Testing for a Curvilinear Relationship

Most knowledgeable people would consider it foolishness to think that higher pretest anxiety will produce higher scores on an exam. A widely held position is that very low level of anxiety will result in poor scores (due to lack of motivation) and that as anxiety scores increase, motivation to do well increases and higher scores result. However, there comes a point when additional anxiety is detrimental to performance, and at the top end of the anxiety scale there would once again be a decrease of performance. Regression analysis (whether it be simple or multiple) measures a *linear* relationship between the independent variable(s) and the dependent variable. In the fictional data set presented earlier there is a fairly strong linear relationship between **anxiety** and the **exam** score ($R = .484$, $p < .0001$). But perhaps the regression equation would generate more accurate predicted values (yielding a better "fit" of the data) if a quadratic equation was employed that included an **anxiety**-squared (**anxiety$^2$**) term.

Usually, before one tests for a curvilinear trend, there needs to be evidence (theoretical or computational) that such a relationship exists. Frankly, in the social sciences, curvilinear trends happen in only a limited number of circumstances, but they can be critical to understanding the data when they do occur. To demonstrate, we produce the scatterplot between exam and anxiety.

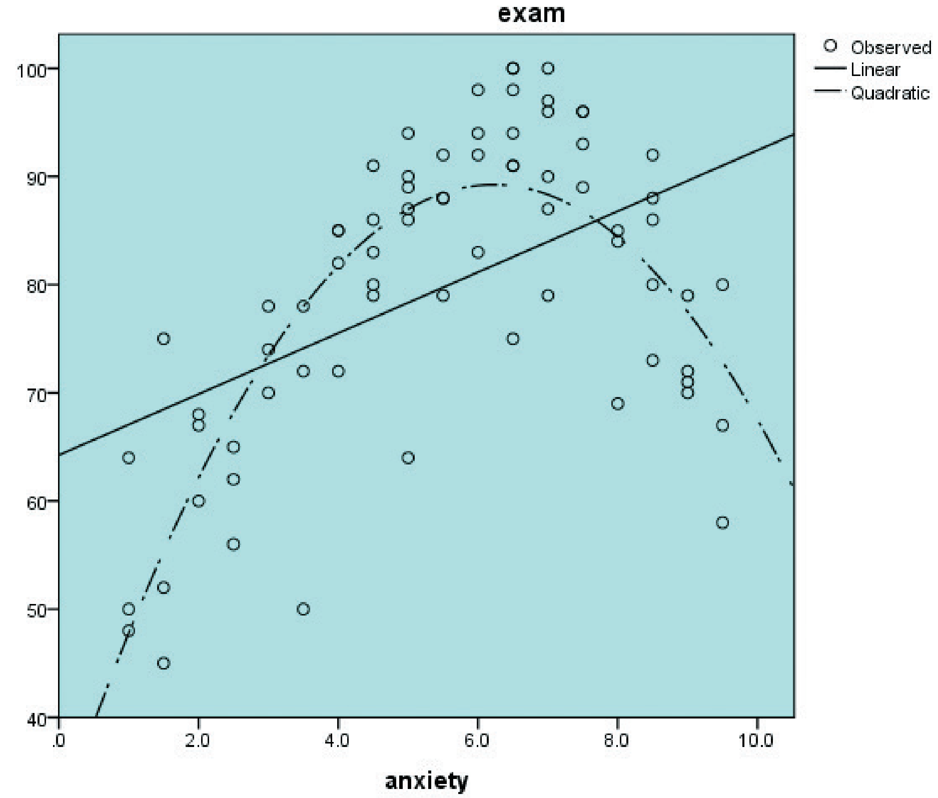
The graph (below) shows the **exam** scores on the vertical axis (the scale ranges from approximately 45 to 105), and **anxiety** on the horizontal axis with a range of 0 to 10. Initial visual inspection reveals what appears to be a curvilinear trend. For mid-range anxiety values the test scores are highest, and at the extremes they tend to be lower. One needs to be cautioned when attempting to read a scattergram. A relationship may appear to exist, but when tested is not statistically significant. More frequently, visual inspection alone does not reveal a curvilinear trend but a statistical test does. When exploring the relationship between two variables, **Regression** is able to reveal whether there is a significant linear trend, a significant curvilinear trend, significant linear *and* curvilinear trends, or neither.

**Screen 15.1** Sample Scatterplot Demonstrating a Curvilinear Trend



A simple procedure offered by SPSS within the context of the **Regression** command is a quick test to check for linear or curvilinear trends titled **Curve Estimation**. You identify the dependent variable (**exam**), the independent variable (**anxiety**), then from the resulting dialog box, select **Linear** and **Quadratic**. An **OK** click will produce a two-line output (shown below) that indicates if linear and/or curvilinear trends exist. The B values are also included so that it is possible to write predicted-value equations for either linear or curvilinear relationships. This process also creates a chart showing the scattergram, the linear regression line (the straight one) and the curvilinear regression line (the curved one). Notice the similarity between the two charts in Screens 15.1 and 15.2. Also note that the constant and coefficients in the equations (below) utilize values from the two lines of output.

**Screen 15.2** Sample Scatterplot Demonstrating a Linear and a Curvilinear Trend



| Dependent | Method | R-square | F | df1 | df2 | Sig of F | Constant | b1 | b2 |
|---|---|---|---|---|---|---|---|---|---|
| EXAM | Linear | .238 | 22.186 | 1 | 71 | .000 | 64.247 | 2.818 | |
| EXAM | Quadratic | .641 | 62.525 | 2 | 70 | .000 | 30.377 | 18.926 | –1.521 |

The linear and curvilinear regression equations now follow:

Linear equation (the straight line): $\textbf{exam}_{(\text{pred})} = 64.25 + 2.82(\textbf{anxiety})$

Quadratic equation (the curved line): $\textbf{exam}_{(\text{pred})} = 30.38 + 18.93(\textbf{anxiety}) + -1.52(\textbf{anxiety})^2$

Observe that in the output (above), the $R^2$ value for the linear equation indicates that **anxiety** explains 23.8% of the **exam** performance, while the $R^2$ value for the quadratic equation (where both the linear and the curvilinear trend influences the outcome) indicates that 64.1% of the variance in **exam** is explained by **anxiety** and the square of

**anxiety**. Under **Sig of F**, the .000 for both the linear and the curvilinear equation indicate that both trends are statistically significant ($p < .001$).

We would like to see if the quadratic equation is more successful at predicting actual scores than was the linear equation. To do so we substitute the **anxiety** values for the same subjects (numbers 2, 43, and 72) used to illustrate the linear equation:

Subject 2: $\textbf{exam}_{\text{(pred)}} = 30.38 + 18.93(\textbf{1.5}) + -1.52(\textbf{1.5})^2 = 55.31$ (actual score, 52)
Subject 43: $\textbf{exam}_{\text{(pred)}} = 30.38 + 18.93(\textbf{7.0}) + -1.52(\textbf{7.0})^2 = 88.30$ (actual score, 87)
Subject 72: $\textbf{exam}_{\text{(pred)}} = 30.38 + 18.93(\textbf{9.0}) + -1.52(\textbf{9.0})^2 = 77.63$ (actual score, 71)

A quick check of the results from the linear equation demonstrates the substantially superior predictive ability of the quadratic equation. Note the table:

| Subject Number | Actual Score | Predicted Linear Score | Predicted Quadratic Score |
|---|---|---|---|
| 2 | 52 | 68.47 | 55.31 |
| 43 | 87 | 83.97 | 88.30 |
| 72 | 71 | 89.61 | 77.63 |

A number of books are available that cover simple, curvilinear, and multiple regression. Several that the authors feel are especially good include Chatterjee and Price (1999), Gonick and Smith (1993), Schulman (1998), Sen and Srivastava (1999), and Weisberg (2005). Please see the reference section for more detailed information on these resources.

# 15.4 Step by Step

## 15.4.1 Simple Linear and Curvilinear Regression

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | → All Programs → IBM SPSS Statistics → IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | → → | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| | Step 2 |
|---|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* | |

*Screens 1 and 2 (displayed on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **anxiety.sav** *file for further analyses:*

| In Screen | Do This | Step 3 |
|---|---|---|
| Front1 | File → Open → Data [*or* ] | |
| Front2 | type `anxiety.sav` → Open [*or* anxiety.sav] | Data |

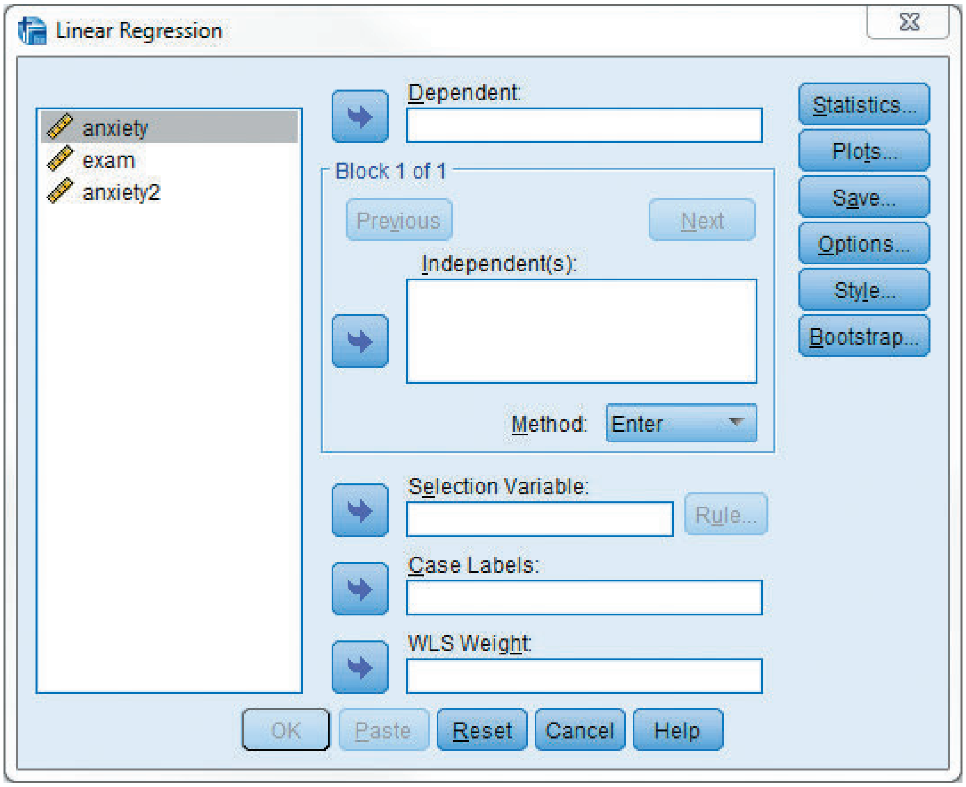
Whether first entering SPSS or returning from earlier operations the standard menu of commands across the top is required. As long as it is visible you may perform any analysis. It is not necessary for the data window to be visible.

*After completion of Step 3, a screen with the desired menu bar appears. When you click a command (from the menu bar), a series of options will appear (usually) below the selected command. With each new set of options, click the desired item. The sequence to access linear regression begins at any screen with the menu of commands visible:*

| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | **Analyze** → **Regression** → **Linear** | 15.3 |

At this point, a new dialog box opens (Screen 15.3, below) that allows you to conduct regression analysis. Because this box is much more frequently used to conduct *multiple* regression analysis than simple linear regression, there are many options available that we will not discuss until next chapter. A warning to parents with young children, however: Under no circumstances allow a child less than 13 years of age to click on the **Statistics** or **Plots** pushbuttons. Windows open with options so terrifying that some have never recovered from the trauma. The list to the left will initially contain only two variables, **anxiety** and **exam**. The procedure is to select **exam** and paste it into the **Dependent** box, select **anxiety** and paste it into the **Independent(s)** box, then click on the **OK** button. The program will then run yielding $R$ and $R^2$ values, $F$ values and tests of significance, the $B$ values that provide constants and coefficients for the regression equation, and Beta ($\beta$) values to show the strength of association between the two variables. Some of the terms may be unfamiliar to you and will be explained in the output section. The step-by-step sequence follows Screen 15.3.

**Screen 15.3** The Initial Linear Regression Window



*To conduct simple linear regression with a dependent variable of* **exam** *and an independent variable of* **anxiety** *perform the following sequence of steps. Begin with Screen 15.3; results from this analysis are in the Output section.*

| In Screen | Do This | Step 5 |
|---|---|---|
| 15.3 | ✳ **exam** → ✳ *upper* ➡ → ✳ **anxiety** → ✳ *second* ➡ → ✳ **OK** | Back1 |

## 15.4.2 Curvilinear Trends

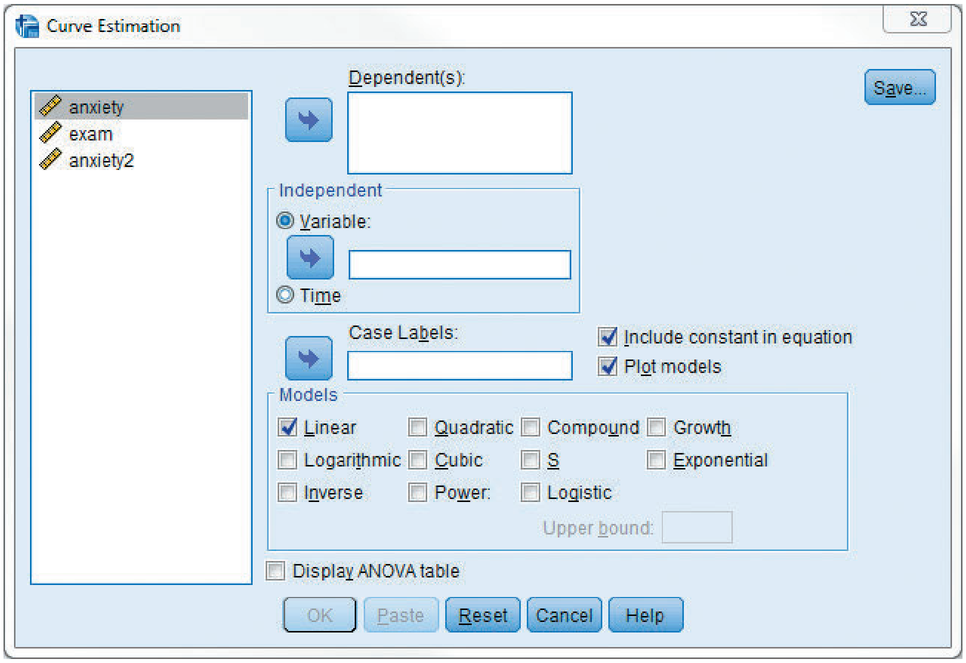
We will show two sequences that deal with curvilinear trends. The first deals with the simple two-line output reproduced in the introduction, and creation of the graph that displays linear and curvilinear trends (also displayed in the introduction). The second considers the more formal procedure of creating a quadratic variable that you may use in a number of different analyses. While much of what takes place in the second procedure could be produced by the first, we feel it is important that you understand what is really taking place when you test for a curvilinear trend. Furthermore, although we present curvilinear trends in the context of simple regression, the same principles (and access procedures) apply to the next chapter on multiple regression.

*To access the* **Curve Estimation** *chart requires a Step-4a sequence.*

| In Screen | Do This | Step 4a |
|---|---|---|
| Menu | ✳ **Analyze** → **Regression** → ✳ **Curve Estimation** | Back1 |

A new dialog box now opens (Screen 15.4) that allows a number of options for curve estimation. The early procedure is identical to that shown in the previous sequence of steps (sequence Step 5): Select **exam** and paste it into the **Dependent(s)** box, select **anxiety** and paste it into the **Independent** box. After the dependent and independent variables are selected, notice that three of the options in this dialog box are already selected by default: **Include constant in equation** provides the constant, necessary to create a regression equation. **Plot models** creates the graph that was illustrated in the introduction. The **Linear** model refers to testing for and including as a line on the graph any linear trend in your data.

**Screen 15.4** The Curve Estimation window



Curve Estimation

anxiety
exam
anxiety2

Dependent(s):

Independent
● Variable:
○ Time

Case Labels:

☑ Include constant in equation
☑ Plot models

Models
☑ Linear ☐ Quadratic ☐ Compound ☐ Growth
☐ Logarithmic ☐ Cubic ☐ S ☐ Exponential
☐ Inverse ☐ Power: ☐ Logistic
Upper bound:

☐ Display ANOVA table

Within the **Models** box a variety of toys exist for the mathematical or statistical wizards. All of the curvilinear models included here have their unique applications, but none are used frequently, and only the **Quadratic** option will be addressed here. For the present analysis, keep the already-selected **Linear** option, and also select **Quadratic**. The steps now follow, beginning with Screen 15.4.

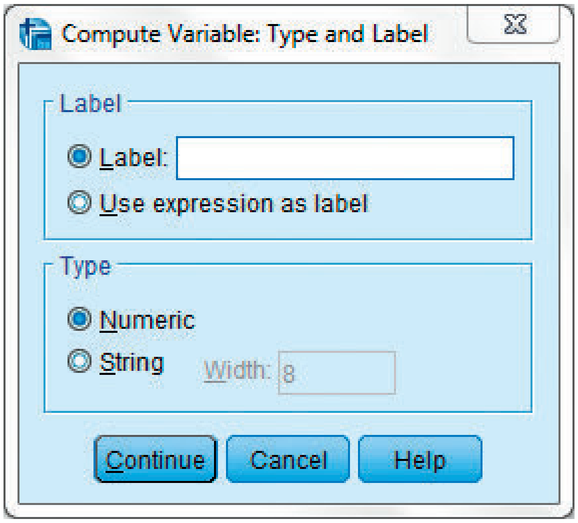
*Beginning with a dependent variable* **exam** *and checking for linear and curvilinear trends in the independent variable (***anxiety***), perform the following sequence of steps. Perform the Step 4a sequence to arrive at Screen 15.4. Results of this analysis are in the Introduction.*

| In Screen | Do This | Step 5a |
|---|---|---|
| 15.4 | ✳ **exam** → ✳ *upper* ➡ → ✳ **anxiety** → ✳ *middle* ➡ → ✳ **Quadratic** → ✳ **OK** | **Back1** |

What actually takes place when you formulate the regression equation that includes the influence of a quadratic term is the creation of a new variable that is the square of the independent variable (**anxiety**), or **anxiety**$^2$. You all vividly remember from your Algebra I days that the inclusion of a variable squared produces a parabola, that is, a curve with a single change of direction that opens either upward (if the coefficient is positive) or downward (if the coefficient is negative). Notice in the equation on page 195 that the coefficient of the squared term is negative [–1.52(**anxiety**)$^2$], and that the parabola opens downward. Also note the influence of the linear trend (a tendency of data points to go from lower left to upper right) is also reflected in the curve. The left end of the curve is lower than the right end of the curve. Thus both linear and curvilinear (quadratic) trends are reflected in the graph. For the other models available in Screen 15.4—exponential, cubic, etc.—pull out your high school algebra homework plots of these functions and see if your scatterplots look similar.

To go beyond this simple test for both linear and curvilinear trends it is necessary to create a new variable, the square of anxiety, assigned a variable name of **anxiety2**. Then when we begin the regression procedure two variables will be included as independent variables, **anxiety** and **anxiety2**. The method of creating new variables is described in Chapter 4, and we will take you step by step through the sequence but will not reproduce the screen here (Screen 4.3 is on page 66 if you wish a visual reference). We do access one window that has not been displayed previously, that is the window that appears when you click on the **Type & Label** button displayed in Screen 4.3. The use of the box is self-explanatory and is reproduced (below) for visual reference only.

**Screen 15.5** The Compute Variable: Type and Label Window



Compute Variable: Type and Label

Label

● Label:

○ Use expression as label

Type

● Numeric

○ String  Width: 8

The step-by-step sequence of creating a new variable, **anxiety2**, and including it along with **anxiety** as independent variables that influence the dependent variable **exam** now follows. This is now *multiple* regression because more than one independent variable is included. (Note: If you downloaded the **anxiety.sav** file, the **anxiety2** variable is already included.)

*To run the regression procedure with a dependent variable of* **exam** *and independent variables of* **anxiety** *and* (**anxiety**)$^2$ *perform the following sequence of steps. Begin at the Data Editor screen with the menu of commands showing. The procedure begins by creating the new variable* **anxiety2**. *Results from this analysis are included in the Output section.*

| In Screen | Do This | Step 5b |
|---|---|---|
| Data | ✳ **Transform** → ✳ **Compute Variable** | |
| 4.3 | **type** `anxiety2` → ✳ **Type & Label** | |
| 15.5 | ✳ *inside the* **Label** *box* → **type** `square of anxiety` → ✳ **Continue** | |
| 4.3 | ✳ **anxiety** → ✳ ➡ → **type** `**` → **type** `2` → ✳ **OK** | |
| Menu | ✳ **Analyze** → ✳ **Regression** → ✳ **Linear** | |
| 15.3 | ✳ **exam** → ✳ *upper* ➡ → ✳ **anxiety** → ✳ *second* ➡ → ✳ **anxiety2** → | |
| | ✳ *second* ➡ → ✳ **OK** | Back1 |

Upon completion of Step 5, 5a, or 5b, the output screen will appear (Screen 1, inside back cover). All results from the just-completed analysis are included in the Output Navigator. Make use of the scroll bar arrows (▲▼▶◀) to view the results. Even when viewing output, the standard menu of commands is still listed across the top of the window. Further analyses may be conducted without returning to the data screen.

## 15.5 Printing Results

Results of the analysis (or analyses) that have just been conducted require a window that displays the standard commands (**File Edit Data Transform Analyze . . .**) across the top. A typical print procedure is shown below beginning with the standard output screen (Screen 1, inside back cover).

*To print results, from the Output screen perform the following sequence of steps:*

| In Screen | Do This | Step 6 |
|---|---|---|
| Back1 | *Select desired output or edit (see pages 16–22)* → ✳ **File** → ✳ **Print** | |
| 2.9 | *Consider and select desired print options offered, then* → ✳ **OK** | |

*To exit you may begin from any screen that shows the* **File** *command at the top.*

| In Screen | Do This | Step 7 |
|---|---|---|
| Menu | ✳ **File** → ✳ **Exit** | 2.1 |

Note: After clicking **Exit**, there will frequently be small windows that appear asking if you wish to save or change anything. Simply click each appropriate response.

# 15.6 Output

## 15.6.1 Simple Linear and Curvilinear Regression Analysis

What follows is slightly condensed output from sequence Step 5 on page 198.

**Variables Entered/Removed[b]**

| Model | Variables Entered | Variables Removed | Method |
|---|---|---|---|
| 1 | anxiety[a] | . | Enter |

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .488[a] | .238 | .277 | 12.215 |

**ANOVA**

| Source of Variation | df | Sum of Squares | Mean Square | F | Signif F |
|---|---|---|---|---|---|
| Regression | 1 | 3310.476 | 3310.476 | 22.186 | .000 |
| Residual | 71 | 10594.209 | 149.214 | | |

**Coefficients**

| | Unstandardized coefficients | | Std Coefficients | | |
|---|---|---|---|---|---|
| | B | Std. Error | Beta | t | Signif of t |
| (Constant) | 64.247 | 3.602 | | 17.834 | .000 |
| ANXIETY | 2.818 | .598 | .488 | 4.710 | .000 |

This output shows that there is a significant linear relationship between exam performance and anxiety such that a higher level of anxiety results in higher scores. The specific meaning of this output is summarized in the definitions of terms that follow.

| Term | Definition/Description |
|---|---|
| **R** | Since there is only one independent variable, this number is the bivariate correlation (*r*) between **exam** and **anxiety**. It is both the test statistic of the regression equation, and the effect size to indicate how strong the relationship is between the dependent variable and the variable that predicts it. |
| **R SQUARE** | The **R SQUARE** value identifies the proportion of variance in **exam** accounted for by **ANXIETY**. In this case 23.8% of the variance in **exam** is explained by **anxiety**. |
| **ADJUSTED R SQUARE** | **R SQUARE** is an accurate value for the sample drawn but is considered an optimistic estimate for the population value. The **ADJUSTED R SQUARE** is considered a better population estimate. |
| **STANDARD ERROR** | The standard deviation of the expected values for the dependent variable, **exam**. |
| **REGRESSION** | Statistics relating to the explained portion of the variance. |
| **RESIDUAL** | Statistics relating to the *un*explained portion of the variance. |
| **DF** | Degrees of freedom: For regression, the number of independent variables (1 in this case). For the residual, the number of subjects (73) minus the number of independent variables (1), minus 1: (73 – 1 – 1 = 71). |
| **SUM OF SQUARES** | For regression this is the between-groups sum of squares; for residual, the within-groups sum of squares. Note that in this case there is a larger portion of unexplained variance than there is of explained variance, a reality also reflected in the $R^2$ value. |

| | |
|---|---|
| **MEAN SQUARE** | Sum of squares divided by degrees of freedom. |
| **F** | Mean square regression divided by mean square residual. |
| **SIGNIF F** | Likelihood that this result could occur by chance. |
| **B** | Coefficient and constant for the linear regression equation:<br>**exam**$_{(pred)}$ **= 64.247 + 2.818(anxiety).** |
| **STD ERROR** | Standard error of *B*: A measure of the stability or sampling error of the *B* values. It is the standard deviation of the *B* values given a large number of samples drawn from the same population. |
| **BETA** | The standardized regression coefficients. This is the *B* value for standardized scores (*z*-scores) of the variable **anxiety**. This value will always vary between ±1.0 in linear relationships. For curvilinear relationships it will sometimes extend outside that range. |
| **t** | *B* divided by the standard error of *B*. |
| **SIGNIF t** | Likelihood that this result could occur by chance. |

## 15.7 A Regression Analysis that Tests for a Curvilinear Trend

What follows is slightly condensed output from sequence Step 5b on page 200.

Variables Entered/Removed[b]

| Model | Variables Entered | Variables Removed | Method |
|---|---|---|---|
| 1 | anxiety2, anxiety | . | Enter |

Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .801[a] | .641 | .631 | 8.443 |

ANOVA

| Source of Variation | df | Sum of Squares | Mean Square | F | Signif F |
|---|---|---|---|---|---|
| Regression | 2 | 8194.538 | 4457.269 | 62.525 | .0000 |
| Residual | 70 | 4990.146 | 71.288 | | |

Coefficients

| | Unstandardized coefficients | | Std Coefficients | | |
|---|---|---|---|---|---|
| | B | Std. Error | Beta | t | Signif of t |
| (Constant) | 30.377 | 4.560 | | 6.662 | .000 |
| ANXIETY | 18.926 | 1.836 | 3.277 | 10.158 | .000 |
| ANXIETY2 | -1.521 | .172 | -2.861 | -8.866 | .000 |

The terms described previously also apply to *this* output. Notice the high multiple *R* value (.80), indicating a very strong relationship between the independent variable (**anxiety**) and its square (**anxiety2**) and the dependent variable, **exam**. An $R^2$ value of .641 indicates that 64.1% of the variance in **exam** is accounted for by **anxiety** and its square. The *F* and associated *p* values (Signif F and Sig T) reflect the strength of the *overall* relationship between both independent variables and **exam** (F and Signif F), and between *each individual* independent variable and **exam** (t and Signif t). In Chapter 16 we will address in some detail the influence of several variables (there are two here) on a dependent variable. Also note that while the betas always vary between ±1 for simple linear equations, they extend beyond that range in this quadratic equation. Beneath the *B* in the lower table are the coefficients for the regression equation. Notice that the regression equation we tested (presented in the Introduction) had a constant of 30.377, an **anxiety** coefficient of 18.926, and an (**anxiety**)$^2$ coefficient of −1.521.

# Exercises

Answers to selected exercises are downloadable at **www.spss-step-by-step.net.**

1. Use the **anxiety.sav** file for exercises that follow (downloadable at the address above).

Perform the 4a–5a sequences on pages 198 and 199.

- Include output in as compact a form as is reasonable
- Write the linear equation for the predicted exam score
- Write the quadratic equation for the predicted exam score

For subjects numbered 5, 13, 42, and 45

- Substitute values into the two equations and solve. Show work on a separate page.
- Then compare in a small table (shown below and similar to that on page 196)
  - → The anxiety score for each subject
  - → Linear equation results,
  - → Quadratic equation results, and
  - → Actual exam scores for sake of comparison.

| Subject # | Anxiety Score | Predicted Linear Score | Predicted Quadratic Score | Actual Exam Score |
|---|---|---|---|---|
| 5 | | | | |
| 13 | | | | |
| 42 | | | | |
| 45 | | | | |

2. Now using the **divorce.sav** file, test for linear and curvilinear relations between:

- physical closeness (**close**) and life satisfaction (**lsatisy**)
- attributional style (**asq**) and life satisfaction (**lsatisy**)

Attributional style, by the way, is a measure of optimism—a low score is "pessimistic" and a high score is "optimistic."

Print graphs and write linear and quadratic equations for both.

For each of the three analyses in problems 3 and 4:

- Print out the results
- Box the Multiple R,
- Circle the R Square,
- Underline the three (3) B values, and
- Double underline the three (3) Signif of t values.

Now identify what is the meaning/function of (a) **Multiple R**, (b) **R Square**, (c) the **B values**, and (d) the **Signif of t** values.

3. Using the **anxiety.sav** file, perform Step 5b (page 200) demonstrating the influence of **anxiety** and anxiety squared (**anxiety2**) on the exam score (**exam**).

4. Now, complete similar procedures for the two relationships shown in problem 2 (from the **divorce.sav** file) and perform the five steps bulleted above: Specifically,

- the influence of closeness (**close**) and closeness squared (**close2**) on life satisfaction (**lsatisy**), and
- the influence of attributional style (**asq**) and the square of attributional style (**asq2**) on life satisfaction (**lsatisy**).

5. A researcher is examining the relationship between stress levels and performance on a test of cognitive performance. She hypothesizes that stress levels lead to an increase in performance to a point, and then increased stress decreases performance. She tests 10 participants, who have the following levels of stress: 10.94, 12.76, 7.62, 8.17, 7.83, 12.22, 9.23, 11.17, 11.88, and 8.18. When she tests their levels of mental performance, she finds the following cognitive performance scores (listed in the same participant order as above): 5.24, 4.64, 4.68, 5.04, 4.17, 6.20, 4.54, 6.55, 5.79, and 3.17. Perform a linear regression to examine the relationship between these variables. What do these results mean?

6. The same researcher tests 10 more participants, who have the following levels of stress: 16, 20, 14, 21, 23, 19, 14, 20, 17, and 10. Their cognitive performance scores are (listed in the same participant order) 5.24, 4.64, 4.68, 5.04, 4.17, 6.20, 4.54, 6.55, 5.79, and 3.17. (Note that in an amazing coincidence, these participants have the same cognitive performance scores as the participants in question 5; this coincidence may save you some typing.) Perform a linear regression to examine the relationship between these variables. What do these results mean?

7. Create a scatterplot (see Chapter 5) of the variables in question 6. How do results suggest that linear regression might not be the best analysis to perform?

8. Perform curve estimation on the data from question 6. What does this tell you about the data that you could not determine from the analysis in question 6?

9. What is different about the data in questions 5 and 6 that leads to different results?

# Chapter 16
# Multiple Regression Analysis

MULTIPLE REGRESSION is the natural extension of simple linear regression presented in Chapter 15. In simple regression, we measured the amount of influence one variable (the independent or predictor variable) had on a second variable (the dependent or criterion variable). We also computed the constant and coefficient for a *regression equation* designed to predict the values of the dependent variable, based on the values of the independent variable. While simple regression shows the influence of *one* variable on another, multiple regression analysis shows the influence of *two or more* variables on a designated dependent variable.

Another way to consider regression analysis (simple or multiple) is from the viewpoint of a slope-intercept form of an equation. When a simple correlation between two variables is computed, the intercept and slope of the *regression line* (or line of best fit) may be requested. This line is based on the regression equation mentioned in the previous chapter (pages 191–193) with the y-intercept determined by the constant value and the slope determined by the coefficient of the independent variable. We describe here a simple regression equation as a vehicle for introducing multiple regression analysis. To assist in this process we present a new example based on a file called **helping1.sav**. This data file is related to a study of helping behavior; it is real data derived from a sample of 81 subjects.

## 16.1 The Regression Equation

In this chapter we introduce the **helping1.sav** data mentioned above. Two of the variables used to illustrate simple regression are **zhelp** (a measure of total amount of time spent helping a friend with a problem, measured in *z* scores) and **sympathy** (the amount of sympathy felt by the helper in response to the friend's problem, measured on a *little* (1) to *much* (7) scale). Although correlation is often bidirectional, in this case **zhelp** is designated as the dependent variable (that is, one's sympathy influences how much help is given *rather than* the amount of help influencing one's sympathy). A significant correlation ($r = .46$, $p < .0001$) was computed, demonstrating a substantial relationship between the amount of sympathy one feels and the amount of help given. In addition, the intercept value (−1.892) and the slope (.498) were also calculated for a regression equation showing the relationship between the two variables. From these numbers we can create the formula to determine the *predicted value* of **zhelp**:

$$\textbf{zhelp}_{\text{(predicted)}} = -1.892 + .498(\textbf{sympathy})$$

If a person measured 5.6 on the **sympathy** scale, the predicted value for that person for **zhelp** would be .897. In other words, if a person measured fairly high in sympathy (5.6 in this case), it is anticipated that he or she would give quite a lot of help (a *z* score of .897 indicates almost one standard deviation more than the average amount of help). The sequence that follows illustrates the computation of this number:

$$\textbf{zhelp}_{(\text{pred})} = -1.892 + .498(\textbf{5.6}) \rightarrow \textbf{zhelp}_{(\text{pred})} = -1.892 + 2.789 \rightarrow \textbf{zhelp}_{(\text{pred})} = \textbf{.897}$$

Multiple regression analysis is similar but allows *more than* one independent variable to have an influence on the dependent variable.

In this example, two other measures were also significantly correlated with **zhelp: anger** (angry or irritated emotions felt by the helper toward the needy friend, on a *none* (1) to *much* (7) scale) and **efficacy** (self-efficacy, or the helper's belief that he or she had the resources to be of help to the friend, on a *little* (1) to *much* (7) scale). The multiple regression analysis generated the following *B* values (The *B* can be roughly translated as the *slope* or *weighted constant* for the variable of interest.): $B_{(\text{sympathy})} = .4941$, $B_{(\text{anger})} = .2836$, and $B_{(\text{efficacy})} = .4125$, and the constant (intercept) = −4.3078. From these numbers a new equation may be generated to determine the predicted value for **zhelp**:

$$\textbf{zhelp}_{(\text{predicted})} = -4.308 + .494(\textbf{sympathy}) + .284(\textbf{anger}) + .412(\textbf{efficacy})$$

Inserting numbers from an actual subject, subject 9 in this case:

$$\textbf{zhelp}_{(\text{predicted})} = -4.308 + .494(\textbf{3.5}) + .284(\textbf{1.0}) + .412(\textbf{2.9}) = \textbf{-1.09}$$

This result suggests that this person who measured midrange on sympathy (3.5), low in anger (1.0), and low in self-efficacy (2.9) would be expected to give little help (a *z* score of −1.09 is more than one standard deviation below average). Subject 9's *actual* **zhelp** score was −.92. The prediction in this case was fairly close to accurate.

A *positive* value for one of the *B* coefficients indicates that a higher score on the associated variable will increase the value of the dependent variable (i.e., more sympathy yields more help). A *negative* coefficient on a predictor variable would *decrease* the value of the dependent variable (the equation above does not illustrate this; an example might be *more* cynicism yields *less* help). The greater the *B* value (absolute values), the greater the influence on the value of the dependent variable. The smaller the *B* value (absolute values), the less influence that variable has on the dependent variable.

However, *B* values often cannot be compared directly because different variables may be measured on different scales, or have different *metrics*. To resolve this difficulty, statisticians have generated a standardized score called *Beta* (β), which allows for direct comparison of the relative strengths of relationships between variables. β varies between ±1.0 and is a *partial correlation*. A partial correlation is the correlation between two variables in which the influences of all other variables in the equation have been partialed out. In the context of the present example, the *Beta* between **anger** and **zhelp** is the correlation between the two variables *after* **sympathy** and **efficacy** have been entered and the variability due to the subjects' sympathy and efficacy have already been calculated. Thus *Beta* is the *unique* contribution of one variable to explain another variable. The *Beta* weight, often called the *standardized regression coefficient*, is not only an important concept in regression analysis but also the construct used in structural equation modeling to show the magnitude and direction of the relationships between all variables in a model. (Structural equation modeling is becoming increasingly popular in social science research but requires the purchase of an additional SPSS module.)

In the previous equation, we find, as expected, that higher **sympathy** and higher **efficacy** scores correlated with higher **zhelp** scores. Contrary to intuition, we find that more **anger** also correlated positively with **zhelp**, that is, when one is angry one helps more! Why this is true is a matter of discussion and interpretation of the researcher. When an unexpected result occurs, the researcher would be well advised to recheck their data to ensure that variables were coded and entered correctly. SPSS gives no clue as to why analyses turn out the way they do.

## 16.2 Regression And $R^2$: The Amount of Variance Explained

In multiple regression analysis, any number of variables *may* be used as predictors, but many variables are not necessarily the ideal. It is important to find variables that *significantly* influence the dependent variable. SPSS has procedures by which only *significant* predictors will be entered into the regression equation. With the **Forward** entry method a dependent variable and any number of predictor (independent) variables are designated. **Regression** will first compute which predictor variable has the highest bivariate correlation with the dependent variable. SPSS will then create a regression equation with this one selected independent variable. This means that after the first step, a regression equation has been calculated that includes the designated dependent variable and *only one* independent variable. Then **Regression** will enter the second variable, a variable that explains the greatest amount of *additional* variance. This second variable will be included only if it explains a *significant* amount of additional variation. After this second step, the regression equation has the same dependent variable but now has *two* predictor variables. Then, if there is a third variable that significantly explains more of the variance, it too will be included in the regression equation. This process will continue until no additional variables significantly explain additional variance. By default, **Regression** will cease to add new variables when $p$ value associated with the inclusion of an additional variable increases above the .05 level of significance. The researcher, however, may designate a different level of significance as a criterion for entry into the equation.

The measure of the strength of relationship between the independent variables (note, plural) and the dependent variable is designated with a capital $R$ and is usually referred to as *multiple R*. This number squared ($R^2$) yields a value that represents the proportion of variation in the dependent variable that is explained by the independent variables. $R^2$ is the most commonly used measure of the overall effect size of the independent or predictor variables on the dependent variable. In the regression analysis that produced the regression equation shown in previous page, the value of multiple $R$ was .616, and the $R^2$ was .380. This indicates that 38% of the variance in **zhelp** was accounted for by **sympathy**, **anger**, and **efficacy**. See the Output section of this chapter for additional information about multiple $R$.

## 16.3 Curvilinear Trends, Model Building, and References

Regression, like correlation, tests for *linear* relationships. In the previous chapter we described the procedure to test for a *curvilinear* relationship. The same process operates with multiple regression. If there is theoretical or statistical evidence that one or more of the predictor variables demonstrates a curvilinear association with the dependent variable, then a quadratic (the variable squared) factor may be added as a predictor. Please refer to the previous chapter for an explanation of this process and other possible nonlinear (curved) relationships.

A final critical item concerns model building, that is, conducting a regression analysis that is conceptually sound. Certain fundamental criteria are necessary for creating a reliable model:

1. Your research must be thoughtfully crafted and carefully designed. The arithmetic of regression will not correct for either meaningless relationships or serious design flaws. Regrettably, many details of what constitutes research that is "thoughtfully crafted and carefully designed" extend well beyond the scope of this book.

2. The sample size should be large enough to create meaningful correlations. There are no hard rules concerning acceptable sample size, but as *N* drops below 50, the validity of your results become increasingly questionable. Also there is the consideration of the number of variables in the regression equation. The more variables involved, the larger the sample size needs to be to produce meaningful results.
3. Your data should be examined carefully for outliers or other abnormalities.
4. The predictor variables should be approximately normally distributed, ideally with skewness and kurtosis values between ±1. However, good results can often be obtained with an *occasional* deviation from normality among the predictor variables, or even the inclusion of a discrete variable (such as gender). A normal distribution for the dependent variable is also urged; but discriminant analysis (Chapter 22) uses a *discrete* measure as the criterion variable in a regression procedure.
5. Be keenly aware of the issue of linear dependency between the predictor variables. Never use two variables when one is partially or entirely dependent upon the other (such as points on the final and total points in the class). Also avoid variables that are conceptually very similar (such as worry and anxiety). To compute a matrix of correlations between potential predictor variables is always wise. Variables that correlate higher than $r = .5$ should be scrutinized carefully before both are included in a regression analysis. The power, stability, and interpretability of results are substantially compromised when variables that are linearly dependent are included in the analysis. This problem is sometimes called "collinearity" (or even "multicollinearity," if "collinearity" is too simple for you.)

Multiple regression analysis is not a simple procedure. Like analysis of variance, a number of thick volumes have been written on the topic, and we are in no way attempting to duplicate those efforts. It is suggested that, before you attempt to conduct multiple regression analysis, you take a course in the subject. A number of books are available that cover both simple and multiple regression. Several that the authors feel are especially good include Chatterjee and Price (1999), Gonick and Smith (1993), Schulman (1998), Sen and Srivastava (1999), and Weisberg (2005). Please see the reference section for more detailed information.

The purpose of the previous four pages has been to remind you of the rationale behind regression if you have been away from it for a while. The purpose of the pages that follow is to explain step by step how to conduct multiple regression analysis with SPSS and how to interpret the output.

The data we use by way of example are from the study already discussed on pages 204–205. The file name is **helping1.sav**, $N = 81$. The following variables will be used in the description; all variables except **zhelp** are measured on a *little* (1) to *much* (7) scale.

- **zhelp**: The dependent variable. The standardized score (*z* score) for the amount of help given by a person to a friend in need on a −3 to +3 scale.
- **sympathy**: Feelings of sympathy aroused in the helper by the friend's need.
- **anger**: Feelings of anger or irritation aroused in the helper by the friend's need.
- **efficacy**: Self-efficacy of the helper in relation to the friend's need.
- **severity**: Helper's rating of how severe the friend's problem was.
- **empatend**: Empathic tendency of the helper as measured by a personality test.

# 16.4 Step by Step

## 16.4.1 Simple Linear and Curvilinear Regression

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | ✳ → ✳ All Programs ▷ → ✳ IBM SPSS Statistics → ✳ IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | ✳ → ✳ → ✳ | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| | Step 2 |
|---|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* | |

*Screens 1 and 2 (displayed on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **helping1.sav** *file for further analyses:*

| In Screen | Do This | Step 3 |
|---|---|---|
| Front1 | ✳ File → Open → ✳ Data [*or* ✳ ] | |
| Front2 | type `helping1.sav` → ✳ Open [*or* ✳ ✳ helping1.sav] | Data |

Whether first entering SPSS or returning from earlier operations the standard menu of commands across the top is required. As long as it is visible you may perform any analyses. It is not necessary for the data window to be visible.

*After completion of Step 3 a screen with the desired menu bar appears. When you click a command (from the menu bar), a series of options will appear (usually) below the selected command. With each new set of options, click the desired item. The sequence to access multiple regression analysis begins at any screen with the menu of commands visible:*

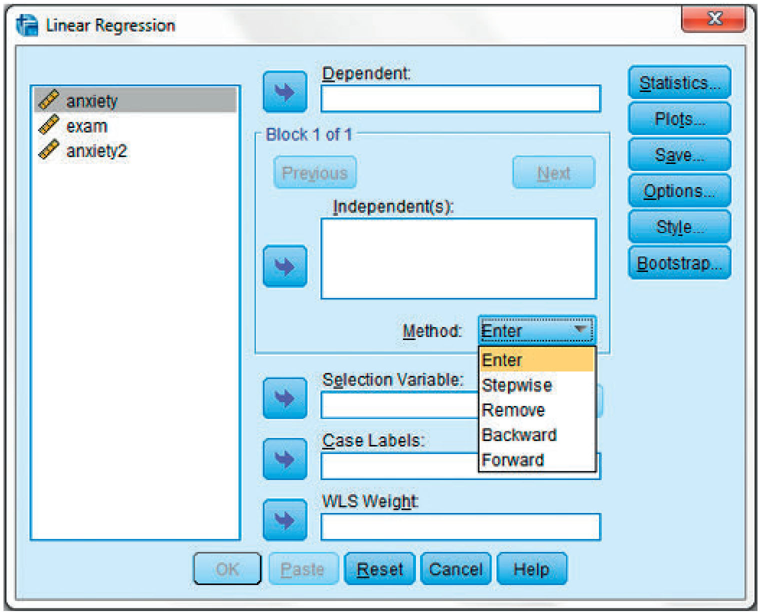
| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | ✳ Analyze → Regression → ✳ Linear | 16.1 |

The screen that appears at this point (Screen 16.1, next page) is identical to the dialog box shown (Screen 15.1) from the previous chapter. We reproduce it here for sake of visual reference and to show the new list of variables in the **helping1.sav** file.

At the top of the window is the **Dependent** box that allows room for a single dependent variable. In the middle of the screen is the **Independent(s)** box where one or more carefully selected variables will be pasted. There are no SPSS restrictions to the *number* of independent variables you may enter; there are common sense restrictions described in the introduction. The **Block 1 of 1** option allows you to create and conduct more than one regression analysis in a single session. After you have chosen the dependent variable and independent variables, and selected all options that you desire related to an analysis, click the **Next** pushbutton located below **Block 1 of 1**. The dependent variable will remain, all selected options and specifications will remain, but the independent variables will revert to the variable list. Then you may select specifications for a second regression analysis. You may specify as many different analyses in a session as you like; a click of the **OK** pushbutton will run them in the same order as they were created. Several additional options are described in the paragraphs that follow.

**Screen 16.1** The Linear Regression Window

**NOTE:** The five options to the right of **Method** are activated.



The menu box labeled **Method** is an important one. It identifies five different methods of entering variables into the regression equation. With a click of the ▼, the five options appear. **Enter** is the default method of entering variables. Each one is described below:

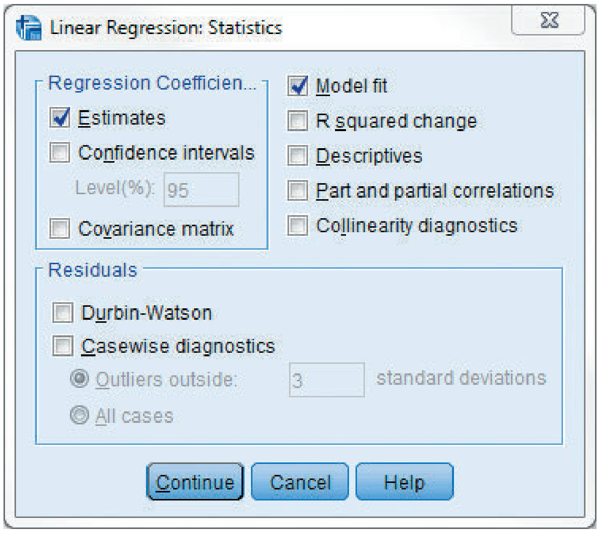
- **Enter:** This is the forced entry option. SPSS will enter at one time all specified variables regardless of significance levels.
- **Stepwise:** This method combines both **Forward** and **Backward** procedures. Due to the complexity of intercorrelations, the variance explained by certain variables will change when new variables enter the equation. Sometimes a variable that qualified to enter loses some of its predictive validity when other variables enter. If this takes place, the **Stepwise** method will remove the "weakened" variable. **Stepwise** is probably the most frequently used of the regression methods.
- **Remove:** This is the forced removal option. It requires an initial regression analysis using the **Enter** procedure. In the next block (**Block 2 of 2**) you may specify one or more variables to remove. SPSS will then remove the specified variables and run the analysis again. It is also possible to remove variables one at a time for several blocks.
- **Backward:** This method enters *all* independent variables at one time and then removes variables one at a time based on a preset significance value to remove. The default value to remove a variable is $p \geq .10$. When there are no more variables that meet the requirement for removal, the process ceases.
- **Forward:** This method will enter variables one at a time, based on the designated significance value to enter. The process ceases when there are no additional variables that explain a significant portion of additional variance.

The **WLS Weight** box (Weighted Least Squares) allows you to select a single variable (not already designated as a predictor variable) that weights the variables prior to analysis. This is an infrequently used option.

The **Plots** option deals with plots of residuals only and will be considered in the final chapter of the book (Chapter 28).

A click on **Statistics** opens a small dialog box (Screen 16.2). Two options are selected by default. **Estimates** will produce the *B* values (used as coefficients for the regression equation), *Beta*s (the standardized regression coefficients) and associated standard errors, *t* values, and significance values. The **Model fit** produces the Multiple *R*, $R^2$, an ANOVA table, and associated *F* ratios and significance values. These two options represent the essence of multiple regression output.

Screen 16.2 The Linear Regression: Statistics Window



Regression Coefficien...

☑ Estimates

☐ Confidence intervals

Level(%): 95

☐ Covariance matrix

☑ Model fit

☐ R squared change

☐ Descriptives

☐ Part and partial correlations

☐ Collinearity diagnostics

Residuals

☐ Durbin-Watson

☐ Casewise diagnostics

● Outliers outside: 3 standard deviations

○ All cases

Other frequently used options include:

- **Confidence intervals:** Produces 95% confidence intervals for the *B* values.
- **Covariance matrix:** Produces a covariance-variance-correlation matrix with covariances below the diagonal, variances on the diagonal, and correlations above the diagonal. This is just one tool used in the regression procedure for identifying collinearity among variables.
- **R squared change:** In the **Forward** and **Stepwise** procedures documents the change in the $R^2$ value as each new variable is entered into the regression equation.
- **Descriptives:** The variable means, standard deviations, and a correlation matrix.
- **Collinearity diagnostics:** Assists in exploring whether collinearity exists among predictor and criterion variables.

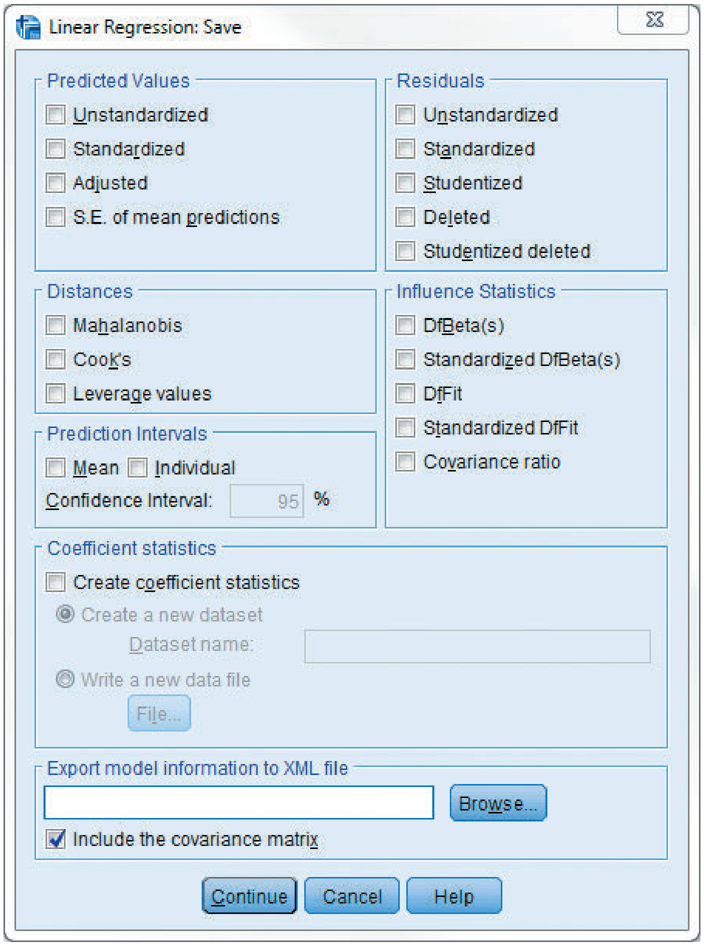
A click on the **Save** button opens a dialog window (Screen 16.3, following page) with many terms unfamiliar even to the mathematically sophisticated. In general, selection of an item within this dialog box will save in the data file one or more new variable-values determined by a number of different procedures. Contents of the five boxes are briefly described below.

- **Residuals:** Some of the five options listed here are covered in Chapter 28.
- **Influence Statistics:** This box deals with the very useful function of what happens to a distribution or an analysis with the deletion of a particular subject or case. It may be necessary to delete even valid data from time to time. I once had a student in a class at UCLA who could run a 4:00 mile (Jim Robbins). If I were measuring physical fitness based on time to run a mile, the class average would probably be about 8:00 minutes. Despite the fact that Jim's result is valid, it is certainly abnormal when compared to the general population. Thus deletion of this value may have a significant influence on the results of an analysis. Unfortunately a more thorough discussion of these procedures extends well beyond the scope of this book.

- **Prediction Intervals:** These are 95% (or any desired percentage) confidence intervals for either the mean of a variable or of an individual case.
- **Distances:** Three different ways of measuring distances between cases, a concept introduced in the chapter on Cluster Analysis (Chapter 21).
- **Predicted Values:** This is the only portion of this entire window that is illustrated in one of the step-by-step sequences that follow. The Unstandardized predicted values are often useful to compare with actual values when considering the validity of the regression equation. Also it is one of the most sophisticated ways to replace missing values (see Chapter 4, page 63. The Standardized predicted values are the predicted values for *z* scores of the selected variables.

Finally we consider the dialog box that appears when you click the **Options** button (Screen 16.4, below). The **Include constant in equation** is selected by default and should remain unless there is some specific reason for deselecting it. The **Stepping Method Criteria**, depending on your setting, may be used fairly frequently. This allows you to select the *p* value for a variable to enter the regression equation (the default is .05) for the **Forward** and **Stepwise** methods, or a *p* value to remove an already entered variable (the default is .10) for the **Stepwise** and **Backward** methods. If you wish you may instead choose *F* values as criteria for entry or removal from the regression equation. The issue of **Missing Values** was discussed in Chapter 4 and will not be considered here.

Screen 16.3 The Linear Regression: Save Window



Linear Regression: Save

Predicted Values
□ Unstandardized
□ Standardized
□ Adjusted
□ S.E. of mean predictions

Residuals
□ Unstandardized
□ Standardized
□ Studentized
□ Deleted
□ Studentized deleted

Distances
□ Mahalanobis
□ Cook's
□ Leverage values

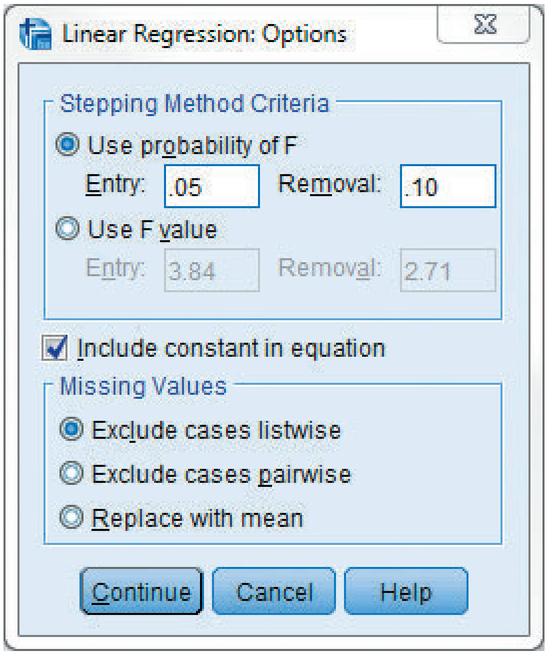
Influence Statistics
□ DfBeta(s)
□ Standardized DfBeta(s)
□ DfFit
□ Standardized DfFit
□ Covariance ratio

Prediction Intervals
□ Mean □ Individual
Confidence Interval: 95 %

Coefficient statistics
□ Create coefficient statistics
● Create a new dataset
Dataset name:
○ Write a new data file
File...

Export model information to XML file
Browse...
☑ Include the covariance matrix

Screen 16.4 The Linear Regression: Options Window



Linear Regression: Options

Stepping Method Criteria
● Use probability of F
Entry: .05 Removal: .10
○ Use F value
Entry: 3.84 Removal: 2.71

☑ Include constant in equation

Missing Values
● Exclude cases listwise
○ Exclude cases pairwise
○ Replace with mean

Below we produce two step-by-step sequences. The first (Step 5) is the simplest possible default regression analysis with **zhelp** as the dependent variable and

**sympathy**, **severity**, **empatend**, **efficacy**, and **anger** as independent (or predictor) variables. We select the **Forward** method of entering variables. The second (Step 5a) is an analysis that includes a number of the options discussed in the previous pages. The italicized introduction to each box will specify what functions are being utilized in each procedure.

*To run the regression procedure with a dependent variable of* **zhelp** *and predictors of* **sympathy**, **severity**, **empatend**, **efficacy**, *and* **anger**, *using the* **forward** *method of selection, perform the following sequence of steps. The starting point is Screen 16.1. Perform the four-step sequence (page* 208*) to arrive at this screen.*

| In Screen | Do This | Step 5 |
|---|---|---|
| 16.1 | ✱ **zhelp** → ✱ *top* ➡ → ✱ **sympathy** → ✱ *second* ➡ → | |
| | ✱ **severity** → ✱ *second* ➡ → ✱ **empatend** → ✱ *second* ➡ → | |
| | ✱ **efficacy** → ✱ *second* ➡ → ✱ **anger** → ✱ *second* ➡ → | |
| | ✱ ▼ *to the right of* **Method** → ✱ **Forward** → ✱ **OK** | Back1 |

The result of this sequence is to produce a regression analysis that identifies which of the predictors (the helpers' sympathy, empathic tendency, self-efficacy, anger, and rating of problem severity) have the greatest influence on the dependent variable, the amount of time spent helping. The forward method of selection will first enter the independent variable having the highest bivariate correlation with help, then enter a second variable that explains the greatest additional amount of variance in time helping, then enter a third variable and so forth until there are no other variables that significantly influence the amount of help given.

*To run the regression procedure with a dependent variable of* **zhelp** *and predictors of* **sympathy**, **severity**, **empatend**, **efficacy**, *and* **anger**, *using the* **Stepwise** *method of selection, including statistics of* **Estimates**, **Descriptives**, *and* **Model fit**, *selecting the* **Unstandardized predicted values** *as an additional saved variable, and establishing an* **Entry value** *of .10 and a* **Removal value** *of .20, perform the following sequence of steps. The starting point is Screen 16.1. (Refer to sequence Step 4 on page* 208.*)*

| In Screen | Do This | Step 5a |
|---|---|---|
| 16.1 | ✱ **zhelp** → ✱ *top* ➡ → ✱ **sympathy** → ✱ *second* ➡ → | |
| | ✱ **severity** → ✱ *second* ➡ → ✱ **empatend** → ✱ *second* ➡ → | |
| | ✱ **efficacy** → ✱ *second* ➡ → ✱ **anger**→ ✱ *second* ➡ → | |
| | ✱ ▼ *to the right of* **Method** → ✱ **Stepwise** → ✱ **Statistics** | |
| 16.2 | ✱ **Descriptives** → ✱ **R squared change** → ✱ **Continue** | |
| 16.1 | ✱ **Save** | |
| 16.3 | ✱ **Unstandardized** (*in the* **Predicted Values** *box*) → ✱ **Continue** | |
| 16.1 | ✱ **Options** | |
| 16.4 | press **TAB** → type `.10` → press **TAB** → type `.20` → ✱ **Continue** | |
| 16.1 | ✱ **OK** | Back1 |

The result of this sequence is to produce a regression analysis that identifies which of the predictors (the helpers' sympathy, empathic tendency, self-efficacy, anger, and rating of problem severity) have the greatest influence on the dependent variable (amount of time spent helping). The stepwise method of selection will first enter the independent variable with the highest bivariate correlation with help, then enter the variable that explains the greatest additional amount of variance, then enter a third variable and so forth until no other variables significantly (significance is specified as $p \leq .10$ for this analysis) influence the amount of help given. If the influence of any variable increases above a significance of .20 after entry into the regression analysis, it will be dropped from the regression equation. We request the saving of a new variable that will include the predicted amount of time helping for each subject. We also request the addition of descriptive statistics and the correlation matrix of all variables.

Upon completion of Step 5, or 5a, the output screen will appear (Screen 1, inside back cover). All results from the just-completed analysis are included in the Output Navigator. Make use of the scroll bar arrows (▲▼▶◀) to view the results. Even when viewing output, the standard menu of commands is still listed across the top of the window. Further analyses may be conducted without returning to the data screen.

## 16.5 Printing Results

Results of the analysis (or analyses) that have just been conducted require a window that displays the standard commands (**File Edit Data Transform Analyze . . .**) across the top. A typical print procedure is shown below beginning with the standard output screen (Screen 1, inside back cover).

*To print results, from the Output screen perform the following sequence of steps:*

| In Screen | Do This | Step 6 |
|---|---|---|
| **Back1** | *Select desired output or edit (see pages 16–22)* → ✳ **File** → ✳ **Print** | |
| **2.9** | *Consider and select desired print options offered, then* → ✳ **OK** | |

*To exit you may begin from any screen that shows the* **File** *command at the top.*

| In Screen | Do This | Step 7 |
|---|---|---|
| **Menu** | ✳ **File** → ✳ **Exit** | **2.1** |

Note: After clicking **Exit**, there will frequently be small windows that appear asking if you wish to save or change anything. Simply click each appropriate response.

## 16.6 Output

### 16.6.1 Multiple Regression Analysis

The following output is produced by sequence Step 5 on page 212. The format of the output below is fairly different (and much neater) than actual SPSS output, but all relevant data are included with identical terminology. As procedures become more complex (in chapters that follow) there is an increasing necessity to condense format to only the most relevant output. Hopefully that way when you get a book-length output from SPSS, you can see where the most essential parts of the output are.

Model Summary

| | |
|---|---|
| R | .616 |
| R square | .380 |
| Adjusted R square | .355 |
| Std Error of Estimate | 1.0058 |

**Model 3** (*Output also displays Models 1 and 2*)
Method: Forward (Criterion: Probability of F to enter <= .050)

Dependent Variable: ZHELP
Predictors: (Constant, SYMPATHY, ANGER, EFFICACY)

ANOVA

| Source of Variation | df | Sum of Squares | Mean Square | F | Signif F |
|---|---|---|---|---|---|
| Regression | 3 | 47.654 | 15.885 | 15.701 | .000 |
| Residual | 77 | 77.899 | 1.012 | | |

Coefficients

| | Unstandardized coefficients | | Std Coefficients | | |
|---|---|---|---|---|---|
| | B | Std. Error | Beta | t | Signif of t |
| (Constant) | −4.308 | .732 | | −5.885 | .000 |
| SYMPATHY | .494 | .100 | .451 | 4.938 | .000 |
| ANGER | .284 | .083 | .310 | 3.429 | .001 |
| EFFICACY | .412 | .132 | .284 | 3.134 | .002 |

Excluded Variables

| | | | | | | Collinearity Stats |
|---|---|---|---|---|---|---|
| Model | Variable | Beta in | t | Signif of t | Partial Correlation | Tolerance |
| 3 | SEVERITY | .077 | .787 | .443 | .090 | .844 |
| | EMPATEND | .119 | 1.269 | .208 | .144 | .909 |

An initial look identifies key elements of the analysis: Three models were tested (only the third is shown here), with the three variables that met the entry requirement included in the final equation (**sympathy**, **anger**, and **efficacy**). Two variables did not meet the entry requirement (**severity** and **empatend**). The multiple *R* shows a substantial correlation between the three predictor variables and the dependent variable **zhelp** (*R* = .616). The *R*-square value indicates that about 38% of the variance in **zhelp** is explained by the three predictor variables. The β values indicate the relative influence of the entered variables, that is, sympathy has the greatest influence on help (β = .45), followed by anger (β = .31), and then efficacy (β = .28). The direction of influence for all three is positive.

## 16.7 Change of Values as Each new Variable is Added

What follows is partial output from sequence Step 5a on page 212. Once again this output is a greatly condensed version of the SPSS output to demonstrate the changes in the variables from step to step as new variables are entered into the regression equation.

The purpose of displaying the output from the three steps in this format is to allow you to see how the computed values change as each new variable is added. Note for instance how the multiple *R*, the *R* square, and the adjusted *R* square *increase* in value with the addition of each new variable. Note also how the standard error and *R*-square change *decrease* in value with the addition of new variables. Similar patterns may be noted for degrees of freedom, sum of squares, and mean squares.

| Dependent Variable | ZHELP | Model # | Model 1 | Model 2 | Model 3 |
|---|---|---|---|---|---|
| | Statistics | Variables in Equation | SYMPATHY | SYMPATHY<br>ANGER | SYMPATHY<br>ANGER<br>EFFICACY |
| | Multiple R | | .455 | .548 | .616 |
| | R Square | | .207 | .300 | .380 |
| | Adjusted R square | | .197 | .282 | .355 |
| | Standard Error | | 1.123 | 1.061 | 1.006 |
| | R Square change | | .207 | .093 | .079 |

| | | | | |
|---|---|---|---|---|
| degrees of freedom | regression | 1 | 2 | 3 |
| | residual | 79 | 78 | 77 |
| SUM of SQUARES | regression | 26.008 | 37.715 | 47.654 |
| | residual | 99.544 | 87.837 | 77.899 |
| MEAN SQUARES | regression | 26.008 | 18.858 | 15.884 |
| | residual | 1.260 | 1.126 | 1.012 |
| Overall F | | 20.641 | 16.746 | 15.701 |
| Significance of F | | .0000 | .0000 | .0000 |

Additional comment is included in the definitions of terms that follow.

| Term | Definition/Description |
|---|---|
| **PROB OF F TO ENTER** | Probability value to enter a variable into the regression equation. In this case *p* = .05, the default value. |
| **PROB OF F TO REMOVE** | For **stepwise** or **backward** procedures only. Probability value to remove an already entered variable from the regression equation. |
| **R** | The multiple correlation between the dependent variable **zhelp** and the three variables in the regression equation, **sympathy**, **anger**, and **efficacy**. |
| **R SQUARE** | The $R^2$ effect size value identifies the portion of the variance accounted for by the independent variables; that is, approximately 38% of the variance in zhelp is accounted for by sympathy, anger, and efficacy. |
| **ADJUSTED R SQUARE** | $R^2$ is an accurate value for the sample drawn but is considered an optimistic estimate for the *population* value. The Adjusted $R^2$ is considered a better population estimate and is useful when comparing the $R^2$ values between models with different numbers of independent variables. |
| **STANDARD ERROR** | STANDARD ERROR is the standard deviation of the expected value of (in this case) zhelp. Note that this value shrinks as each new variable is added to the equation (see output from the bottom output table above). |
| **REGRESSION** | Statistics relating to the explained portion of the variance. |
| **RESIDUAL** | Statistics relating to the *un*explained portion of the variance. |
| **DF** | For regression, the number of independent variables in the equation. For residual, *N*, minus the number of independent variables entered, minus 1: 81 – 3 – 1 = 77. |
| **SUM OF SQUARES** | The regression sum of squares corresponds to the between-group sum of squares in ANOVA, and the residual sum of squares corresponds to the within-group sum of squares in ANOVA. Note that in this case there is a larger portion of *un*explained variance than there is of explained variance, a reality also reflected in the R SQUARE value. |
| **MEAN SQUARE** | Sum of squares divided by degrees of freedom. |
| **F** | Mean square regression divided by mean square residual. |
| **SIGNIF F** | Probability that this *F* value could occur by chance. The present results show that the likelihood of the given correlation occurring by chance is less than 1 in 10,000. |

| | |
|---|---|
| **B** | The coefficients and constant for the regression equation that measures predicted values for ZHELP:<br>**zhelp**$_{\text{(predicted)}}$ = –4.308 + .494 (**sympathy**) + .284 (**anger**) + .412 (**efficacy**) |
| **STD ERROR** | Standard error of *B*: A measure of the stability or sampling error of the *B* values. It is the standard deviation of the *B* values given a large number of samples drawn from the same population. |
| **BETA (β)** | The standardized regression coefficients. This is the *B* value for standardized scores (*z* scores) of the independent variables. This value will always vary between ±1.0 in linear relationships. For curvilinear relationships it will sometimes extend outside that range. |
| **t** | *B* divided by the standard error of *B*. This holds both for variables in the equation and those variables *not* in the equation. |
| **SIGNIF t** | Likelihood that these *t* values could occur by chance. |
| **BETA IN** | The beta values for the *excluded* variables if these variables were actually in the regression equation. |
| **PARTIAL CORRELATION** | The partial correlation coefficient with the dependent variable **zhelp**, adjusting for variables already in the regression equation. For example, the simple correlation (*r*) between **zhelp** and **severity** is .292. After **sympathy**, **anger**, and **efficacy** have "consumed" much of the variance, the *partial* correlation between **zhelp** and **severity** is only .089. **empatend**, on the other hand, shrinks only a little, from .157 to .144. |
| **MINIMUM TOLERANCE** | Tolerance is a commonly used measure of collinearity. It is defined as 1 – $R_i^2$, where $R_{\text{i}}$ is the *R* value of variable *i* when variable *i* is predicted from all other independent variables. A low tolerance value (near 0) indicates extreme collinearity, that is, the given variable is almost a linear combination of the other independent variables. A high value (near 1) indicates that the variable is relatively independent of other variables. This measure deals with the issue of linear dependence that was discussed in the Introduction. When variables are included that are linearly dependent, they inflate the standard errors, thus weakening the power of the analysis. It also conceals other problems, such as curvilinear trends in the data. |
| **R SQUARE CHANGE** | This is simple subtraction of the $R^2$ value for the given line minus the $R^2$ value from the previous line. Note that the value for **anger** .093 = .300 – .207 (within rounding error). The number (.09324) indicates that the inclusion of the second variable (**anger**) explains an additional 9.3% of the variance. |

# Exercises

Answers to selected exercises are downloadable at **www.spss-step-by-step.net.**

Use the **helping3.sav** file for the exercises that follow (downloadable at the address shown above).

Conduct the following THREE regression analyses: Criterion variables:

1. **thelplnz:** time spent helping
2. **tqualitz:** quality of the help given
3. **tothelp:** a composite help measure that includes both time and quality

Predictors (use the same predictors for each of the three dependent variables):

**age:** ranges from 17 to 89

**angert:** amount of anger felt by the helper toward the needy friend

**effict:** helper's feeling of self-efficacy (competence) in relation to the friend's problem

**empathyt:** helper's empathic tendency as rated by a personality test

**gender:** 1 = female, 2 = male

**hclose:** helper's rating of how close the relationship was

**hcontrot:** helper's rating of how controllable the cause of the problem was

**hcopet:** helper's rating of how well the friend was coping with his or her problem

**hseveret:** helper's rating of the severity of the problem

**obligat:** the feeling of obligation the helper felt toward the friend in need

**school:** coded from 1 to 7 with 1 being the lowest education and 7 the highest (>19 years)

**sympathi:** The extent to which the helper felt sympathy toward the friend

**worry:** amount the helper worried about the friend in need

- Use **entry value** of **.06** and **removal value** of **.11**.
- Use **stepwise** method of entry.

Create a table (example below) showing for each of the three analyses Multiple *R*, $R^2$, then each of the variables that significantly influence the dependent variables. Following the $R^2$, list the name of each variable and then (in parentheses) list its, β value. Rank order them from the most influential to least influential from left to right. Include only significant predictors.

| Dependent Variable | Multiple R | $R^2$ | 1st var (β) | 2nd var (β) | 3rd var (β) | 4th var (β) | 5th var (β) | 6th var (β) |
|---|---|---|---|---|---|---|---|---|
| Time helping | | | | | | | | |
| Help quality | | | | | | | | |
| Total help | | | | | | | | |

4. A researcher is examining the relationship between **stress** levels, **self-esteem**, **coping skills**, and **performance** on a test of cognitive performance (the dependent measure). His data are shown below. Perform multiple regression on these data, entering variables using the stepwise procedure. Interpret the results.

| Stress | Self-esteem | Coping skills | Performance |
|---|---|---|---|
| 6 | 10 | 19 | 21 |
| 5 | 10 | 14 | 21 |
| 5 | 8 | 14 | 22 |
| 3 | 7 | 13 | 15 |
| 7 | 14 | 16 | 22 |
| 4 | 9 | 11 | 17 |
| 6 | 9 | 15 | 28 |
| 5 | 9 | 10 | 19 |
| 5 | 11 | 20 | 16 |
| 5 | 10 | 17 | 18 |

# Chapter 17
# Nonparametric Procedures

THIS CHAPTER deals with (as the title suggests) nonparametric tests. Before we become involved in *non* parametric tests, it might be well to consider first, what is a *parametric* test? A parametric test is one that is based on a parameter, also known as "some descriptive statistic about a population." The most commonly used (and useful) parameter is a mean—drawn from a population that we assume is normal. The standard deviation, close friends with the mean, is another commonly used parameter. Although some operations are based on other assumptions or parameters (e.g., binomial or Poisson distributions of data), the **Nonparametric Tests** procedure deals primarily with populations that are neither normally distributed nor based on continuous data (so means don't mean anything) and considers how to conduct statistical tests if the assumption of normality is violated. For example, if you have 99 people that are listed in rank order, the mean rank will be 50. But that doesn't tell you anything, except that half-way through the list of ranks is the number 50. So, if you want to do statistics on ranks, you need to use nonparametric procedures.

One word of warning: The procedures in this chapter are, in general, not as powerful as other procedures (in other words, it is harder to find significance using these procedures than with procedures that assume normality). So people really don't like using these procedures if they can avoid it, and researchers will go to great lengths to justify using their data that's not technically normal in another, more powerful procedure. Sometimes they will even transform their data (by doing some math on the data to make it more normal, as in "I wonder if I take the natural logarithm of my data if it will look normal enough to avoid this chapter?"). Whether or not you can get away with this depends on how much of a statistical purist you (or your teacher or advisor or editor) are, just how far away from normal your data actually are, and what test you are planning on using (for example, the ANOVA is considered "robust" because it doesn't break too easily when it's assumptions are violated). Because this decision is not something that can be made in a step-by-step format, we will assume that you are here because (a) you have decided that these procedures are best or (b) you want to learn how to use these procedures in case you need them later.

There are already two nonparametric procedures that have been covered in this book. The most popular nonparametric procedure by far is the chi-square test (Chapter 8): If your data consists of frequencies in categories, and you can calculate expected frequencies for the categories, then you should probably use the chi-square procedure. Also, if you are calculating correlations using continuous values that are not normally distributed, Spearman's rho (Chapter 10) is probably the best test for you.

To demonstrate nonparametric tests, we will make use, once more, of the original data file first presented in Chapters 1 and 2: the **grades.sav** file. This is an excellent file to demonstrate nonparametric tests because the contents of the file are so easily understood. Tests will be conducted on the following variables within the file: **gender**, the five 10-point quizzes (**quiz1** to **quiz5**), **ethnic** (the ethnic makeup of the sample), and **section** (membership of subjects in the three sections of the class). The *N* for this sample is 105.

One of the most important steps when doing nonparametric tests is specifying exactly what your research question is, and what kind of nonparametric test you want to calculate. The problem is that, with the SPSS nonparametrics procedure, you can easily calculate 18 tests with 3 clicks. But testing 18 things when you really only want to test 1 is a very bad idea, as it's too tempting to just look for things that are significant rather than what you really care about. There are five basic kinds of questions that you can ask:

1. Are my observed values distributed differently than my hypothesized distribution? (For example, if I hypothesize that half of the students will be men and half will be women, does my observed distribution of men and women differ from that hypothesized distribution?)
2. Is the order of my observed values different than what they would be if they were merely random? (For example, is the order of males and females in the class random, or are males and females segregated somehow?)
3. Are my observed distributions of a continuous variable different depending on what group they are in? (For example, is the distribution of **quiz3** different between the three **sections** of the class?) This is the logical equivalent of a *t* test or one-way ANOVA, but uses ranks so it works with data that aren't normally distributed.
4. Are the medians of a continuous variable different for different groups? (For example, the median quiz score different for different class sections?) This is also the logical equivalent of a *t* test or one-way ANOVA, but instead of using ranks (as in question 3), it only looks at the median instead of the entire distribution.
5. Finally, if you are using repeated measures (also known as dependent measures or within-subjects variables), you can ask: Are my observed distributions of several variables different than each other? (For example, are the different distributions of the five quizzes in a class different than each other?)

This will be one of only two analysis chapters (Chapters 6–27) that deviates substantially from the traditional format. In this chapter, we begin the Step by Step section with the (by now) familiar first three steps. After the first three steps are presented, the Step by Step and Output sections are repeated five times, once for each of the different kinds of questions that you can test using the nonparametric procedures. Each section will have its own introduction, followed by Step 4, then the output, an interpretation of the output, and definitions of terms. If you are sure which of the five kinds of questions that you want to ask, you can go directly to that section of the chapter; if not, you will probably want to read the introduction to each section of the chapter to help you figure out which procedure you want to use.

We have covered here the most commonly used nonparametric tests. SPSS can conduct several other procedures, but they are not as frequently used (or as easy to use). For additional information, the most comprehensive book on nonparametric statistics discovered by the authors is a text by Siegel and Castellan (1988). Please see the reference section for additional detail.

# 17.1 Step by Step

## 17.1.1 Nonparametric Tests

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | → All Programs → IBM SPSS Statistics → IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | ✳ 🚀 → ✳ ⬛ → ✳ Σ | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| | Step 2 |
|---|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* | |

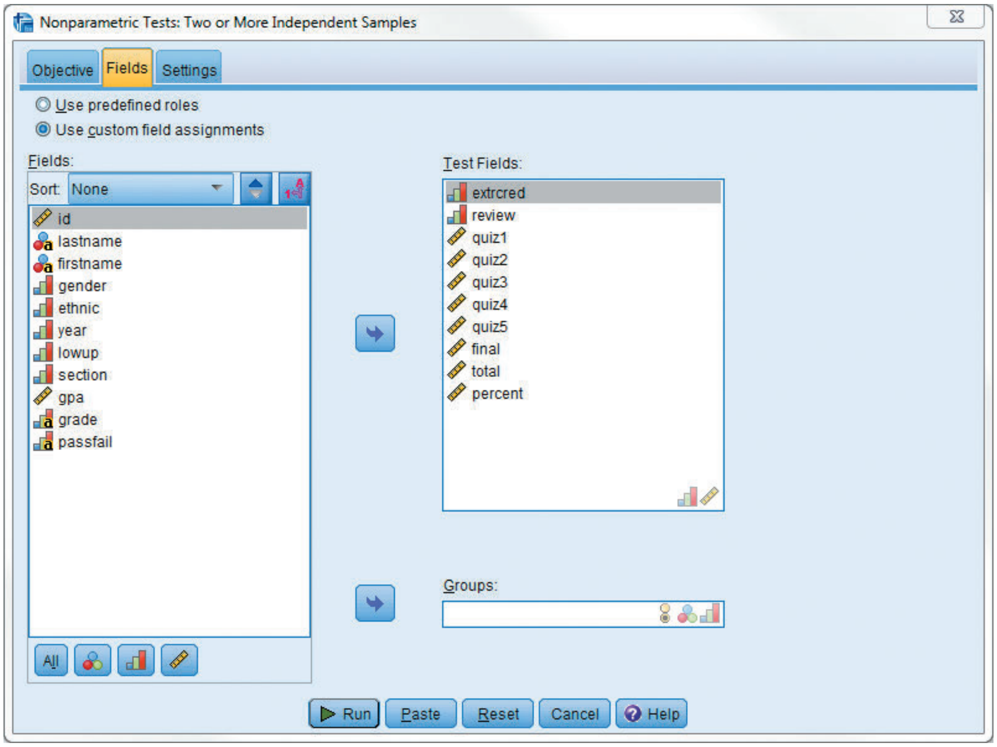
*Screens 1 and 2 (displayed on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **grades.sav** *file for further analyses:*

| In Screen | Do This | Step 3 |
|---|---|---|
| Front1 | ✳ **File** → **Open** → ✳ **Data** [*or* ✳ 📂] | |
| Front2 | type `grades.sav` → ✳ **Open** [*or* ✳ ✳ **grades.sav**] | Data |

Whether first entering SPSS or returning from earlier operations the standard menu of commands across the top is required. As long as it is visible you may perform any analyses. It is not necessary for the data window to be visible.

## 17.1.2 Screen Used in All of the Nonparametric Procedures

### Screen 17.1



Nonparametric Tests: Two or More Independent Samples

○ Use predefined roles
● Use custom field assignments

Fields:

Sort: None

id, lastname, firstname, gender, ethnic, year, lowup, section, gpa, grade, passfail

Test Fields:

extrcred, review, quiz1, quiz2, quiz3, quiz4, quiz5, final, total, percent

Groups:

In all of the nonparametric procedures, you need to specify which variable you want to analyze. Unfortunately, SPSS defaults to analyzing all of the variables in your data file that can be analyzed; this is rarely a good idea. And even if you really, really want to examine a lot of variables at once, it is still probably a good idea to do them one at a time as occasionally SPSS will do something strange if you are using too many variables at once. So you will need to regularly select "Use custom field assignments" and select which fields you want to test on Screen 17.1. (Depending on the test you are conducting, the "Groups" field may or may not be included.)

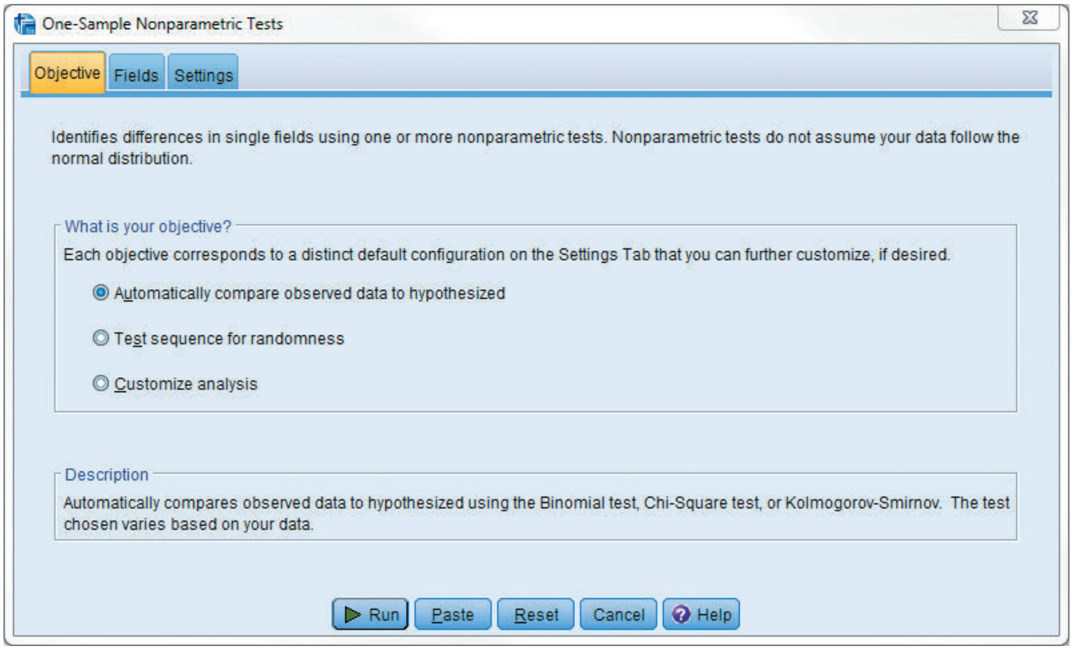
## 17.2 Are Observed Values Distributed Differently than a Hypothesized Distribution?

With this procedure, you are testing one variable (that could be categorical or continuous), and seeing if its distribution is significantly different than whatever distribution you expect it to have. For example, if you expect 50% of the people in your classes to be men and 50% to be women, is your actual distribution of 41 men and 64 women significantly different than what you had expected? Or, if you expected students to be evenly distributed between three sections of a class that you are teaching, are they? Or, are the final scores for a class normally distributed? Note that SPSS defaults to a uniform distribution if the variable you are testing is categorical (all categories are equal), and to a normal distribution if the variable you are testing is continuous. But you can specify different predictions; if you expect that there are twice as many women as men in a class, you can test that as well. In all cases, you are comparing a distribution that you observed with a distribution that you expected, to see if they are significantly different.

If you've been paying close attention, this might sound familiar . . . isn't the chi-square procedure for comparing observed and expected frequencies? Indeed, if you are examining several different categories of a variable, then in fact this *is* the procedure to calculate a one-way chi-square test.

In this example, we will examine the **ethnic** variable to see if ethnicities are evenly distributed in the **grades.sav** file.

Screen 17.2



One-Sample Nonparametric Tests

Identifies differences in single fields using one or more nonparametric tests. Nonparametric tests do not assume your data follow the normal distribution.

What is your objective?

Each objective corresponds to a distinct default configuration on the Settings Tab that you can further customize, if desired.

● Automatically compare observed data to hypothesized

○ Test sequence for randomness

○ Customize analysis

Description

Automatically compares observed data to hypothesized using the Binomial test, Chi-Square test, or Kolmogorov-Smirnov. The test chosen varies based on your data.

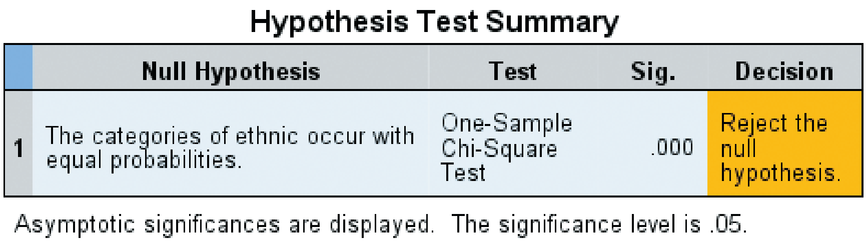
*Perform the following steps to determine if* **ethnic** *is equally distributed.*

| In Screen | Do This | Step 5 |
|---|---|---|
| Menu | ✳ **Analyze** → **Nonparametric Tests** → ✳ **One-Sample** | |
| 17.2 | ✳ **Fields** | |
| 17.1 | ✳ **Use custom field assignments** → ✳ ✳ **ethnic** → ✳ **Run** **Back1** | |

If you want to test if a distribution is different than something other than a uniform distribution (each category has the same frequency, in the case of categorical variables) or is different than an exponential, uniform, or Poisson distribution (in the case of continuous variables), then you should first do the simple analysis described in this step-by-step procedure, and then (once you have seen which test SPSS recommends for your variable) do it again. But before clicking on **Run**, click on **Settings**, **Customize tests**, select whichever test SPSS did when you first ran the command, then click **Options**, followed by **Customize expected probability**. You can then tell SPSS what distribution you expected to find.

### 17.2.1 Output

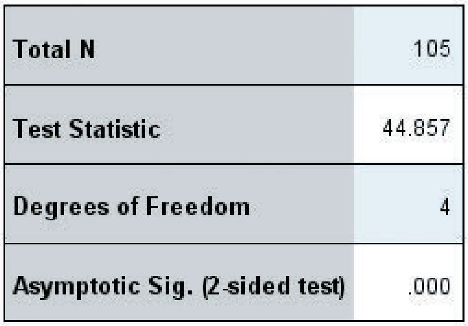
Important trick! In the Output of the nonparametrics procedure, SPSS only displays part of the output. Be sure to double-click on the output from the procedure so that you can see all of the output.



**Hypothesis Test Summary**

| | Null Hypothesis | Test | Sig. | Decision |
|---|---|---|---|---|
| 1 | The categories of ethnic occur with equal probabilities. | One-Sample Chi-Square Test | .000 | Reject the null hypothesis. |

Asymptotic significances are displayed. The significance level is .05.



| | |
|---|---|
| Total N | 105 |
| Test Statistic | 44.857 |
| Degrees of Freedom | 4 |
| Asymptotic Sig. (2-sided test) | .000 |

Because SPSS is choosing the test automatically, it helpfully tells you what the null hypothesis it is testing. In this case, it is seeing if it can confidently reject the null hypothesis that the categories of ethnicity occur with equal probabilities. It did a one-sample chi-square test to see if the observed values were different than the expected values, and indeed, with a *p* value of less than .001, we can reject the null hypothesis: The categories of ethnicity are not distributed equally. Depending on what kind of variable you analyzed, SPSS may conduct a binomial test, a one-sample chi-square test, or a Kolmogorov-Smirnov test.

Once you double-click on the Hypothesis Test Summary on the left, the output on the right appears. It always includes *N* (the number of cases analyzed), the test statistic (sometimes with a standardized test statistic as well), and the *p* value for the test to see if your observed distribution was different than your expected distribution (the asymptotic significance, or two-sided test). It may also include degrees of freedom or other values used in calculating the test statistic and *p* value.

| Test Run | What the Test Statistic Is | Also Report | How to Report the Effect Size |
|---|---|---|---|
| **ONE-SAMPLE BINOMIAL TEST** | Not really a test statistic, just how many people were in one of your categories. Just report the frequencies and/or percentages in each category. | | Just report the frequencies and/or percentages of people in each category. |
| **ONE-SAMPLE CHI-SQUARE TEST** | Chi-squared ($\chi^2$). The formula is described in Chapter 8. | Degrees of freedom (# of categories minus 1). | Calculate Cramer's *V*, by taking the $\chi^2$ value and dividing it by the product *N* times the degrees of freedom, then take the square root of that value:<br>$\text{Cramer's } V = \sqrt{\frac{\chi^2}{N(\text{degrees of freedom})}}$ |
| **ONE-SAMPLE KOLMOGOROV-SMIRNOV TEST** | Kolmogorov-Smirnov Statistic or Kolmogorov-Smirnov *Z* (with Lilliefors correction). | | Most Extreme Differences, Absolute: This identifies the greatest difference between the sample and the hypothesized distribution, in *z* scores. |

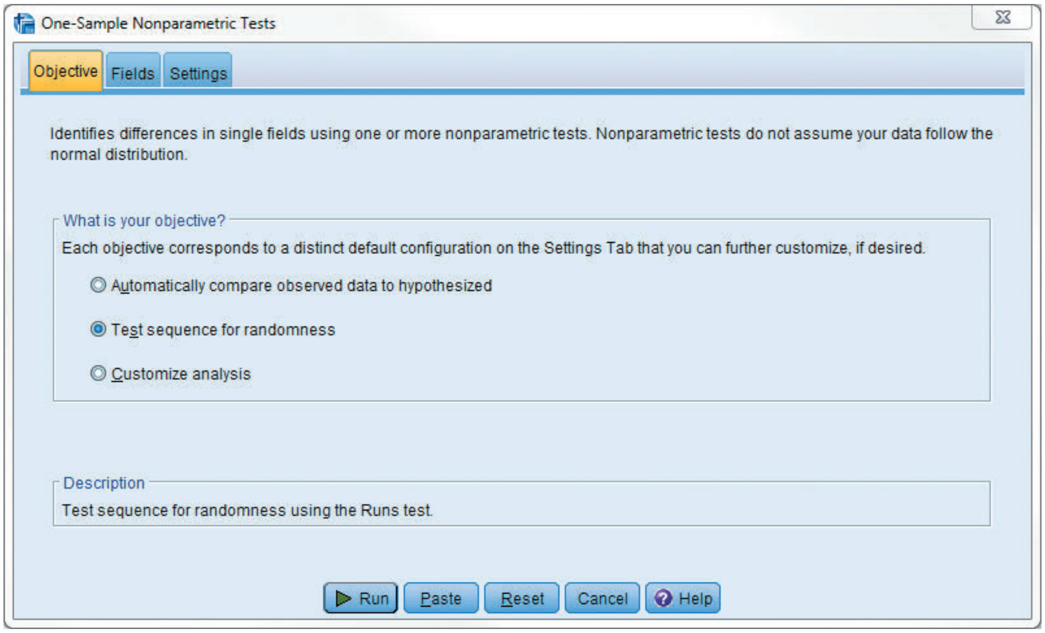
It is rarely appropriate to report a test statistic without saying what that test statistic is. The following table describes what the test statistic really is, along with other things that need to be reported, for each of the possible tests using this procedure.

## 17.3 Is the Order of Observed Values Non-Random?

With this procedure, you are testing whether the sequence of a categorical variable (with two categories) is significantly different than you would get if the order is random. Note that, unlike most SPSS tests, the order of your data is critical, as that's what is being tested.

This test is called the "runs" test, because it examines how many "runs" exist with the same value of your variable occurring several times in a row. If the sequence HHTHTTHTTHTHTTTTHTH resulted from flipping a coin, does this sequence differ significantly from randomness? In other words, are we flipping a biased coin? Unfortunately this procedure works only with dichotomous data (exactly two possible outcomes). It is not possible to test, for instance, if we are rolling a loaded die. Sticking fiercely by our determination to use the **grades.sav** file to demonstrate all procedures in this chapter, we test whether the males and females in our file are distributed randomly throughout our data set.

Screen 17.3



Identifies differences in single fields using one or more nonparametric tests. Nonparametric tests do not assume your data follow the normal distribution.

What is your objective?

Each objective corresponds to a distinct default configuration on the Settings Tab that you can further customize, if desired.

○ Automatically compare observed data to hypothesized

● Test sequence for randomness

○ Customize analysis

Description

Test sequence for randomness using the Runs test.

*Perform the following steps to determine if* **gender** *is randomly distributed in* **grades.sav**.

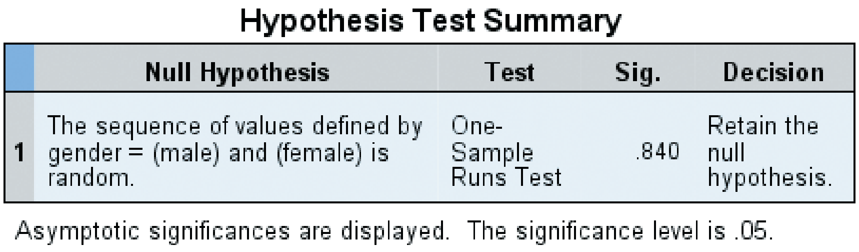
| In Screen | Do This | Step 5 |
|---|---|---|
| Menu | ✳ **Analyze** → **Nonparametric Tests** → ✳ **One-Sample** | |
| 17.2 | ✳ **Test sequence for randomness** → ✳ **Fields** | |
| 17.1 | ✳ **Use custom field assignments** → ✳ ✳ **gender** → ✳ **Run** **Back1** | |

### 17.3.1 Output

Important trick! In the Output of the nonparametrics procedure, SPSS only displays part of the output. Be sure to double-click on the output from the procedure so that you can see all of the output.

Once you double-click on the first output (left), additional output appears (right). It always includes $N$ (the number of cases analyzed), the number of runs (called the test statistic), the standard error of the test statistic, the standardized test statistic (the number of runs converted to a $z$ value), and the $p$ value for the test to see if your observed sequence of men and women differs from random.

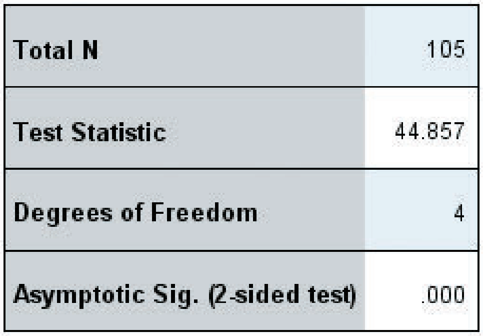
In this case, SPSS tested if the sequence of values defined by gender (male or female) is different than random; because the significance is high (.84), we cannot be confident that the distribution of males and females in the data files is different than random.



**Hypothesis Test Summary**

| | Null Hypothesis | Test | Sig. | Decision |
|---|---|---|---|---|
| 1 | The sequence of values defined by gender = (male) and (female) is random. | One-Sample Runs Test | .840 | Retain the null hypothesis. |

Asymptotic significances are displayed. The significance level is .05.



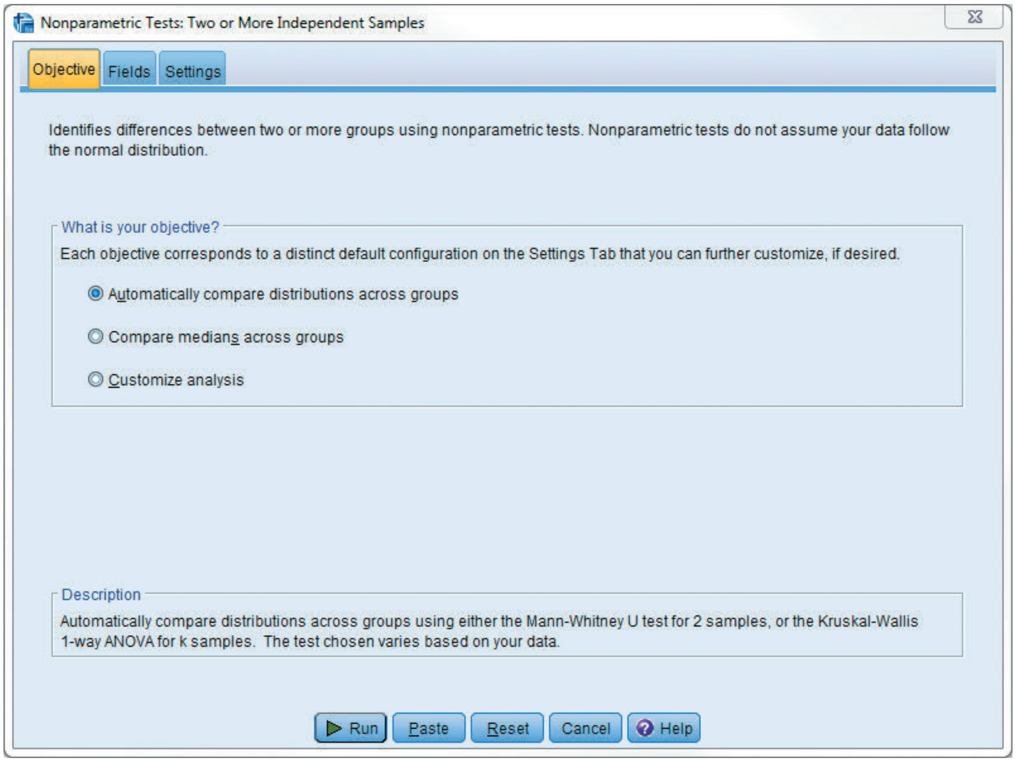
| | |
|---|---|
| Total N | 105 |
| Test Statistic | 44.857 |
| Degrees of Freedom | 4 |
| Asymptotic Sig. (2-sided test) | .000 |

| Test Run | What the Test Statistic Is | How to Report the Effect Size |
|---|---|---|
| **ONE-SAMPLE RUNS TEST** | The number of runs. Reporting the runs test *z* is usually more useful. | The *z* for the runs test gives a measure of how far away from random the sequence of values is. |

## 17.4 Is a Continuous Variable Different in Different Groups?

If you wanted to perform a *t* test or a one-way ANOVA but you can't because your data aren't normally distributed, this is the procedure for you. It examines the distribution of a level of a continuous variable that isn't normally distributed (e.g., the scores on **quiz3**) to see how the distribution is different for different levels of a second categorical variable (e.g., class **section**); it does this by ranking the values, and analyzing the ranks rather than the raw scores. If the distributions do *not* differ significantly from normal, the *t* test or one-way ANOVA should be used because they have greater power.

Screen 17.4



Nonparametric Tests: Two or More Independent Samples

Identifies differences between two or more groups using nonparametric tests. Nonparametric tests do not assume your data follow the normal distribution.

What is your objective?

Each objective corresponds to a distinct default configuration on the Settings Tab that you can further customize, if desired.

● Automatically compare distributions across groups

○ Compare medians across groups

○ Customize analysis

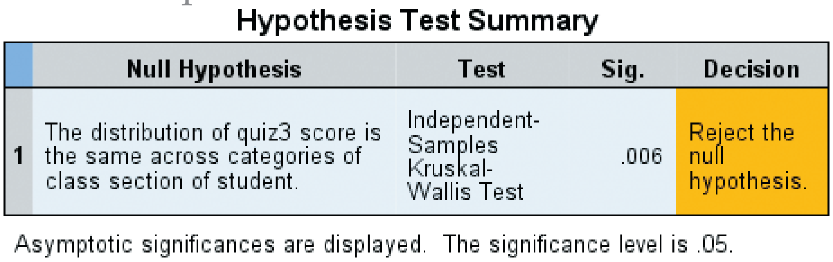
Description

Automatically compare distributions across groups using either the Mann-Whitney U test for 2 samples, or the Kruskal-Wallis 1-way ANOVA for k samples. The test chosen varies based on your data.

*Perform the following steps to determine if the distribution for* **quiz3** *differs in different* **sections**.

| In Screen | Do This | Step 5 |
|---|---|---|
| Menu | ✳ Analyze → Nonparametric Tests → ✳ Independent Samples | |
| 17.4 | ✳ Fields | |
| 17.1 | ✳ Use custom field assignments → ✳ quiz3 → ✳ *top* ➡ → | |
| | ✳ section → ✳ *bottom* ➡ → ✳ Run | Back1 |

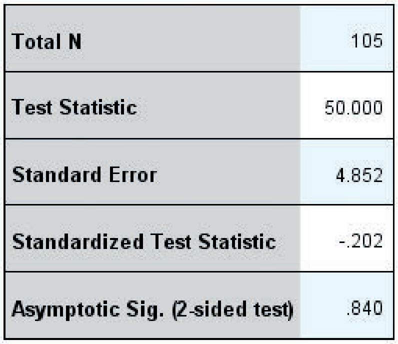
## 17.4.1 Output



**Hypothesis Test Summary**

| | Null Hypothesis | Test | Sig. | Decision |
|---|---|---|---|---|
| 1 | The distribution of quiz3 score is the same across categories of class section of student. | Independent-Samples Kruskal-Wallis Test | .006 | Reject the null hypothesis. |

Asymptotic significances are displayed. The significance level is .05.



| | |
|---|---|
| Total N | 105 |
| Test Statistic | 50.000 |
| Standard Error | 4.852 |
| Standardized Test Statistic | -.202 |
| Asymptotic Sig. (2-sided test) | .840 |

Note: Be sure to double-click on the output to the left so you can see the output on the right.

Because SPSS is choosing the test automatically, it helpfully tells you what test has been done on the data. In this case, SPSS has selected the Kruskal-Wallis test, because there are three different categories of **section** (so it's like a one-way ANOVA). If we had asked SPSS to see if **quiz3** was different for women and men, then SPSS would have chosen the Mann-Whitney U test instead. In this case, the $p$ value is less than .001 and we can reject the null hypothesis: The quiz scores are different between the three class sections.

Once you double-click on the left output, the output on the right appears. It always includes $N$ (the number of cases analyzed), the test statistic (sometimes with a standardized test statistic as well), and the $p$ value for the test to see if your groups were different (the asymptotic significance, or two-sided test). It may also include degrees of freedom or other values used in calculating the test statistic and $p$ value.

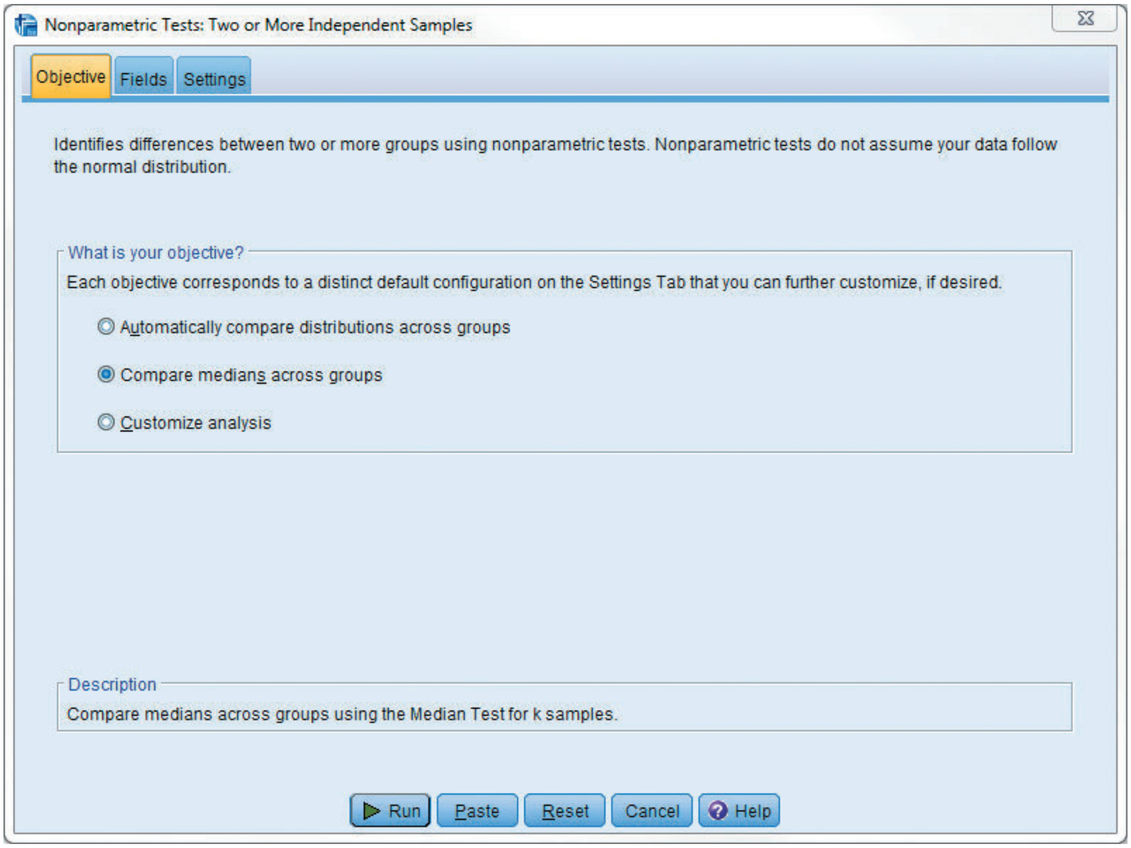
It is rarely appropriate to report a test statistic without saying what that test statistic is. The following table describes what the test statistic really is, along with other things that need to be reported.

| Test Run | Test Statistic | Also Report This | How to Report the Effect Size |
|---|---|---|---|
| **INDEPENDENT SAMPLES KRUSKAL-WALLIS TEST** | Kruskal-Wallis $H$ | Degrees of freedom (# of groups –1) | There is no standard way to calculated effect size in this case. But if you want to examine whether two particular groups in your data are different than each other, you can select cases to examine only those two groups and then run the test again . . . you will then have a Mann-Whitney $U$, and you can calculate the effect size for that comparison between two groups at a time. |
| **INDEPENDENT SAMPLES MANN-WHITNEY U TEST** | Mann-Whitney $U$ | | Using the standardized test statistic from the SPSS output (which is a $Z$), calculate an $r$ effect size: $r = \frac{Z}{\sqrt{N}}$ |

## 17.5 Are the Medians of a Variable Different for Different Groups?

If your data is such that medians are the only thing that you feel you can trust (because your data is really not normal or has a very strange distribution), but you think that the medians are still useful, this test will compare medians across groups and see if the medians are different from each other. It's not the most powerful test, because it is only comparing medians of each group to the grand median of all groups. But if all you can trust is a median, then this test is your friend. This procedure examines the medians of a variable (e.g., the scores on **quiz3**) to see how the medians are different for different levels of a second categorical variable (e.g., class **section**). If the distributions do *not* differ significantly from normal, the *t* test or one-way ANOVA should be used because they have much greater power.

### Screen 17.5



Identifies differences between two or more groups using nonparametric tests. Nonparametric tests do not assume your data follow the normal distribution.

What is your objective?

Each objective corresponds to a distinct default configuration on the Settings Tab that you can further customize, if desired.

○ Automatically compare distributions across groups

● Compare medians across groups

○ Customize analysis

Description

Compare medians across groups using the Median Test for k samples.

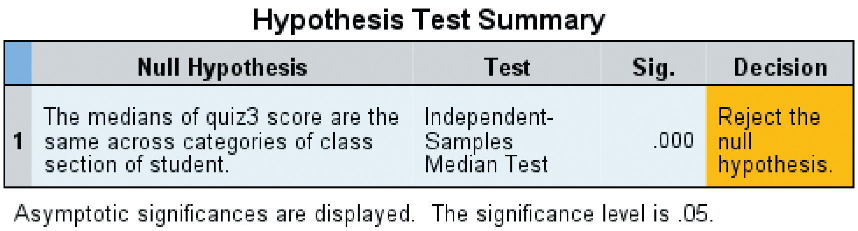
*Perform the following steps to determine if the median scores for* **quiz3** *differ in different* **sections**.

| In Screen | Do This | Step 5 |
|---|---|---|
| Menu | ✱ Analyze → Nonparametric Tests → ✱ Independent Samples | |
| 17.4 | ✱ Compare medians across groups → ✱ Fields | |
| 17.1 | ✱ Use custom field assignments → ✱ quiz3 → ✱ *top* ➔ → | |
| | ✱ section → ✱ *bottom* ➔ → ✱ Run | Back1 |

### 17.5.1 Output

Note: Output is on the following page. In this case, the *p* value is less than .001 and we can reject the null hypothesis: The median quiz scores are different between the three class sections.

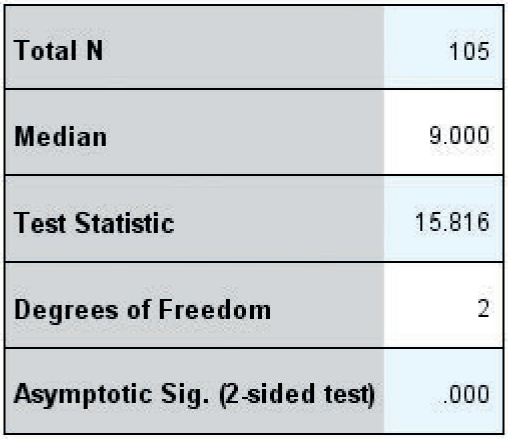
Once you double-click on the output on the left below, the output on the right appears. It includes *N* (the number of cases analyzed), the median (across all groups), the test statistic (which in this case is actually a $\chi^2$ value), the degrees of freedom (number of groups minus 1), and the *p* value for the test to see if the medians for your groups were different (the asymptotic significance, or two-sided test).



Hypothesis Test Summary

| | Null Hypothesis | Test | Sig. | Decision |
|---|---|---|---|---|
| 1 | The medians of quiz3 score are the same across categories of class section of student. | Independent-Samples Median Test | .000 | Reject the null hypothesis. |

Asymptotic significances are displayed. The significance level is .05.



| | |
|---|---|
| Total N | 105 |
| Median | 9.000 |
| Test Statistic | 15.816 |
| Degrees of Freedom | 2 |
| Asymptotic Sig. (2-sided test) | .000 |

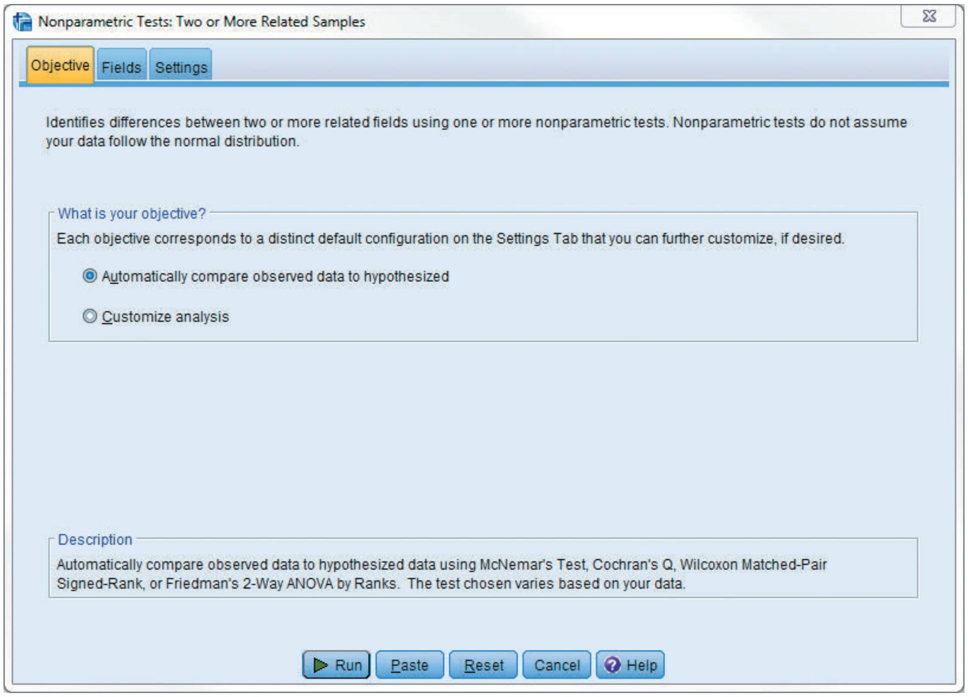
Note: Be sure to double-click on the output to the left so you can see the output on the right.

| Test Run | Test Statistic | Also Report | How to Report the Effect Size |
|---|---|---|---|
| **INDEPENDENT SAMPLES MEDIAN TEST** | $\chi^2$ (chi square) | Degrees of freedom (# of groups –1) | There's not a clean and clear way to do this, but if you want to compare one Independent Samples Median Test to another you can calculate Cramer's *V*, by taking the $\chi^2$ value (test statistic) and dividing it by the product *N* times the degrees of freedom, then take the square root of that value:<br>$\textit{Cramer's } V = \sqrt{\frac{\chi^2}{N(\textit{degrees of freedom})}}$ |

## 17.6 Are My Within-Subjects (Dependent Samples or Repeated Measures) Measurements Different?

With this procedure, you are typically testing one variable that has been measured for one person several times (that could be categorical/ordinal or continuous), and seeing if its distribution is significantly different between the different times that the variable was measured. For example, if people's learning was measured five times with five different quizzes, you might ask "Are the scores different for the quizzes?" This test will also work if the variable you are measuring comes from different people, as long as the people's responses are related to each other (for example, if you are measuring married couples, you might expect them to be similar on many characteristics).

### Screen 17.6



Nonparametric Tests: Two or More Related Samples

Identifies differences between two or more related fields using one or more nonparametric tests. Nonparametric tests do not assume your data follow the normal distribution.

What is your objective?

Each objective corresponds to a distinct default configuration on the Settings Tab that you can further customize, if desired.

● Automatically compare observed data to hypothesized

○ Customize analysis

Description

Automatically compare observed data to hypothesized data using McNemar's Test, Cochran's Q, Wilcoxon Matched-Pair Signed-Rank, or Friedman's 2-Way ANOVA by Ranks. The test chosen varies based on your data.

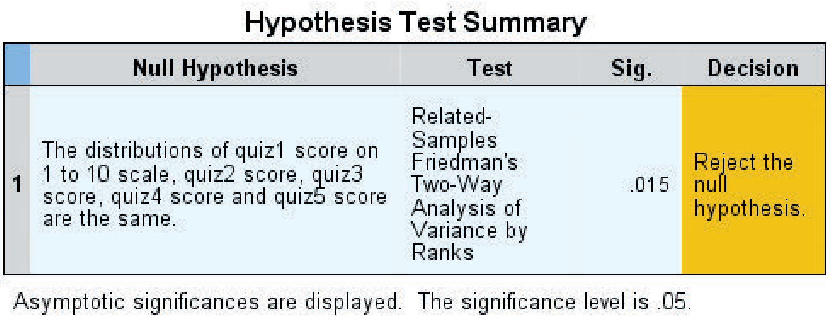
This procedure will work when your dependent variable is continuous (though if it's normal, you'd be better to conduct a dependent-samples *t* test or a one-way ANOVA), or when your dependent variable is ordinal.

In this example, we will examine the five **quiz** variables to see if the distribution of the quiz scores are different in the **grades.sav** file.

*Perform the following steps to determine if the distribution of five quiz scores differ.*

| In Screen | Do This | Step 5 |
|---|---|---|
| Menu | 🖱 **Analyze** → **Nonparametric Tests** → 🖱 **Related Samples** | |
| 17.4 | 🖱 **Fields** | |
| 17.1 | 🖱 **Use custom field assignments** → 🖱 🖱 **quiz1** → 🖱 🖱 **quiz2** | |
| | → 🖱 🖱 **quiz3** → 🖱 🖱 **quiz4** → 🖱 🖱 **quiz5** → 🖱 **Run** | Back1 |

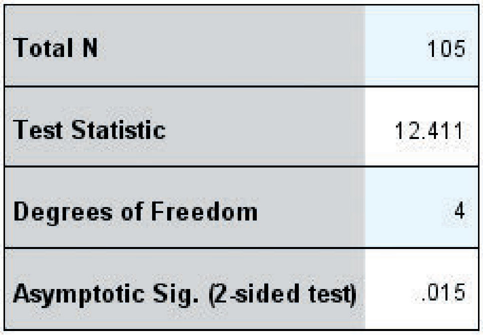
## 17.6.1 Output



**Hypothesis Test Summary**

| | Null Hypothesis | Test | Sig. | Decision |
|---|---|---|---|---|
| 1 | The distributions of quiz1 score on 1 to 10 scale, quiz2 score, quiz3 score, quiz4 score and quiz5 score are the same. | Related-Samples Friedman's Two-Way Analysis of Variance by Ranks | .015 | Reject the null hypothesis. |

Asymptotic significances are displayed. The significance level is .05.



| | |
|---|---|
| Total N | 105 |
| Test Statistic | 12.411 |
| Degrees of Freedom | 4 |
| Asymptotic Sig. (2-sided test) | .015 |

Note: Be sure to double-click on the output to the left so you can see the output on the right.

Because SPSS is choosing the test automatically, it helpfully tells you what the null hypothesis it is testing. In this case, it is seeing if there is a difference between the way the five quizzes are distributed (with a null hypothesis that they are all the same). It conducted a related-samples Friedman's two-way ANOVA by Ranks, and because the significance is less than .05 we can fairly confidently reject the null hypothesis. Depending on what kind of variable you analyzed (categorical and ordinal, or continuous) and how many different groups there are, SPSS may conduct:

- A Wilcoxon signed-rank test, in which differences in a continuous variable between two groups are calculated and then those differences are ranked and analyzed.
- A Friedman two-way ANOVA by ranks, in which ranks of continuous data are analyzed between three or more groups. Because it is called a "two-way ANOVA," one might expect that there are two independent variables. In fact, there is only one independent variable. (Calling it a two-way ANOVA probably made sense when Friedman developed the test in 1937 and the whole idea of ANOVA was young, but it doesn't fit with the way the term is used now, and indeed most people just refer to this as a Friedman ANOVA.)
- McNemar's test, in which the differences in the distribution of a categorical variable with two categories that has been measured twice (e.g., before and after you do something with the participants) to see if they are different. This is similar to a $\chi^2$ (chi-square) test, but it is using repeated measures (within subjects) data.
- Cochran's *Q*, in which the differences in the distribution of a categorical variable with two categories has been measured three or more times (e.g., before, during, and after you do something with the participants) to see if they are different. This is similar to a $\chi^2$ (chi-square) test, but it is using repeated measures (within subjects) data.

Once you double-click on the left output, the output on the right appears. It always includes *N* (the number of cases analyzed), the test statistic, the degrees of freedom (number of groups minus 1), and the *p* value for the test to see if you can safely reject the null hypothesis (i.e., if the variable was different across measurements).

It is rarely appropriate to report a test statistic without saying what that test statistic is. The following table describes what the test statistic really is, along with other things that need to be reported, for each of the possible tests using this procedure.

| Test Run | Test Statistic | Also Report | How to Report the Effect Size |
|---|---|---|---|
| **RELATED-SAMPLES WILCOXON SIGNED-RANK TEST** | Wilcoxon *W* (positive ranks) | | The standardized test statistic is a *Z* value based on the test statistic, and can give an indication of the size of the effect. |
| **RELATED SAMPLES FRIEDMAN TWO-WAY ANOVA BY RANKS** | Friedman $\chi^2$ | Degrees of Freedom (# of groups minus one) | There is no standard way to calculate effect size in this case. But if you want to examine whether two particular measurements in your data are different than each other, you can run the test again selecting only those variables. You will then have a Wilcoxon Signed-Rank test, and you can calculate the effect size for that comparison between two groups at a time. |
| **RELATED SAMPLES MCNEMAR TEST** | McNemar $\chi^2$ | | Although there are measures that can be used to get at effect size (such as odds ratios), the easiest approach is to report your frequencies and percentages. |
| **RELATED SAMPLES COCHRAN'S Q TEST** | Cochran's *Q* | Degrees of Freedom (# of measurements) | We recommend reporting your frequencies and percentages. If you want to get fancy, Serlin, Carr, and Marascuillo's (1982) recommend: $\eta_Q^2 = \frac{Q}{(\text{number pepole})(\text{number of groups})}$ |

## 17.7 Printing Results

Results of the analysis (or analyses) that have just been conducted require a window that displays the standard commands (**File Edit Data Transform Analyze . . .**) across the top. A typical print procedure is shown below beginning with the standard output screen (Screen 1, inside back cover).

*To print results, from the Output screen perform the following sequence of steps:*

| In Screen | Do This | Step 6 |
|---|---|---|
| Back1 | *Select desired output or edit (see pages 16–22)* → ✳ **File** → ✳ **Print** | |
| 2.9 | *Consider and select desired print options offered, then* → ✳ **OK** | |

*To exit you may begin from any screen that shows the* **File** *command at the top.*

| In Screen | Do This | Step 7 |
|---|---|---|
| Menu | ✳ **File** → ✳ **Exit** | 2.1 |

Note: After clicking **Exit**, there will frequently be small windows that appear asking if you wish to save or change anything. Simply click each appropriate response.

# Chapter 18
# Reliability Analysis

MANY CONSTRUCTS are measured in which a subset of relevant items is selected, administered to subjects, and scored—and then inferences are made about the true population values. Examples abound: An introductory psychology course administers a final exam of 100 questions to test students' knowledge; performance on the final is designed to be representative of knowledge of the entire course content. The quantitative section of the GRE is designed to measure general mathematical ability. The 11 subscales of the Wechsler Intelligence Scale are designed to measure general intellectual aptitude—intelligence. Thirty-three questions on the *Mehrabian-Epstein Scale of Empathic Tendency* are designed to measure the subject's general empathic tendency. And this process continues outside of academia: When Reynaldo Nehemiah, former world record holder in the 110-meter hurdles, expressed interest in playing football with the San Francisco 49ers, they measured his 40-yard speed, his vertical leaping ability, and his strength in a number of lifts as indicators of his ability to play football successfully.

Of the thousands of measurement scales that have been constructed, two critical questions are asked of each: "Is it reliable?" and "Is it valid?" The question of *reliability* (the topic of this chapter) addresses the issue of whether this instrument will produce the same results each time it is administered to the same person in the same setting. Instruments used in the sciences are generally considered reliable if they produce similar results regardless of who administers them and regardless of which forms are used. The tests given to Nehemiah were certainly reliable: The same tests given many times over several weeks would yield a 40-yard dash of about 4.2 seconds, a vertical leap of about 53 inches, and a bench press of about 355 pounds. But this raised the second question. The cognoscenti of football coined the phrase, "Yeah, but can he take a hit?" You see, in track and field it is considered impolite to tackle your opponent while running a hurdles race! They acknowledged Nehemiah's exceptional physical skills, but he hadn't played football since high school. Were these measures of physical skill good predictors of his ability to play in the NFL? They were concerned about the *validity* of the measure used by the 49ers. In a general sense, validity asks the question, "Does it actually measure what it is trying to measure?" In the case of Nehemiah, they wondered if measures of 40 yards speed, leaping ability, and strength actually measured his ability to play professional football.

This chapter deals with the issue of *reliability*. We addressed the issue of *validity* briefly because the two words are so often linked together, and many budding researchers get the two mixed up. We do not have a chapter on validity, unfortunately, because validity is frequently determined by nonstatistical means. Construct validity is sometimes assessed with the aid of factor analysis, and discriminant analysis can assist in creating an instrument that measures well what the researcher is attempting to measure; but face validity is established by observational procedures, and construct validity (although factor analysis may be used to assist) is often theory based.

The two types of *reliability* discussed in this chapter are Chronbach's alpha (also referred to as *coefficient alpha* or α) and *split-half reliability*. There are other measures of reliability (these will be mentioned but not explained in the Step by Step section), but α is the measure that is most widely used. In this chapter, we will first explain the theoretical and statistical procedure on which coefficient alpha is based and then, more briefly, do the same for split-half reliability. We will then present the example that will be used to demonstrate coefficient alpha and split-half reliability.

## 18.1 Coefficient Alpha (α)

Chronbach's alpha is designed as a measure of internal consistency; that is, do all items within the instrument measure the same thing? Alpha is measured on the same scale as a Pearson *r* (correlation coefficient) and typically varies between 0 and 1. Although a negative value is possible, such a value indicates a scale in which some items measure the opposite of what other items measure. The closer the alpha is to 1.00, the greater the internal consistency of items in the instrument being assessed. At a more conceptual level, coefficient alpha may be thought of as the correlation between a test score and all other tests of equal length that are drawn randomly from the same population of interest. For instance, suppose 1,000 questions existed to test students' knowledge of course content. If ten 100-item tests were drawn randomly from this set, coefficient alpha would approximate the average correlation between all pairs of tests. The formula that determines alpha is simple and makes use of the number of items in the scale (*k*) and the average correlation between pairs of items (*r*):

$$\alpha = \frac{kr}{1 + (k - 1)r}$$

As the number of items in the scale (*k*) increases, the value of α becomes larger. Also, if the intercorrelation between items is large, the corresponding α will also be large.

## 18.2 Split-Half Reliability

Split-half reliability is most frequently used when the number of items is large and it is possible to create two halves of the test that are designed to measure the same thing. An example for illustration is the *16 Personality Factor Questionnaire* (16PF) of Raymond Cattell. This 187-item test measures 16 different personality traits. There are 10 to 14 questions that measure each trait. If you wished to do a split-half reliability measure with the 16PF, it would be foolish to compare the first half of the test with the second half, because questions in each half are designed to measure many different things. It would be better to compute reliabilities of a single trait. If you wished to check the reliability of **aggression** (assessed by 14 questions), the best procedure would be to select 7 questions randomly from the 14 and compare them to the other 7. Once item selection is completed, SPSS conducts split-half reliability by computing correlations between the two parts. The split-half procedure can also be utilized if two different forms of a test are taken, or if the same test is administered more than once.

## 18.3 The Example

We use real data from the **helping2.sav** file ($N = 517$) to demonstrate reliability analysis. One segment of this file contains 14 questions that measure the subjects' empathic tendency. These are questions from the *Mehrabian-Epstein Scale of Empathic Tendency* mentioned earlier. The variables are named **empathy1** to **empath14**.

# 18.4 Step by Step

## 18.4.1 Reliability Analysis

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | 🖱 ⊞ → 🖱 All Programs ▷ → 🖱 IBM SPSS Statistics → 🖱 IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | 🖱 🚀 → 🖱 ■ → 🖱 Σ | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| In Screen | Do This | Step 2 |
|---|---|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* | | |

*Screens 1 and 2 (displayed on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **helping2.sav** *file for further analyses:*

| In Screen | Do This | Step 3 |
|---|---|---|
| Front1 | 🖱 File → Open → 🖱 Data [*or* 🖱 📂] | |
| Front2 | type `helping2.sav` → 🖱 Open [*or* 🖱 🖱 helping2.sav] | Data |

Whether first entering SPSS or returning from earlier operations the standard menu of commands across the top is required. As long as it is visible you may perform any analyses. It is not necessary for the data window to be visible.

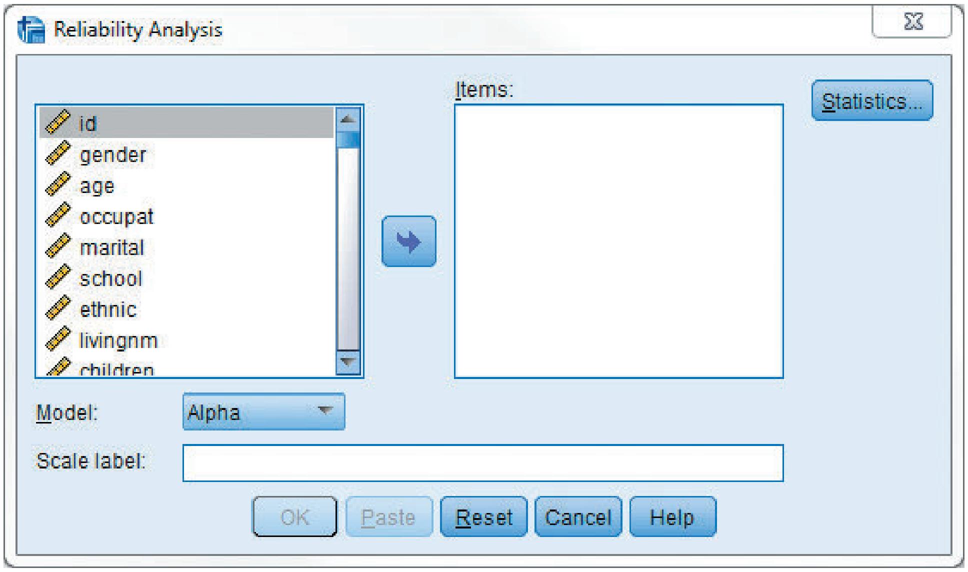
*After completion of Step 3 a screen with the desired menu bar appears. When you click a command (from the menu bar), a series of options will appear (usually) below the selected command. With each new set of options, click the desired item. The sequence to access reliability analysis begins at any screen with the menu of commands visible:*

| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | 🖱 Analyze → Scale → 🖱 Reliability Analysis | 18.1 |

**Reliability Analysis** is a popular and frequently used procedure and the SPSS method of accessing reliability analysis is user friendly and largely intuitive. There are only two dialog boxes that provide different selections for the types of analyses offered: the main dialog box (Screen 18.1, on the following page) and a variety of **Statistics** options (Screen 18.2, page 235).

Within the main dialog window several options are available. Similar to other main dialog windows, the variables are listed to the left and the five standard-function pushbuttons beneath. A single **Items** box provides room to enter the variables you wish to include in reliability analysis. When you place variables into the active box (the **Items** box for reliability analysis), you may select them in any order. If they are already in consecutive order this makes your job that much easier. In the menu of items to the right of **Model**, the default (**Alpha**) appears in the small window, but a click on the ▾ will reveal several others.

Screen 18.1 The Reliability Analysis Window



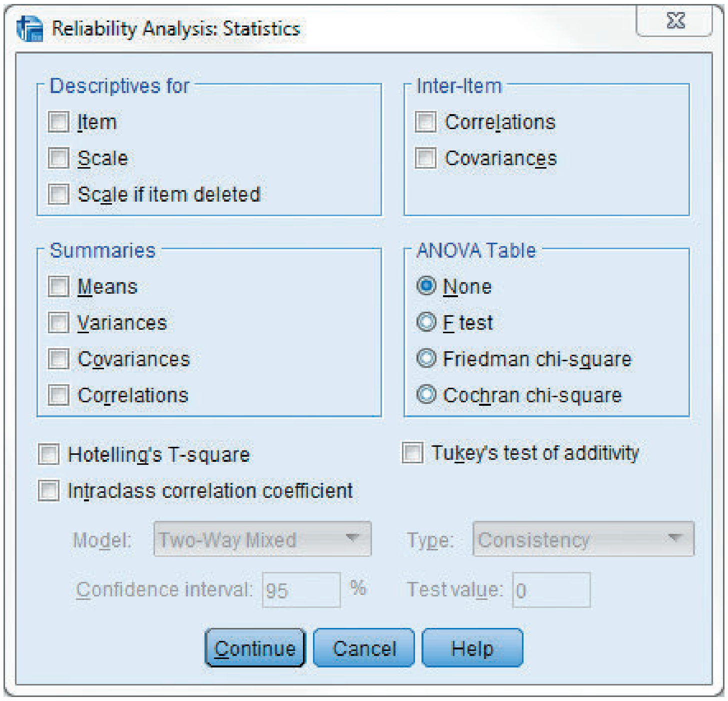
- **Split-half:** From a single list of variables the split-half procedure will compare variables in the first half of the list (the first half is one larger if there are an odd number of items) with variables in the second half of the list.
- **Guttman:** This option calculates reliability based on a lower bounds procedure.
- **Parallel:** Computes reliability based on the assumption that all included items have the same variance.
- **Strict parallel:** Computes a reliability based on the assumption that all items have the same mean, the same true score variance, and the same error variance over multiple administrations of the instrument.

A click on the **Statistics** pushbutton opens a dialog box (Screen 18.2, following page) that presents the remaining options for this procedure. All items will be briefly described here except for items in the **ANOVA Table** box. The ANOVA and Chi-square options provided there are described in some detail elsewhere in this book, and, frankly, are rarely used in reliability analysis. First the **Descriptives for** items will be described, followed by **Summaries**, **Inter-Item**, **Hotelling's T-square**, and **Tukey's test of additivity**.

- **Descriptives for Item:** This provides simple means and standard deviations for each variable included in the analysis.
- **Descriptives for Scale:** This option provides the mean, variance, standard deviation, and *N* for the *sum* of all variables in the scale. More specifically, for each subject the scores for the 14 empathy questions are summed. Descriptives are then provided for this single summed variable.
- **Descriptives for Scale if item deleted:** For each variable this option identifies the resulting alpha value if that item were deleted from the scale. Almost always this option should be selected.
- **Summaries for Means:** This option computes the mean for all subjects for each of the entered variables. In the present example, it computes the mean value for **empathy1**, the mean value for **empathy2**, the mean value for **empathy3**, and so forth up to the mean value of **empath14**. Then it lists the *mean* of these 14 means, the *minimum* of the 14 means, the *maximum* of the 14 means, the *range*, the *maximum divided by the minimum*, and the *variance* of the 14 means. Although a bit complex, this is a central construct of reliability analysis.

- **Summaries for Variances:** The same as the previous entry except for the variances.
- **Summaries for Covariances:** This procedure computes the covariance between each entered variable and the sum of all other variables. In the present example, a covariance would be computed between **empathy1** and the sum of the other 13, between **empathy2** and the sum of the other 13, and so forth for all 14 variables. Then the final table would provide the mean of these 14 covariances, the minimum, maximum, range, minimum divided by maximum, and variance.
- **Summaries for Correlations:** The same as the procedure for covariances except that all values are computed for correlations instead. In the present example, the mean of these 14 correlations is the *r* in the alpha formula.
- **Inter-Item Correlations:** The simple correlation matrix of all entered variables.
- **Inter-Item Covariances:** The simple covariance matrix of all entered variables.
- **Hotelling's T-square:** This is a multivariate T-test that tests whether all means of the entered variables are equal.
- **Tukey's test of additivity:** This is essentially a test for linear dependency between entered variables.

Screen 18.2 Reliability Analysis: Statistics Window



Reliability Analysis: Statistics

Descriptives for
□ Item
□ Scale
□ Scale if item deleted

Inter-Item
□ Correlations
□ Covariances

Summaries
□ Means
□ Variances
□ Covariances
□ Correlations

ANOVA Table
● None
○ F test
○ Friedman chi-square
○ Cochran chi-square

□ Hotelling's T-square
□ Tukey's test of additivity
□ Intraclass correlation coefficient

Model: Two-Way Mixed
Type: Consistency
Confidence interval: 95 %
Test value: 0

Three different step-by-step sequences will be presented: Step 5 will describe the simplest default procedure for conducting reliability analysis. In some settings a default procedure may be quite useful. Often you may want a quick check of how reliable some measure is without thought about further analyses. Step 5a will create a reliability analysis that includes a number of options discussed earlier. It turns out that when using all 14 of the empathy questions the reliability is fairly low. By deleting five of the items that hurt the internal consistency of the measure, nine items are left that make a reliable scale. This second analysis will include only those variables that contribute to the maximum alpha value. Finally, Step 5b will demonstrate how to access a split-half reliability with the nine (rather than the 14) variables. The italicized text before each sequence box will identify the specifics of the analysis. All three analyses are included in the Output section.

*To conduct a reliability analysis with the 14 empathic tendency questions (***empathy1** *to* **empath14***) with all the default options, perform the following sequence of steps. The starting point is Screen 18.1. Carry out the Step-4 sequence (page 233) to arrive at this screen.*

| In Screen | Do This | Step 5 |
|---|---|---|
| 18.1 | ✳ **empathy1** → press **Shift** *and* ✳ **empath14** → ✳ ➡ → ✳ **OK** | Back1 |

A click and drag operation is no longer allowed with the variable list. You are now required to use key strokes along with **Shift** and **Ctrl** keys to select several variables. The output from this sequence includes the alpha value, the number of subjects and the number of variables.

*To conduct a reliability analysis based on coefficient alpha with the nine empathic tendency questions that yielded the highest alpha* (**empathy1-3-4-5-6-7-9-11-12**), *to request* **Scale**, **Scale if item deleted**, **Means**, **Variances**, **Correlations**, **Inter-Item Correlations** *and a list of the value labels, perform the following sequence of steps. The starting point is Screen 18.1. Carry out the four-step sequence (page 233) to arrive at this screen.*

| In Screen | Do This | Step 5a |
|---|---|---|
| 18.1 | *Hold down* **Ctrl** *key while you* ✳ **empathy1** → ✳ **empathy3** → ✳ **empathy4** → | |
| | ✳ **empathy5** → ✳ **empathy6** → ✳ **empathy7** → ✳ **empathy9** → | |
| | ✳ **empath11** → ✳ **empath12** → ✳ ➡ → ✳ **Statistics** | |
| 18.2 | ✳ **Scale** → ✳ **Scale if item deleted** → ✳ **Means** → ✳ **Correlations** → | |
| | ✳ **Variances** → ✳ **Inter-Item Correlations** → ✳ **Continue** → | |
| 18.1 | ✳ **OK** | Back1 |

*If you wish to perform an identical procedure to that demonstrated in sequence Step 5a but conducting* **Split-half** *reliability rather than using* **Alpha**, *perform the following sequence of steps. The starting point is Screen 18.1.*

| In Screen | Do This | Step 5b |
|---|---|---|
| 18.1 | *Hold down* **Ctrl** *key while you* ✳ **empathy1** → ✳ **empathy3** → ✳ **empathy4** → | |
| | ✳ **empathy5** → ✳ **empathy6** → ✳ **empathy7** → ✳ **empathy9** → | |
| | ✳ **empath11** → ✳ **empath12** → ✳ ➡ → ✳ ▼ *next to* **Model** → | |
| | ✳ **Split-half** → ✳ **Statistics** | |
| 18.2 | ✳ **Scale** → ✳ **Scale if item deleted** → ✳ **Means** → ✳ **Correlations** → | |
| | ✳ **Variances** → ✳ **Inter-Item Correlations** → ✳ **Continue** → | |
| 18.1 | ✳ **OK** | Back1 |

Upon completion of Step 5, 5a, or 5b, the output screen will appear (Screen 1, inside back cover). All results from the just-completed analysis are included in the Output

Navigator. Make use of the scroll bar arrows (▲▼▶◀) to view the results. Even when viewing output, the standard menu of commands is still listed across the top of the window. Further analyses may be conducted without returning to the data screen.

## 18.5 Printing Results

Results of the analysis (or analyses) that have just been conducted require a window that displays the standard commands (**File Edit Data Transform Analyze...**) across the top. A typical print procedure is shown below beginning with the standard output screen (Screen 1, inside back cover).

*To print results, from the Output screen perform the following sequence of steps:*

| In Screen | Do This | Step 6 |
|---|---|---|
| Back1 | *Select desired output or edit (see pages 16–22)* → 🖱 **File** → 🖱 **Print** | |
| 2.9 | *Consider and select desired print options offered, then* → 🖱 **OK** | |

*To exit you may begin from any screen that shows the* **File** *command at the top.*

| In Screen | Do This | Step 7 |
|---|---|---|
| Menu | 🖱 **File** → 🖱 **Exit** | 2.1 |

Note: After clicking **Exit**, there will frequently be small windows that appear asking if you wish to save or change anything. Simply click each appropriate response.

## 18.6 Output

### 18.6.1 Reliability Analysis

What follows is partial output from the three sequence steps (5, 5a, and 5b) presented on page 236. We have included an **Item-total statistics** for all 14 variables (even though it was not selected in the default procedure) for sake of illustration. A slightly condensed version of the output in the standard format is reproduced here. We begin with the list of variable names and labels. This is followed by the item-total statistics for all 14 variables, used to illustrate the criteria by which certain variables may be eliminated to improve the alpha. This is followed by scale statistics, inter-item correlations, and the item-total statistics for the nine remaining items after five have been removed from the analysis. The section concludes with output from a split-half analysis. Definitions and text are included between sections to aid in interpretation. Below are variable names and labels for your reference.

| # | Variable | Label |
|---|---|---|
| 1. | EMPATHY1 | Sad to see lonely stranger |
| 2. | EMPATHY2 | Annoyed at sorry-for-self people (*neg) |
| 3. | EMPATHY3 | Emotionally involved with friends problem |
| 4. | EMPATHY4 | Disturbed when bring bad news |
| 5. | EMPATHY5 | Person crying upsetting |
| 6. | EMPATHY6 | Really involved in a book or movie |
| 7. | EMPATHY7 | Angry when someone ill-treated |
| 8. | EMPATHY8 | Amused at sniffling at movies (*neg) |
| 9. | EMPATHY9 | Do not feel ok when others depressed |

*(Continued)*

| # | Variable | Label |
|---|---|---|
| 10. | EMPATH10 | Hard to see why others so upset (*neg) |
| 11. | EMPATH11 | Upset to see animal in pain |
| 12. | EMPATH12 | Upset to see helpless old people |
| 13. | EMPATH13 | Irritation rather than sympathy at tears (*neg) |
| 14. | EMPATH14 | Remain cool when others excited (*neg) |

**Item-Total Statistics**

|  | Scale Mean if Item Deleted | Scale Variance if Item Deleted | Corrected Item-Total Correlation | Squared Multiple Correlation | Cronbach's Alpha if Item Deleted |
|---|---|---|---|---|---|
| empathy1 | 62.27 | 81.313 | .403 | .326 | .654 |
| empathy2 | 63.37 | 94.520 | −.053 | .120 | .718 |
| empathy3 | 62.57 | 81.323 | .457 | .350 | .648 |
| empathy4 | 62.17 | 81.702 | .453 | .415 | .649 |
| empathy5 | 62.40 | 79.322 | .473 | .372 | .644 |
| empathy6 | 62.31 | 81.852 | .385 | .232 | .657 |
| empathy7 | 61.22 | 84.746 | .398 | .240 | .659 |
| empathy8 | 62.13 | 84.895 | .228 | .272 | .680 |
| empathy9 | 62.83 | 85.395 | .271 | .211 | .672 |
| empath10 | 62.75 | 91.286 | .048 | .121 | .704 |
| empath11 | 61.60 | 82.256 | .400 | .325 | .655 |
| empath12 | 61.51 | 81.987 | .436 | .411 | .651 |
| empath13 | 61.45 | 84.628 | .303 | .250 | .668 |
| empath14 | 63.34 | 88.461 | .164 | .101 | .686 |

**Reliability Statistics**

| Cronbach's Alpha | Cronbach's Alpha Based on Standardized Items | N of Items |
|---|---|---|
| .685 | .698 | 14 |

The first chart lists the variables for which a reliability check is being conducted. Note that five of the items are marked with a "(*neg)". These are questions that have been phrased negatively to control for response bias. The scoring on these items is reversed before beginning calculations. The second chart is designed to designate similarities and differences between each variable and composites of the other variables. Some terminology (defined below) will aid understanding.

| Term | Definition/Description |
|---|---|
| **SCALE MEAN IF ITEM DELETED** | For each subject the 13 variables (excluding the variable to the left) are summed. The values shown are the means for the 13 variables across all 517 subjects. |
| **SCALE VARIANCE IF ITEM DELETED** | The variance of summed variables when the variable to the left is deleted. |
| **CORRECTED ITEM-TOTAL CORRELATION** | Correlation of the designated variable with the sum of the other 13. |
| **ALPHA IF ITEM DELETED** | The resulting alpha if the variable to the left is deleted. |

Notice that for four of the variables, the correlations between each of them and the sum of all other variables is quite low; in the case of **empathy2**, it is even negative (−.053). Correspondingly, the **Alpha** value would increase (from the .685 shown on the lower chart) if these items were deleted from the scale. In the analyses that follow, all items that *reduce* the alpha value have been deleted. After removing the four designated variables, a second analysis was run, and it was found that a fifth variable's removal would also increase the value of **Alpha**. Five variables were eventually dropped from the scale: variables 2, 8, 10, 13, and 14. Please be vividly aware that variables are not *automatically* dropped just because a higher **Alpha** results. There are often theoretical or practical reasons for keeping them.

Interpretation of the descriptive statistics and correlation matrix are straightforward and not shown. The output that follows involves analyses *after* the five variables were removed.

**Scale Statistics**

| Mean | Variance | Std. Deviation | N of Items |
|---|---|---|---|
| 44.75 | 71.131 | 8.434 | 9 |

**Summary Item Statistics**

| | Mean | Minimum | Maximum | Range | Max/Min | Variance | N of Items |
|---|---|---|---|---|---|---|---|
| Item Means | 4.973 | 4.242 | 5.849 | 1.607 | 1.379 | .286 | 9 |
| Item Variances | 2.317 | 1.632 | 2.663 | 1.031 | 1.631 | .110 | 9 |
| Inter-Item Correlations | .303 | .120 | .521 | .401 | 4.343 | .009 | 9 |

These items were described and explained earlier; the definitions of terms follow.

| Term | Definition/Description |
|---|---|
| **SCALE STATISTICS** | There are a total of nine variables being considered. This line lists descriptive information about the *sum* of the nine variables for the entire sample of 517 subjects. |
| **ITEM MEANS** | This is descriptive information about the 517 subjects' means for the nine variables. In other words, the $N = 9$ for this list of descriptive statistics. The "mean of means" = 4.973. The minimum of the nine means = 4.242, and so forth. |
| **ITEM VARIANCES** | A similar construct as that used in the line above (**ITEM MEANS**). The first number is the mean of the nine variances, the second is the lowest of the nine variances, and so forth. |
| **INTER-ITEM CORRELATIONS** | This is descriptive information about the correlation of each variable with the sum of the other eight. There are nine correlations computed: the correlation between the first variable and the sum of the other eight variables, the correlation between the second variable and the sum of the other eight, and so forth. The first number listed is the mean of these nine correlations (.303), the second is the lowest of the nine (.120), and so forth. The mean of the inter-item correlations (.303) is the *r* in the $\alpha = rk/[1+(k-1)r]$ formula. |

The table that follows is the same as that presented earlier except that values shown relate to 9 rather than 14 variables, and an additional column is defined: Squared Multiple Correlation. Further the **Alpha** has increased to .795 (from .685). Also notice that the alpha value cannot be improved by deleting another variable. Four of the columns have been described on the previous page. Terms new to this chart are defined after the following output.

**Item-Total Statistics**

| | Scale Mean if Item Deleted | Scale Variance if Item Deleted | Corrected Item-Total Correlation | Squared Multiple Correlation | Cronbach's Alpha if Item Deleted |
|---|---|---|---|---|---|
| empathy1 | 39.95 | 55.847 | .522 | .316 | .770 |
| empathy3 | 40.26 | 57.477 | .509 | .337 | .772 |
| empathy4 | 39.85 | 56.141 | .590 | .400 | .762 |
| empathy5 | 40.08 | 54.658 | .572 | .368 | .763 |
| empathy6 | 39.99 | 58.630 | .399 | .221 | .787 |
| empathy7 | 38.91 | 61.109 | .420 | .209 | .783 |
| empathy9 | 40.51 | 59.014 | .399 | .185 | .787 |
| empath11 | 39.28 | 58.630 | .432 | .314 | .782 |
| empath12 | 39.20 | 57.263 | .527 | .403 | .770 |

**Reliability Statistics**

| Cronbach's Alpha | Cronbach's Alpha Based on Standardized Items | N of Items |
|---|---|---|
| .795 | .796 | 9 |

| Term | Definition/Description |
|---|---|
| **SQUARED MULTIPLE CORRELATION** | These values are determined by creating a multiple regression equation to generate the predicted correlation based on the correlations for the other eight variables. The numbers listed are these predicted correlations. |
| **ALPHA** | Based on the formula: $\alpha = rk/[1+(k-1)r]$, where *k* is the number of variables considered (9 in this case) and *r* is the mean of the inter-item correlations (.303, from previous page). The alpha value is inflated by a larger number of variables; so there is no set interpretation as to what is an acceptable alpha value. A rule of thumb that applies to most situations is:<br>$\alpha > .9$—excellent<br>$\alpha > .8$—good<br>$\alpha > .7$—acceptable<br>$\alpha > .6$—questionable<br>$\alpha > .5$—poor<br>$\alpha < .5$—unacceptable |
| **ALPHA BASED ON STANDARDIZED ITEMS** | This is the alpha produced if the composite items (in the chart on the previous page, not the scores for the 517 subjects) are changed to *z* scores before computing the **Alpha**. The almost identical values produced here (.795 vs. .796) indicate that the means and variances in the original scales do not differ much, and thus standardization does not make a great difference in the **Alpha**. |

## 18.6.2 Split-Half Reliability

The output from a split-half reliability follows. Recall that we are conducting a split-half analysis of the same nine empathy questions. This would typically be considered too small a number of items upon which to conduct this type of analysis. However, it serves as an adequate demonstration of the procedure and does not require description of a different data set. Much information in the **Split-half** output is identical to that for the **Alpha** output. Only output that is different is included here.

**Scale Statistics**

| | Mean | Variance | Std. Deviation | N of Items |
|---|---|---|---|---|
| Part 1 | 23.64 | 31.112 | 5.578 | 5 |
| Part 2 | 21.12 | 17.108 | 4.136 | 4 |
| Both Parts | 44.75 | 71.131 | 8.434 | 9 |

**Summary Item Statistics**

| | | Mean | Minimum | Maximum | Range | Max/Min | Variance | N of Items |
|---|---|---|---|---|---|---|---|---|
| Item Means | Part 1 | 4.727 | 4.499 | 4.901 | .402 | 1.089 | .023 | 5 |
| | Part 2 | 5.279 | 4.242 | 5.849 | 1.607 | 1.379 | .505 | 4 |
| | Both Parts | 4.973 | 4.242 | 5.849 | 1.607 | 1.379 | .286 | 9 |
| Item Variances | Part 1 | 2.446 | 2.120 | 2.663 | .543 | 1.256 | .069 | 5 |
| | Part 2 | 2.155 | 1.632 | 2.474 | .842 | 1.516 | .138 | 4 |
| | Both Parts | 2.317 | 1.632 | 2.663 | 1.031 | 1.631 | .110 | 9 |
| Inter-Item Correlations | Part 1 | .389 | .260 | .504 | .244 | 1.940 | .005 | 5 |
| | Part 2 | .332 | .226 | .521 | .295 | 2.306 | .010 | 4 |
| | Both Parts | .303 | .120 | .521 | .401 | 4.343 | .009 | 9 |

**Reliability Analysis—Split-half**

| | |
|---|---|
| Alpha For Part 1 (5 Items) | .759 |
| Alpha For Part 2 (4 Items) | .662 |
| Correlation Between Forms | .497 |
| Equal-Length Spearman-Brown | .664 |
| Unequal-Length Spearman-Brown | .666 |
| Guttman Split-Half | .644 |

The Statistics for Item Means, Item Variances, and Inter-Item Correlations sections are identical to those shown on page 239 and it has equivalent numbers for the two halves of the scale as well as for the entire scale. The Scale Statistics values are also identical to those on page 239. Please refer there for an explanation. Notice that the alpha values (last line of output) for each half are lower than that for the whole scale. This reflects the reality that a scale with a fewer number of items produces a deflated alpha value. Definitions of new terms follow.

| Term | Definition/Description |
|---|---|
| **CORRELATION BETWEEN FORMS** | An estimate of the reliability of the measure if it had five items. |
| **EQUAL-LENGTH SPEARMAN-BROWN** | The reliability of a 10-item test (.6636) if it was made up of equal parts that each had a five-item reliability of .4965. |
| **GUTTMAN SPLIT-HALF** | Another measure of the reliability of the overall test, based on a lower-bounds procedure. |
| **UNEQUAL-LENGTH SPEARMAN-BROWN** | The reliability acknowledging that this test has 9 and not 10 items. |

# Exercises

Answers to selected exercises are downloadable at **www.spss-step-by-step.net**.

Measure the internal consistency of the following sets of variables. Compute coefficient alpha for the following sets of variables and then delete variables until you achieve the highest possible alpha value. Print out relevant results.

Use the **helping3.sav** file (downloadable at the address shown above) for exercises 1–10. An "h" in front of a variable name refers to assessment by the help giver; an "r" in front of a variable name refers to assessment by the help recipient.

1. Measure of problem severity: **hsevere1, hsevere2, rsevere1, rsevere2**
2. Measure of helper's sympathy: **sympath1, sympath2, sympath3, sympath4**
3. Measure of helper's anger: **anger1, anger2, anger3, anger4**
4. How well the recipient is coping: **hcope1, hcope2, hcope3, rcope1, rcope2, rcope3**
5. Helper rating of time spent helping: **hhelp1-hhelp15**
6. Recipient's rating of time helping: **rhelp1-rhelp15**
7. Helper's rating of empathy: **empathy1-empath14**
8. Quality of help: **hqualit1, hqualit2, hqualit3, rqualit1, rqualit2, rqualit3**
9. Helper's belief of self-efficacy: **effic1-effic15**
10. Controllability of the cause of the problem: **hcontro1, hcontro2, rcontro1, rcontro2**

Use the **divorce.sav** file (downloadable at the address shown above) for exercises 11–13.

11. Factors disruptive to divorce recovery: drelat-dadjust (16 items)
12. Factors assisting recovery from divorce: arelat-amain2 (13 items)
13. Spirituality measures: sp8-sp57 (18 items)

# Chapter 19
# Multidimensional Scaling

MULTIDIMENSIONAL SCALING is useful because a picture is often easier to interpret than a table of numbers. In multidimensional scaling, a matrix made up of dissimilarity data (for example, ratings of how different products are, or distances between cities) is converted into a one-, two-, or three-dimensional graphical representation of those distances. Although it is possible to have more than three dimensions in multidimensional scaling, that is rather rare.

A particular advantage of multidimensional scaling graphs is that the distances between two points can be interpreted intuitively. If two points on a multidimensional scaling graph are far apart from each other, then they are quite dissimilar in the original dissimilarity matrix; if two points are close together on the graph, then they are quite similar in the original matrix.

Multidimensional scaling has some similarities with factor analysis (see Chapter 20). Like factor analysis, multidimensional scaling produces a number of dimensions along which the data points are plotted; unlike factor analysis, in multidimensional scaling the meaning of the dimensions or factors is not central to the analysis. Like factor analysis, multidimensional scaling reduces a matrix of data to a simpler matrix of data; whereas factor analysis typically uses correlational data, multidimensional scaling typically uses dissimilarity ratings. Finally, factor analysis produces plots in which the angles between data points are the most important aspect to interpret; in multidimensional scaling, the location in space of the data points (primarily the distance between the points) is the key element of interpretation.

There are variations on the multidimensional scaling theme that use similarity data instead of dissimilarity data. These are included in the SPSS Categories option, which is not covered in this book.

There are also a number of similarities between multidimensional scaling and cluster analysis (see Chapter 21). In both procedures, the proximities or distances between cases or variables may be examined. In cluster analysis, however, a qualitative grouping of the cases into clusters is typical; in multidimensional scaling, rather than ending with clusters we end with a graph in which we can both visually and quantitatively examine the distances between each case.

The SPSS multidimensional scaling procedure (historically known as ALSCAL) is in fact a collection of related procedures and techniques rather than a single procedure. In this chapter, we will focus on three analyses that exemplify several key types of data that may be analyzed, and attempt to briefly yet clearly define the key terms involved. If this is your first experience with multidimensional scaling, we would suggest that you refer to Davidson (1992) or Young and Hamer (1987); references are provided on page 372.

In the first example, a sociogram is computed for a selection of students in a class; in this case, ratings of how much they dislike each other are converted into a graph of how far apart they want to be from each other. In the second example, we will let SPSS compute dissimilarity scores for students' quiz scores, and produce a graph of how far apart the students are in their patterns of quiz scores. Finally, in the third example a small group of student' ratings of dissimilarities between television shows are analyzed.

## 19.1 Square Asymmetrical Matrixes (The Sociogram Example)

Imagine that an instructor wanted to create the perfect seating chart of a class. In order to do this, she asked each student to rate each other student, asking them, "On a scale of 1 (not at all) to 5 (very much), how much do you dislike _____?" She asked each of the 20 students in her class to rate the other 19 students (they did not rate themselves). The raw data for this fictitious example is included in the **grades-mds.sav** file. The teacher would like a visual representation of how far apart each student wants to sit from each other student, so she performs a multidimensional scaling analysis.

In this case, the teacher is starting from a 20 × 20 matrix of dissimilarities. She would like to convert this matrix to a two-dimensional picture indicating liking or disliking. To do this, multidimensional scaling will develop a much simpler matrix (20 students × 2 dimensions) that will, as much as possible, represent the actual distances between the students. This is a square asymmetrical dissimilarity matrix. To translate:

- **Square matrix:** The rows and columns of the matrix represent the same thing. In this case, the rows represent raters, and the columns represent the people being rated. This is a square matrix because the list of students doing the rating is the same as the list of people being rated . . . they are rating each other. The matrix would be called rectangular if the raters were rating something else other than themselves. The analysis and interpretation of rectangular matrices is beyond the scope of this book.
- **Asymmetrical matrix:** The matrix is called asymmetrical because it is quite possible that, for example, Shearer dislikes Springer less than Springer dislikes Shearer. If the numbers below the diagonal in the matrix can be different than the numbers above the diagonal, as in this example, then the matrix is asymmetrical; if the numbers below the diagonal in the matrix are always a mirror image of the numbers above the diagonal in the matrix, then the matrix is symmetric. Correlation tables are a good example of a matrix that is always symmetric.
- **Dissimilarity matrix:** In the data, higher numbers mean more dissimilarity; in this case, higher numbers mean more disliking. If a similarity matrix is used (for example, if students rated how much they liked each other), students that were farthest apart would be the ones that liked each other most (that's why you shouldn't use multidimensional scaling to analyze similarity matrixes).

### 19.1.1 Square Symmetrical Matrixes with Created Distances from Data (The Quiz Score Example)

After successfully using her seating chart for one term, our fictional instructor decides that she wants to use multidimensional scaling to create a seating chart again; this time, she decides to seat students according to their quiz scores on the first five quizzes. In this case, because the quiz scores are ***not*** dissimilarity data, SPSS will be used to calculate dissimilarities. Therefore, we start from a data matrix containing 20 students × 5 quizzes, computing a data matrix containing 20 students × 20 students dissimilarities, and then using multidimensional scaling to produce a 20 students × 2 dimensions matrix, represented visually on a graph, from which we can derive our seating chart. We will again be using data included in the **grades-mds.sav** file, and the analysis will involve a square symmetrical matrix:

- **Square matrix:** Although our original matrix is not square (because we have cases that are students, and variables that are quizzes), we are asking SPSS to compute dissimilarities. The dissimilarity matrix that SPSS computes will have students in rows and in columns; because the rows and columns represent the same thing, the matrix is square.

- **Symmetrical matrix:** The matrix is symmetrical because the dissimilarity between, for example, Liam and Cha is identical to the dissimilarity between Cha and Liam.

### 19.1.2 Individual Difference Models (The TV Show Example)

In each of the previous examples, there was one dissimilarity matrix that was used to compute distances and create a multidimensional mapping of the data. In some cases, however, it is possible to have several dissimilarity matrixes used in a single analysis. This can happen, for example, if you want to replicate your experiment in three different samples; in this case, you would have three dissimilarity matrixes (this is known as replicated multidimensional scaling). You may also have several dissimilarity matrixes if you have several people rating the dissimilarity between things. In our example, we will focus on the later situation (known as individual difference models, or weighted multidimensional scaling models).

Imagine that, because our fictitious teacher has worn herself out developing seating charts, she decides to have her students watch television for an hour each day. After watching several episodes each of *ER*, *60 Minutes*, *The Simpsons*, and *Seinfeld*, she has five of her students rate how different each of the programs are from each of the other programs. These five symmetrical dissimilarity matrixes will be examined to determine how each of the television shows is perceived.

Because SPSS is rather exacting as to the format of the data when using individual difference models, the data for this analysis is included in the a separate **grades-mds2.sav** file. In particular, with individual difference data, you must have several matrixes of the same size in your data file; in our example, because we have five students rating dissimilarity between four television shows, our data file needs to have (5) 4 × 4 matrixes. The first four rows of the file will be the dissimilarity matrix for the first student, the second four rows of the file will be the dissimilarity matrix for the second student, and so forth, for a total of 20 rows in the file.

In the Step by Step section, different versions of Step 5 will be used to describe how to complete each of the three sample analyses.

## 19.2 Step by Step

### 19.2.1 Multidimensional Scaling

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | → All Programs ▷ → IBM SPSS Statistics → IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | → → | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| Step 2 |
|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* |

*Screens 1 and 2 (on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **grades-mds.sav** *file for further analyses:*

| In Screen | Do This | Step 3 |
|---|---|---|
| Front1 | ✳ File → Open → ✳ Data [or ✳ 📂] | |
| Front2 | type `grades-mds.sav` → ✳ Open [or ✳ ✳ grades-mds.sav] | Data |

Whether first entering SPSS or returning from earlier operations the standard menu of commands across the top is required. As long as it is visible you may perform any analyses. It is not necessary for the data window to be visible.

*After completion of Step 3 a screen with the desired menu bar appears. When you click a command (from the menu bar), a series of options will appear (usually) below the selected command. With each new set of options, click the desired item. The sequence to access multidimensional scaling begins at any screen with the menu of commands visible*:

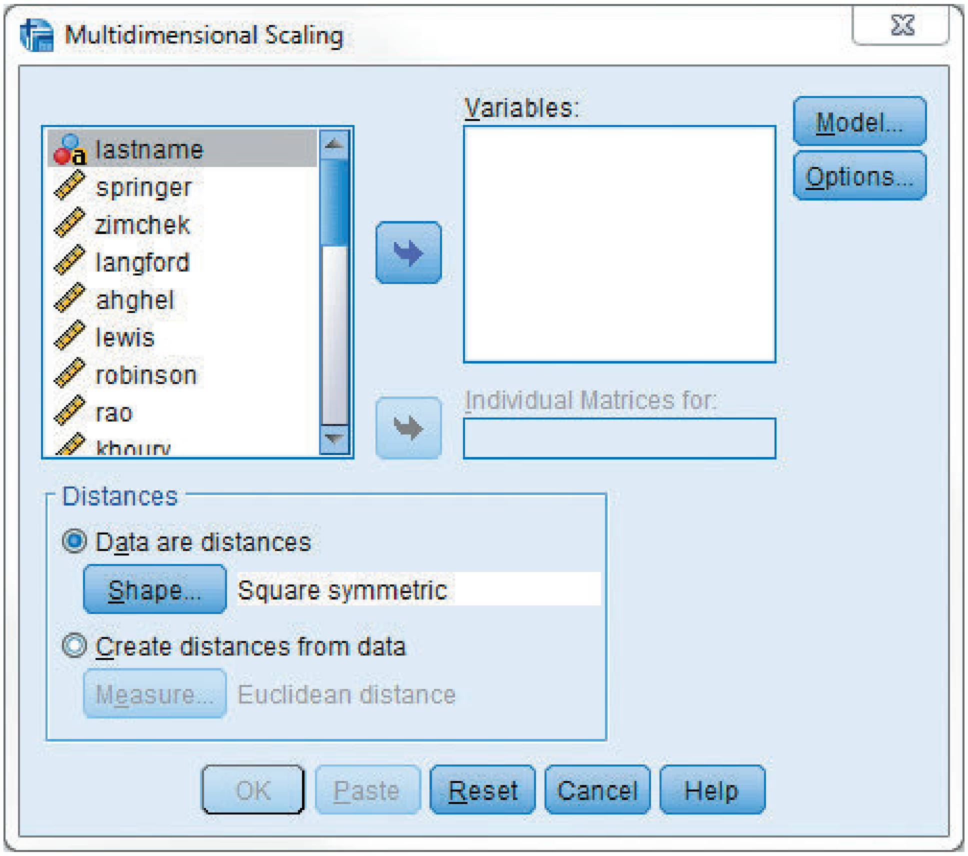
| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | ✳ Analyze → Scale → ✳ Multidimensional Scaling (ALSCAL) | 19.1 |

After this sequence step is performed, a window opens (Screen 19.1, below) in which you select the variables that you want to analyze. There are two different ways of doing a multidimensional scaling analysis in SPSS: You can either provide a dissimilarity matrix or you can ask SPSS to create one.

If you are providing a dissimilarity matrix, then you select which columns in your data matrix you want to analyze. It is possible to have more columns in your data file than you want to include in your analysis, but it is not possible to include more rows in the data file than you want to analyze (unless you use the **Select Cases** command; see pages 73–75). If you are analyzing a square matrix (in which the variables correspond to the cases in your data file, as in our sociogram example), and your data file contains dissimilarity data (rather than having SPSS compute the dissimilarities), then the number of variables that you select should be equal to the number of rows in your data file.

If you want SPSS to calculate a dissimilarity matrix from other data in your file, then SPSS can calculate distances between cases (using the variables you select), or distances between variables (using all the cases).

**Screen 19.1** Multidimensional Scaling Dialog Window



● Data are distances

Shape... Square symmetric

○ Create distances from data

Measure... Euclidean distance

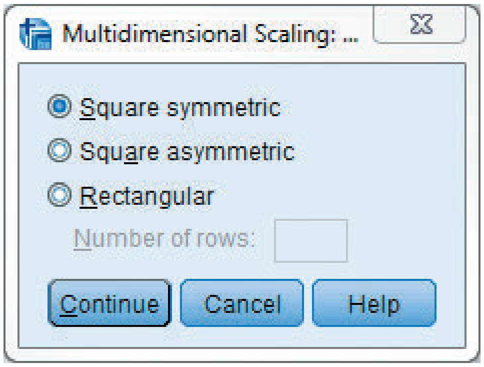
In either case, once you have selected the variables from the list at the left that you want to analyze, click the ➔ button to move them to the **Variables** box.

If you want SPSS to compute distances from your data, and you are also planning to use a replication model or an individual differences (weighted) model, then you can select a variable for the **Individual Matrices for** box. This is a fairly complex procedure, so we do not further discuss this option here.

Once you have selected which variables you want to include in the analysis, it is time to tell SPSS exactly what the structure of your data file is. This is done through the options and dialog buttons in the lower left of Screen 19.1, labeled **Distances**. If your data file contains dissimilarity data (that is, it contains data indicating how different or how far apart cases and variables are) then you should keep the default setting, **Data are distances**. If your data matrix is square symmetric (that is, the dissimilarity values above the diagonal are equal to the dissimilarity values below the diagonal), then you do not need to further specify what kind of data you have. If you are using square asymmetric or rectangular data, then click on the **Shape . . .** button and Screen 19.2 will pop up.

In this screen, you select the shape of your dissimilarity matrix. If your matrix is square (that is, if the items in the rows correspond to the items in the columns) and symmetric (that is, the values below the diagonal in the data matrix are equivalent to the values above the diagonal), then you may leave the **Square symmetric** option selected. This is the type of dissimilarity matrix used in the TV show example, because there is no difference, for example, in ratings of how different *Seinfeld* and *60 Minutes* are rated as compared to how different *60 Minutes* and *Seinfeld* are rated. If your data matrix is square asymmetric (that is, the values below the diagonal in the data matrix are not the same as the values above the diagonal in the data matrix), then select **Square asymmetric**. This is the type of dissimilarity matrix used in the sociogram example. If your data matrix includes different types of things in the rows and the columns, then select a **Rectangular** shape for your data matrix. This type of analysis is rare, and goes beyond the scope of this book.

**Screen 19.2** Multidimensional Scaling: Shape of Data Window



Multidimensional Scaling: ...

● Square symmetric

○ Square asymmetric

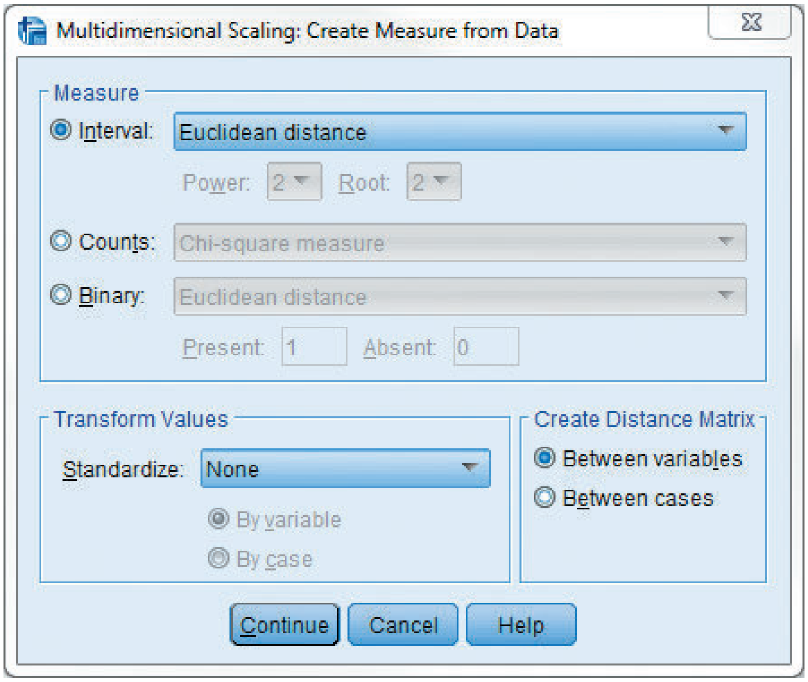
○ Rectangular

Number of rows:

If your data file is not a dissimilarity matrix, but instead you want to calculate dissimilarities from other variables in your data file, then in Screen 19.1 you would select **Create distances from data**. SPSS will automatically create a dissimilarity matrix from the variables you have selected. If you want to use Euclidian distance between data points (think high school geometry and the Pythagorean theorem: $a^2 + b^2 = c^2$) and you want SPSS to calculate distances between variables, then you do not need to further define the **Distances**. If you want to use a different kind of geometry, or if you want to have SPSS calculate the distances between cases instead of between variables (as in our quiz score example), then click the **Measure** button and Screen 19.3 will appear.

In this dialog window you select the **Measure**, whether to **Transform Values**, or whether to **Create Distance Matrix** for variables or cases. In addition to Euclidean distance (described above), other popular distance equations include Minkowski (similar to Euclidian distance, except that you specify the power; if you set power to 4, for example, then $a^4 + b^4 = c^4$) and block (think of a city block, in which you have to walk only on the streets and can't cut across within a block; this may be appropriate

Screen 19.3 Multidimensional Scaling: Create Measures from Data Window



● Interval: Euclidean distance

Power: 2 Root: 2

○ Counts: Chi-square measure

○ Binary: Euclidean distance

Present: 1 Absent: 0

Standardize: None

● By variable

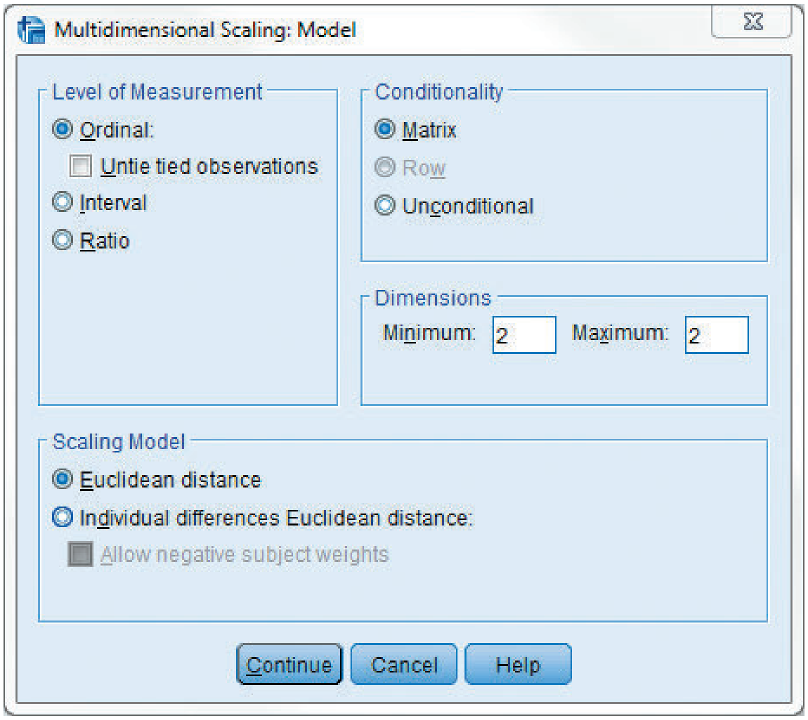
○ By case

● Between variables

○ Between cases

if your variables are measuring different constructs). If all of your variables are measured in the same scale (e.g., they are all using a 1–5 scale), then you do not need to **Transform Values**, and may leave the **Standardize** option at the default setting (None). If your variables are measured using different scales (e.g., some 1–5 scales, some 1–20 scales, and some distances in meters), then you may want to **Transform Values** (with a *Z* score transformation being the most common). Finally, by default SPSS calculates the distances **Between variables**. It is quite possible that you want the distances **Between cases**; if this is the case, select the **Between cases** option.

Screen 19.4 Multidimensional Scaling: Model Window



● Ordinal:

□ Untie tied observations

○ Interval

○ Ratio

● Matrix

○ Row

○ Unconditional

Minimum: 2 Maximum: 2

● Euclidean distance

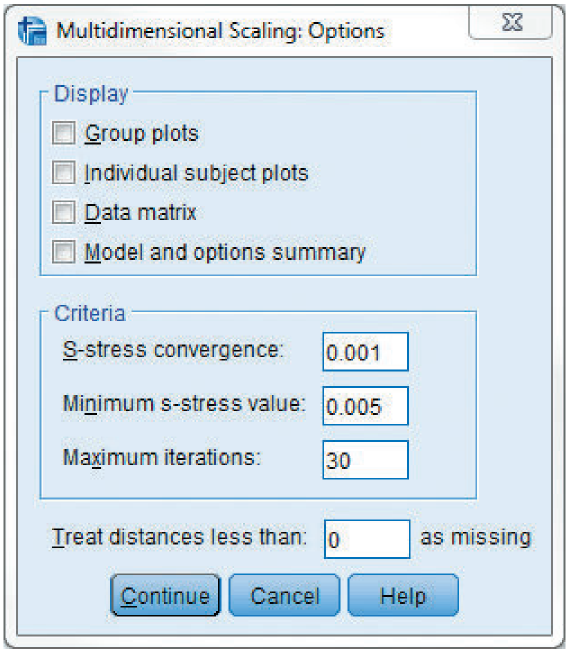
○ Individual differences Euclidean distance:

□ Allow negative subject weights

Once you have defined the **Distances**, you can now tell SPSS what type of **Model** you wish to use. Once you click the **Model** button, Screen 19.4 appears. Of key interest here are the **Dimensions** and the **Scaling Model**. Unlike factor analysis (see Chapter 20), multidimensional scaling cannot automatically determine the number of dimensions that you should use. Since typically only one, two, or three dimensions are desired, this is not a major problem, as you can always do multidimensional scaling for a **Minimum** of one dimension through a **Maximum** of three dimensions and examine each solution to see which makes the most sense (here's where some of the art comes into multidimensional scaling). In most cases, you will want to use a **Euclidian distance Scaling Model** (in which you either have one dissimilarity matrix or have several replicated scaling matrixes), but if you have ratings from several participants and wish to examine the differences between them, then you should click **Individual differences Euclidian distance**. This is the case in the TV shows example, in which five students rated four television shows. **Conditionality** allows you to tell SPSS which comparisons are or are not meaningful: With **Matrix** conditionality, cells with a given matrix may be compared with each other (this is the case in our TV shows example, as student's ratings may be assumed to be comparable with their own ratings but not necessarily with other students' ratings). In the case of **Row** conditionality, all cells within a given row may be compared with each other (this is the case in our sociogram example, as each row represents a particular rater; raters may use different scales, but presumably an individual rater is using the same scale throughout). In **Unconditional** conditionality, SPSS may compare any cell with any other cell. For **Level of Measurement**, you can tell SPSS how to treat your dissimilarity data; there are not often major differences between the models produced when treating data as **Ordinal**, **Interval**, or **Ratio**.

Once the **Model** is specified and you return to Screen 19.1, a click on the **Options** button will bring up Screen 19.5. In this window, you may request **Group plots** (you will almost always want this, as without selecting this option you will have to get out the graph paper to draw the picture yourself!), the **Data matrix** (this shows your raw data and is primarily useful if you are having SPSS compute the dissimilarity matrix for you and you want to take a peek), and **Model and options summary** (this is useful to remind yourself later what type of model you ran). Other options in this screen are rarely needed or advised.

**Screen 19.5** Multidimensional Scaling: Options Window



□ Group plots
□ Individual subject plots
□ Data matrix
□ Model and options summary

S-stress convergence: 0.001
Minimum s-stress value: 0.005
Maximum iterations: 30

Treat distances less than: 0 as missing

*To perform a multidimensional scaling analysis with two dimensions for the 20 students' disliking ratings (the sociogram example), perform the following sequence of steps. This analysis is for a square asymmetric dissimilarity matrix of ordinal data. The starting point is Screen 19.1. Please perform whichever of Steps 1–4 (pages 245–246) are necessary to arrive at this screen.*

| In Screen | Do This | Step 5 |
|---|---|---|
| 19.1 | 🖱 **springer** → 🖱 *hold* **Shift** *and* 🖱 **shearer** → 🖱 *top* ➡ → 🖱 **Shape** | |
| 19.2 | 🖱 **Square asymmetric** → 🖱 **Continue** | |
| 19.1 | 🖱 **Model** | |
| 19.4 | 🖱 **Row** → 🖱 **Continue** | |
| 19.1 | 🖱 **Options** | |
| 19.5 | 🖱 **Group plots** → 🖱 **Continue** | |
| 19.1 | 🖱 **OK** | **Back1** |

*If you wish to perform an analysis with two dimensions using a square symmetric matrix than SPSS computes using other variables in your data file, perform the following sequence of steps. This analysis has SPSS calculate distances based on quiz scores (the quiz score example) and uses that square symmetrical dissimilarity matrix (treated as interval data) to perform multidimensional scaling. The starting point is Screen 19.1. Please perform whichever of Steps 1–4 (pages 245–246) are necessary to arrive at this screen.*

| In Screen | Do This | Step 5a |
|---|---|---|
| 19.1 | 🖱 **quiz1** → *hold* **Shift** *and* 🖱 **quiz5** → 🖱 *top* ➡ → 🖱 **Create distances from data** → 🖱 **Measure** | |
| 19.3 | 🖱 **Between cases** → 🖱 **Continue** | |
| 19.1 | 🖱 **Model** | |
| 19.4 | 🖱 **Interval** → 🖱 **Continue** | |
| 19.1 | 🖱 **Options** | |
| 19.5 | 🖱 **Group plots** → 🖱 **Continue** | |
| 19.1 | 🖱 **OK** | **Back1** |

*If you wish to perform an analysis with two dimensions using a square symmetric matrix with an individual differences model, perform the following sequence of steps. Please note that this example uses the* **grades-mds2.sav** *file; if you wish to try this procedure, please open this file in Step 3, page 246. In this analysis, five participants' ratings of dissimilarity between four television shows are analyzed; the matrixes are square symmetric, and the data are assumed to be ratio data. The starting point is Screen 19.1. Please perform whichever of Steps 1–4 (pages 245–246) are necessary to arrive at this screen.*

| In Screen | Do This | Step 5b |
|---|---|---|
| 19.1 | ✳ **er** → *hold* **Shift** *and* ✳ **seinfeld** → ✳ *top* ➡ → ✳ **Model** | |
| 19.4 | ✳ **Ratio** → ✳ **Individual differences Euclidian distance** → ✳ **Continue** | |
| 19.1 | ✳ **Options** | |
| 19.5 | ✳ **Group plots** → ✳ **Continue** | |
| 19.1 | ✳ **OK** | Back1 |

Upon completion of Step 5, the output screen will appear (Screen 1, inside back cover). All results from the just-completed analysis are included in the Output Navigator. Make use of the scroll bar arrows (▲▼▶◀) to view the results. Even when viewing output, the standard menu of commands is still listed across the top of the window. Further analyses may be conducted without returning to the data screen.

## 19.3 Printing Results

Results of the analysis (or analyses) that have just been conducted require a window that displays the standard commands (**File Edit Data Transform Analyze . . .**) across the top. A typical print procedure is shown below beginning with the standard output screen (Screen 1, inside back cover).

*To print results, from the Output screen perform the following sequence of steps:*

| In Screen | Do This | Step 6 |
|---|---|---|
| Back1 | *Select desired output or edit (see pages 16–22)* → ✳ **File** → ✳ **Print** | |
| 2.9 | *Consider and select desired print options offered, then* → ✳ **OK** | |

*To exit you may begin from any screen that shows the* **File** *command at the top.*

| In Screen | Do This | Step 7 |
|---|---|---|
| Menu | ✳ **File** → ✳ **Exit** | 2.1 |

Note: After clicking **Exit**, there will frequently be small windows that appear asking if you wish to save or change anything. Simply click each appropriate response.

## 19.4 Output

### 19.4.1 Multidimensional Scaling

What follows is partial output from sequence Step 5 (the sociogram example) on page *250*. In several cases in which the output differs substantially with Step 5a (the quiz score example) or Step 5b (the TV shows example), selected output from those procedures is included.

### 19.4.2 Iteration History

Iteration history for the 2 dimensional solution (in squared distances)

```
Iteration         S-stress        Improvement
    1              .34159
    2              .32376            .01783
    .                 .                 .
    .                 .                 .
    .                 .                 .
    8              .30867            .00120
    9              .30779            .00088
Iterations stopped because S-stress improve-
ment is less than .001000
```

SPSS attempts a first model, and the S-Stress formula indicates how far off the model is from the original dissimilarity matrix (lower numbers mean less stress and a better model). SPSS then tries to improve the model, and it keeps improving (each improvement is called an iteration) until the S-Stress doesn't improve very much. If you reach 30 iterations, that probably means that your data have some problems.

### 19.4.3 Stress and Squared Correlations in Distances

Stress and squared correlation (RSQ) in distances

```
Stimulus           Stress            RSQ
    1               .328            .773
    2               .220            .892
    3               .216            .890
    .                 .               .
    .                 .               .
    .                 .               .
         Averaged (rms) over stimuli
           Stress = .259 RSQ = .764
```

For each of the rows in a square matrix that you have specified as row conditional, for each matrix in an individual differences model, and in all cases for the overall model, SPSS calculates the stress and $R^2$. The stress is similar in concept to the S-Stress output above, but uses a different equation that makes comparisons between different analyses with different computer programs easier. RSQ ($R^2$) indicates how much of the variance in the original dissimilarity matrix is accounted for by the multidimensional scaling model (thus, higher numbers are good). If you are running the analysis with several different dimensions (for example, seeing how the graph looks in one, two, or three dimensions), then the stress and $R^2$ values can help you decide which model is most appropriate.

### 19.4.4 Stimulus Coordinates

```
Stimulus     Stimulus      Dimension
 Number        Name            1          2
    1        SPRINGER        .8770     -.0605
    2        ZIMCHEK        1.1382      .1515
    3        LANGFORD       1.0552      .2871
    .            .             .          .
    .            .             .          .
    .            .             .          .
```

For each of the items on the multidimensional scaling graph, the coordinates on the graph are provided (in case you want to create your own graph, or if you want to further analyze the data). For a two-dimensional graph, these are the *x* and *y* coordinates on the graph.

### 19.4.5 Subject Weights (Individuals Difference or Weighted Models Only)

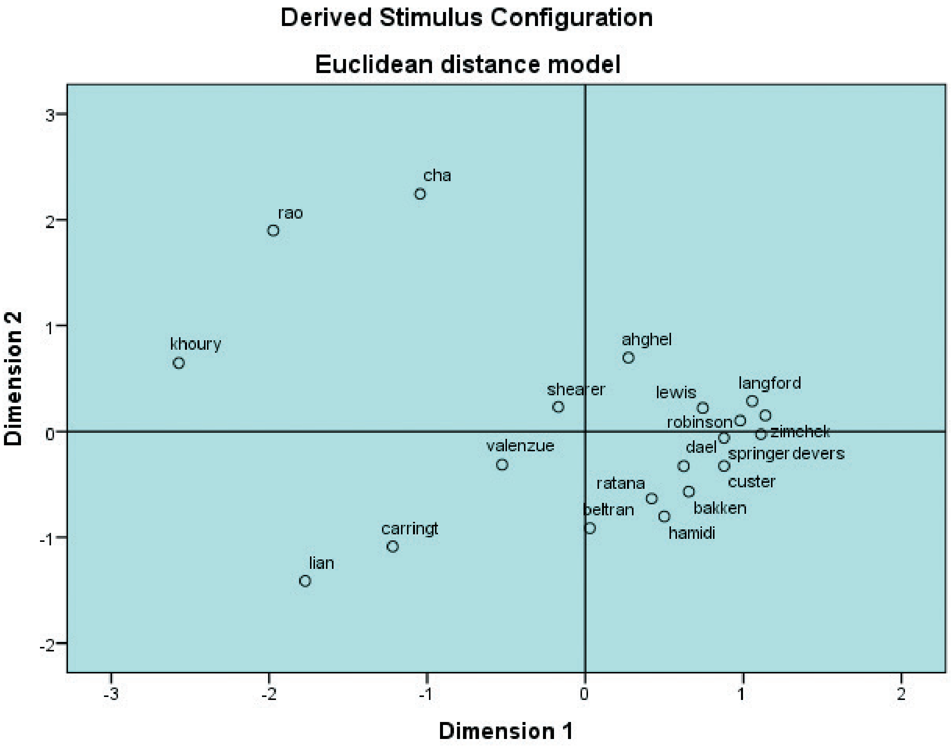
If you are performing an individual differences model or weighted multidimensional scaling, then SPSS will calculate weights indicating the importance of each dimension to each subject. In this TV show example (table below), each of the five subjects has weights on each of the two dimensions. The weights indicate how much each student's variance in his or her ratings was related to each of the two dimensions. An overall weighting of each dimension involving all of the students is also included. "Weirdness" indicates the imbalance in the weights for a particular subject; in this case, subject number 4 (who seemed to base his or her ratings almost totally along dimension 1) has the highest possible weirdness score of 1.

Subject Weights (Individual Difference or Weighted Models Only)

```
  Subject      Weird-          Dimension
  Number       ness          1            2
     1         .2277       .8587        .3679
     2         .3897       .7276        .1124
     3         .1698       .9155        .3564
     4        1.0000       .9584        .0000
     5         .3463       .8339        .4395
Overall importance of
each dimension:            .7437        .0936
```

### 19.4.6 Derived Stimulus Configuration

This graph displays the final multidimensional scaling model. For the sociogram example, this displays the 20 students in the class arranged so that the farther apart the cases are in the graph, the farther apart they are in the original dissimilarity matrix. In this case, we can see which students are most unpopular (Khoury, Rao, and Cha) and which students are most popular (those in the cluster in the middle right of the graph). Note that if you ask SPSS to compute distances for you (as in the quiz scores example), the points on the graph will be labeled "VAR1," "VAR2," and so forth; you will need to go back to your original data file to determine the meaning of each case. The two dimensions do not necessarily have an obvious meaning; the main thing of interest here is the distance between the points.



In this case, one group of students (the clique in the middle right of the graph) is so close together that it is difficult to read their names. If this occurs, you should either (a) look back at the stimulus coordinates above or (b) double click on the graph to go into the output graph editor (see Chapter 5) and use the ⊞ button to select which labels you want displayed.

## Chapter 20

# Factor Analysis

OVER THE past 60 or 70 years, factor analysis has gained increasing acceptance and popularity. Raymond B. Cattell drew attention to the procedure when he used factor analysis to reduce a list of more than 4,500 trait names to fewer than 200 questions that measured 16 different personality traits in his personality inventory called the *16 Personality Factor Questionnaire* (16PF). Cattell's use of factor analysis underlines its primary usefulness, that is, to take a large number of observable instances to measure an unobservable construct or constructs. For instance: "Attends loud parties," "talks a lot," "appears comfortable interacting with just about anyone," and "is usually seen with others" are four behaviors that can be observed that may measure an unobservable construct called "outgoing." Factor analysis is most frequently used to identify a small number of factors (e.g., outgoing) that may be used to represent relationships among sets of interrelated variables (e.g., the four descriptors).

Four basic steps are required to conduct a factor analysis:

1. Calculate a correlation matrix of all variables to be used in the analysis.
2. Extract factors.
3. Rotate factors to create a more understandable factor structure.
4. Interpret results.

The first three steps are covered in this introduction. For Step 4, we give a conceptual feel of how to interpret results, but most of the explanation will take place in the Output section.

## 20.1 Create a Correlation Matrix

Calculating a correlation matrix of all variables of interest is the starting point for factor analysis. This starting point provides some initial clues as to how factor analysis works. Even at this stage it is clear that factor analysis is derived from some combinations of intercorrelations among descriptor variables. It is not necessary to type in a correlation matrix for factor analysis to take place. If you are starting from raw data (as we have in all chapters up to this point), the **Factor** command will automatically create a correlation matrix as the first step. In some instances the researcher may not have the raw data, but only a correlation matrix. If so, it is possible to conduct factor analysis by inserting the correlation matrix into an SPSS syntax command file. The process is complex, however, and a description of it extends beyond the scope of this book. Please consult *SPSS Command Syntax Reference* for specifics.

## 20.2 Factor Extraction

The purpose of the factor-extraction phase is to extract the factors. *Factors* are the underlying constructs that describe your set of variables. Mathematically, this procedure is similar to a forward run in multiple regression analysis. As you recall (Chapter 16), the first step in multiple regression is to select and enter the *independent* variable that significantly explains the greatest amount of variance observed in the *dependent* variable. When this is completed, the next step is to find and enter the independent variable that significantly explains the greatest *additional* amount of variation in the dependent variable. Then the procedure selects and enters the variable that significantly explains the *next* greatest additional amount of variance and so forth, until there are no variables that significantly explain further variance.

With factor analysis the procedure is similar, and a conceptual understanding of the factor-extraction phase could be gained by rewriting the previous paragraph and omitting *dependent variable*, *significantly*, and changing *independent variable* to *variables* (plural). Factor analysis does not begin with a dependent variable. It starts with a measure of the total amount of variation observed (similar to total sum of squares) in all variables that have been designated for factor analysis. Please note that this "variation" is a bit difficult to understand conceptually (Whence all this variance floating about? How does one see it? smell it? grasp it? but I digress . . .), but it is quite precise mathematically. The first step in factor analysis is for the computer to select the combination of variables whose shared correlations explain the greatest amount of the total variance. This is called Factor 1 (or, *Component 1*—the words are used interchangeably). Factor analysis will then extract a second factor. This is the combination of variables that explains the greatest amount of the variance that remains, that is, variation *after* the first factor has been extracted. This is called Factor 2 (or Component 2). This procedure continues for a third factor, fourth factor, fifth factor, and so on, until as many factors have been extracted as there are variables.

In the default SPSS procedure, each of the variables is initially assigned a *communality* value of 1.0. Communalities are designed to show the proportion of variance that the factors contribute to explaining a particular variable. These values range from 0 to 1 and may be interpreted in a matter similar to a Multiple *R*, with 0 indicating that common factors explain none of the variance in a particular variable, and 1 indicating that all the variance in that variable is explained by the common factors. However, for the default procedure at the initial extraction phase, each variable is assigned a communality of 1.0.

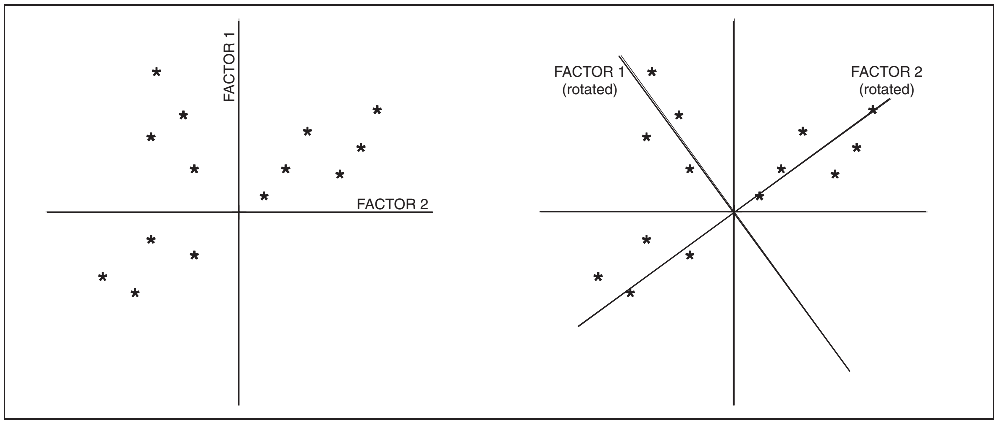
After the first factor is extracted, SPSS prints an eigenvalue to the right of the factor number (e.g., Factor number = 1; eigenvalue = 5.13). Eigenvalues are designed to show the proportion of variance accounted for by each factor (*not* each *variable* as do communalities). The first eigenvalue will always be largest (and always be greater than 1.0) because the first factor (by the definition of the procedure) always explains the greatest amount of total variance. It then lists the percent of the variance accounted for by this factor (the eigenvalue divided by the total number of variables), and this is followed by a cumulative percent. For each successive factor, the eigenvalue printed will be smaller than the previous one, and the cumulative percent (of variance explained) will total 100% after the final factor has been calculated. The absence of the word *significantly* is demonstrated by the fact that the **Factor** command extracts as many factors as there are variables, regardless of whether or not subsequent factors explain a significant amount of additional variance.

## 20.3 Factor Selection and Rotation

The factors extracted by SPSS are almost never *all* of interest to the researcher. If you have as many factors as there are variables, you have not accomplished what factor analysis was created to do. The goal is to explain the phenomena of interest with fewer than the original number of variables, usually substantially fewer. Remember Cattell? He started with 4,500 descriptors and ended up with 16 traits.

The first step is to decide *which* factors you wish to retain in the analysis. The common-sense criterion for retaining factors is that each retained factor must have some sort of face validity or theoretical validity; but prior to the rotation process, it is often impossible to interpret what each factor means. Therefore, the researcher usually selects a mathematical criterion for determining which factors to retain. The SPSS default is to keep any factor with an eigenvalue larger than 1.0. If a factor has an eigenvalue less than 1.0 it explains less variance than an original variable and is usually rejected. (Remember, SPSS will print out as many factors as there are variables, and usually for only a few of the factors will the eigenvalue be larger than 1.0.) There are other criteria for selection (such as the scree plot) or conceptual reasons (based on your knowledge of the data) that may be used. The procedure for selecting a number other than the default will be described in the Step by Step section.

Once factors have been selected, the next step is to rotate them. Rotation is needed because the original factor structure is mathematically correct but is difficult to interpret. The goal of rotation is to achieve what is called *simple structure*, that is, high *factor loadings* on one factor and low loadings on all others. Factor loadings vary between ±1.0 and indicate the strength of relationship between a particular variable and a particular factor, in a way similar to a correlation. For instance, the phrase "enjoys loud parties" might have a high loading on an "outgoing" factor (perhaps > .6) and have a low loading on an "intelligence" factor (perhaps < .1). This is because an enjoys-loud-parties statement is thought to be related to outgoingness, but unrelated to intelligence. Ideally, simple structure would have variables load entirely on one factor and not at all on the others. On the second graph (below), this would be represented by all asterisks being *on* the rotated factor lines. In social science research, however, this never happens, and the goal is to rotate the axes to have data points as close as possible to the rotated axes. The following graphs are good representations of how an unrotated structure and a rotated structure might look. SPSS will print out graphs of your factor structure (after rotation) and include a table of coordinates for additional clarity.



Rotation does not alter the mathematical accuracy of the factor structure, just as looking at a picture from the front rather than from the side does not alter the picture, and changing the measure of height in inches to height in centimeters does not alter how tall a person is. Rotation was originally done by hand, and the researcher would position the axes in the location that appeared to create the optimal factor structure. Hand rotation is not possible with SPSS, but there are several mathematical procedures available to rotate the axes to the best simple structure. **Varimax** is the default procedure used by SPSS, but there are several others (mentioned in the Step by Step section).

**OBLIQUE ROTATIONS** Varimax rotations are called *orthogonal rotations* because the axes that are rotated remain at right angles to each other. Sometimes it is possible to achieve a better simple structure by diverging from perpendicular. The **Direct Oblimin** and **Promax** procedures allow the researcher to deviate from orthogonal to achieve a better simple structure. Conceptually, this deviation means that factors are no longer uncorrelated with each other. This is not necessarily disturbing, because few factors in the social sciences *are* entirely uncorrelated. The use of oblique rotations can be quite tricky, and (here we insert our standard disclaimer) you should not attempt to use them unless you have a clear understanding of what you are doing. Let's expand: You should probably not attempt to conduct factor analysis *at all* unless you have had a course in it and/or have a clear conceptual understanding of the procedure. The technique for specifying **Direct Oblimin** or **Promax** rotation is described in the Step by Step section. We do not demonstrate this procedure in the Output section because it requires more attention than we can accord it here.

## 20.4 Interpretation

In an ideal world, each of the original variables will load highly (e.g., > .5) on one of the factors and low (e.g., < .2) on all others. Furthermore, the factors that have the high loadings will have excellent face validity and appear to be measuring some underlying construct. In the real world, this rarely happens. There will often be two or three irritating variables that end up loading on the "wrong" factor, and often a variable will load onto two or three different factors. The output of factor analysis requires considerable understanding of your data, and it is rare for the arithmetic of factor analysis alone to produce entirely clear results. In the Output section we will clarify with a real example. The example will be introduced in the following paragraphs.

We draw our example from real data; it is the same set of data used to demonstrate reliability analysis (Chapter 18), the **helping2.sav** file. It has *N* of 517 and measures many of the same variables as the **helping1.sav** and **helping3.sav** files used in earlier chapters. In the **helping2.sav** file, self-efficacy (belief that one has the ability to help effectively) was measured by 15 questions, each paired with an amount-of-help question that measured a particular type of helping. An example of one of the paired questions follows:

**9a) Time spent expressing sympathy, empathy, or understanding.**

| | | | | | | |
|---|---|---|---|---|---|---|
| none | 0–15<br>minutes | 15–30<br>minutes | 30–60<br>minutes | 1–2<br>hours | 2–5<br>hours | ____<br>hours |

**9b) Did you believe you were capable of expressing sympathy, empathy, or understanding to your friend?**

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| not at all | | | some | | very much so | |

There were three categories of help represented in the 15 questions: 6 questions were intended to measure empathic types of helping, 4 questions were intended to measure informational types of helping, 4 questions were intended to measure instrumental ("doing things") types of helping, and the 15th question was open-ended to allow any additional type of help given to be inserted. Factor analysis was conducted on the 15 self-efficacy questions to see if the results would yield the three categories of self-efficacy that were originally intended.

A final word: In the Step by Step section, there will be two Step-5 variations. Step 5 will be the simplest default factor analysis that is possible. Step 5a will be the sequence of steps that includes many of the variations that we present in this chapter. Both will conduct a factor analysis on the 15 questions that measure self-efficacy in the **helping2.sav** file.

For additional information on factor analysis, the best textbooks on the topic known by the authors are by Comrey and Lee (1992) and Gorsuch (1983). Please see the reference section for additional information about these books.

## 20.5 Step by Step

### 20.5.1 Factor Analysis

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | → All Programs → IBM SPSS Statistics → IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | → → | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| | Step 2 |
|---|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* | |

*Screens 1 and 2 (on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **helping2.sav** *file for further analyses:*

| In Screen | Do This | Step 3 |
|---|---|---|
| Front1 | File → Open → Data [*or* ] | |
| Front2 | type `helping2.sav` → Open [*or* helping2.sav] | Data |

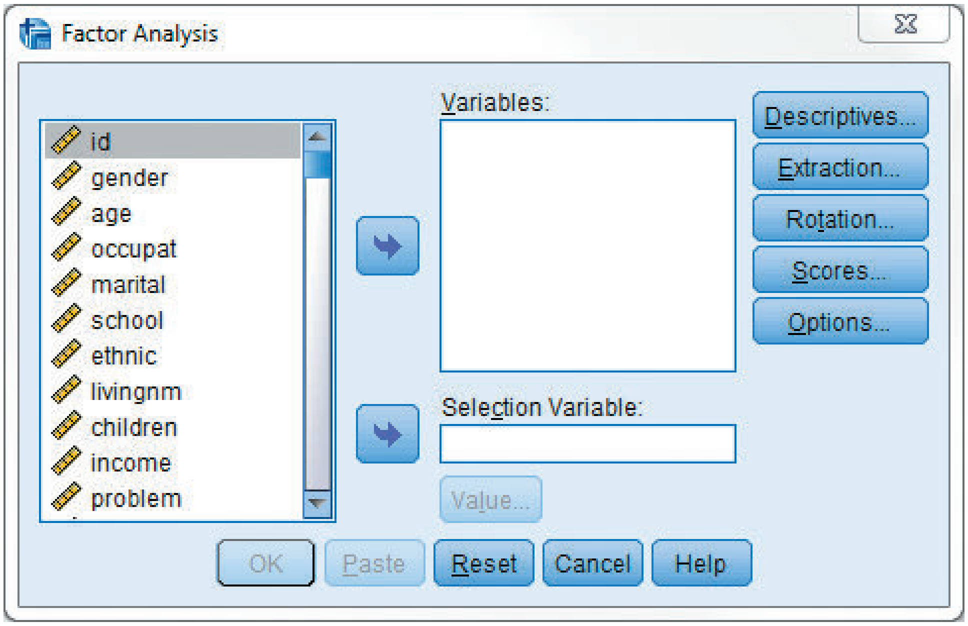
Whether first entering SPSS or returning from earlier operations the standard menu of commands across the top is required. As long as it is visible you may perform any analyses. It is not necessary for the data window to be visible.

After completion of Step 3 a screen with the desired menu bar appears. When you click a command (from the menu bar), a series of options will appear (usually) below the selected command. With each new set of options, click the desired item. The sequence to access Factor Analysis begins at any screen with the menu of commands visible:

| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | Analyze → Dimension Reduction → Factor | 20.1 |

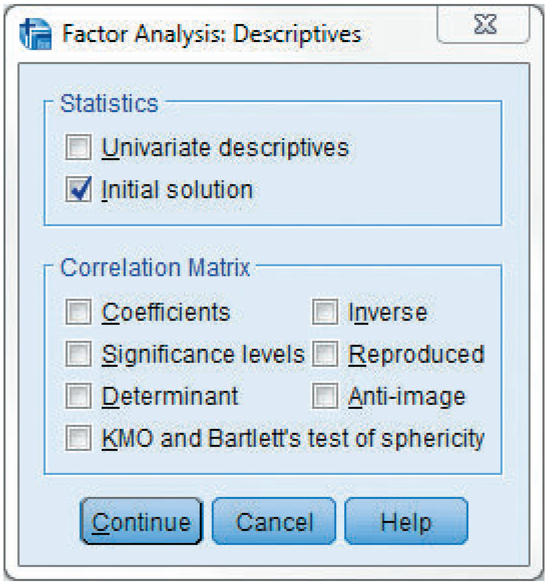
The click on **Factor** opens the main dialog window for factor analysis (Screen 20.1, following page). The initial screen looks innocent enough: the box of available variables to the left, a single active (**Variables**) box to the right, and five push buttons representing different options.

**Screen 20.1** The Factor Analysis Opening Dialog Window



Although the arithmetic of factor analysis is certainly complex, a factor analysis may be conducted by simply pasting some variables into the active box, selecting **Varimax** rotations, and clicking **OK**. Indeed, *any* factor analysis will begin with the pasting of carefully selected variables into the **Variables** box. For the present illustration we will select the 15 self-efficacy questions (**effic1** to **effic15**) for analysis. The efficacy questions are displayed in the variable list box in Screen 20.1 above. The five pushbuttons (to the right) represent many available options and may be processed in any order. We begin with **Descriptives**.

**Screen 20.2** The Factor Analysis: Descriptives Window


Factor Analysis: Descriptives

Statistics

- □ Univariate descriptives
- ☑ Initial solution

Correlation Matrix

- □ Coefficients
- □ Inverse
- □ Significance levels
- □ Reproduced
- □ Determinant
- □ Anti-image
- □ KMO and Bartlett's test of sphericity

This option opens up Screen 20.2, shown above. For this and other windows we will describe only those selections that are most frequently used by researchers. The **Univariate descriptives** option is quite useful. It lists in four neat columns the variable names, the means, the standard deviations, and the always handy variable labels. This chart will be referred to frequently during the course of analysis. The **Initial solution** is selected by default and lists the variable names, the initial communalities (1.0 by default), the factors, the eigenvalues, and the percent and cumulative percent accounted for by each factor. The correlation matrix is the starting point of any factor analysis. We describe four of the frequently used statistics related to the correlation matrix:

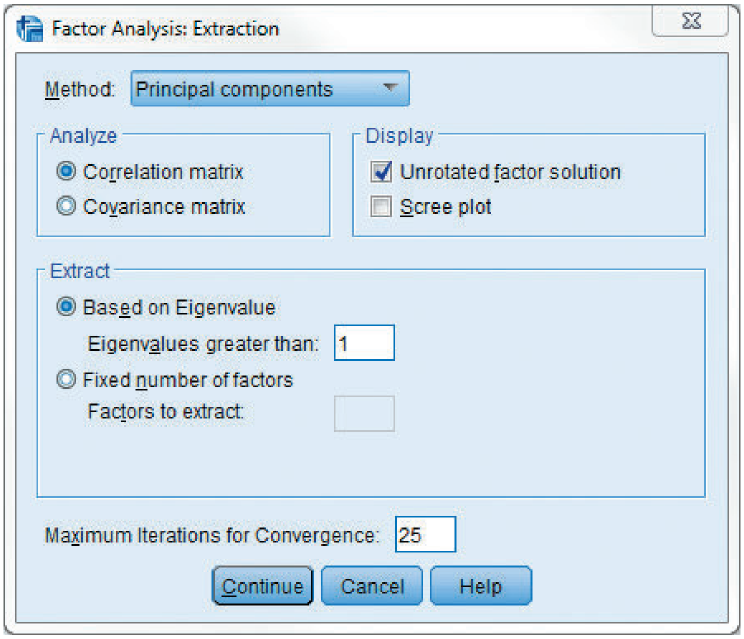
- **Coefficients**: This is simply the correlation matrix of included variables.
- **Significance levels:** These are the *p* values associated with each correlation.
- **Determinant:** This is the determinant of the correlation matrix. It is used in computing values for tests of multivariate normality.
- **KMO and Bartlett's test of sphericity**: The KMO test and Bartlett's test of sphericity are both tests of multivariate normality and sampling adequacy (the adequacy of your variables for conducting factor analysis). This test is selected by default. This option is discussed in greater detail in the Output section.

A click on the **Extraction** button opens an new dialog box (Screen 20.3, below) that deals with the method of extraction, the criteria for factor selection, the display of output related to factor extraction, and specification of the number of iterations for the procedure to converge to a solution. The **Method** of factor extraction includes seven options. **Principal components** is the default procedure; a click of the ▼ reveals the six others:

- **Unweighted least squares**
- **Generalized least squares**
- **Maximum likelihood**
- **Principal-axis factoring**
- **Alpha factoring**
- **Image factoring**

The principal components method is, however, the most frequently used (with maximum likelihood the next most common), and space limitations prohibit discussion of the other options.

Screen 20.3 The Factor Analysis: Extraction Window



Factor Analysis: Extraction

Method: Principal components

Analyze
● Correlation matrix
○ Covariance matrix

Display
☑ Unrotated factor solution
□ Scree plot

Extract
● Based on Eigenvalue
Eigenvalues greater than: 1
○ Fixed number of factors
Factors to extract:

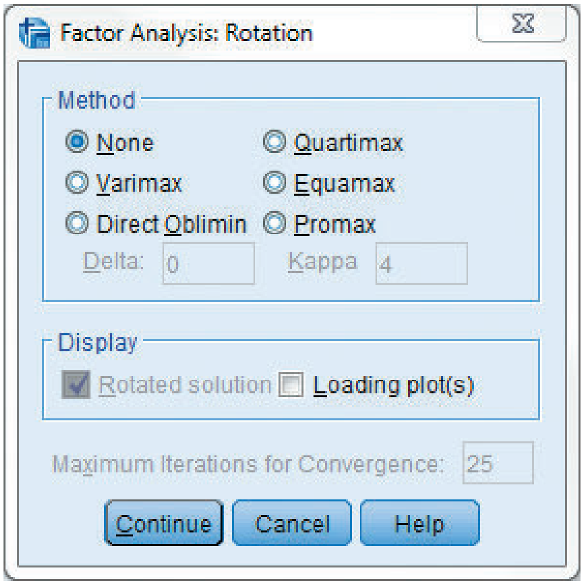
Maximum Iterations for Convergence: 25

The **Analyze** box allows you to select either the **Correlation matrix** or the **Covariance matrix** as the starting point for your analysis. Under **Extract**, selection of factors for rotation is based on the eigenvalue being greater than 1 (the default), selection of a different eigenvalue, or simple identification of how many factors you wish to select for rotation.

Under **Display**, the **Unrotated factor solution** is selected by default, but unless you are quite mathematically sophisticated, the unrotated factor solution rarely reveals much. Most researchers would choose to deselect this option. You may also request a **Scree plot** (illustrated in the Output section). Finally you may designate the number of iterations you wish for convergence. The default of 25 is almost always sufficient.

A click on **Rotation** moves on to the next step in factor analysis, rotating the factors to final solution. Screen 20.4 displays available options. There are three different *orthogonal* methods for rotation, **Varimax** (the most popular method, yes, but using it requires weathering the disdain of the factor analytic elite), **Equamax**, and **Quartimax**. The **Direct Oblimin** and **Promax** procedures allow for a nonorthogonal rotation of selected factors. Both oblique procedures' parameters ($\delta$ and $k$) can usually be left at the default values. As suggested in the introduction, don't even think of attempting oblique rotations without a solid course in factor analysis.

**Screen 20.4** The Factor Analysis: Rotation Window



Factor Analysis: Rotation

Method

● None ○ Quartimax
○ Varimax ○ Equamax
○ Direct Oblimin ○ Promax
Delta: 0 Kappa 4

Display

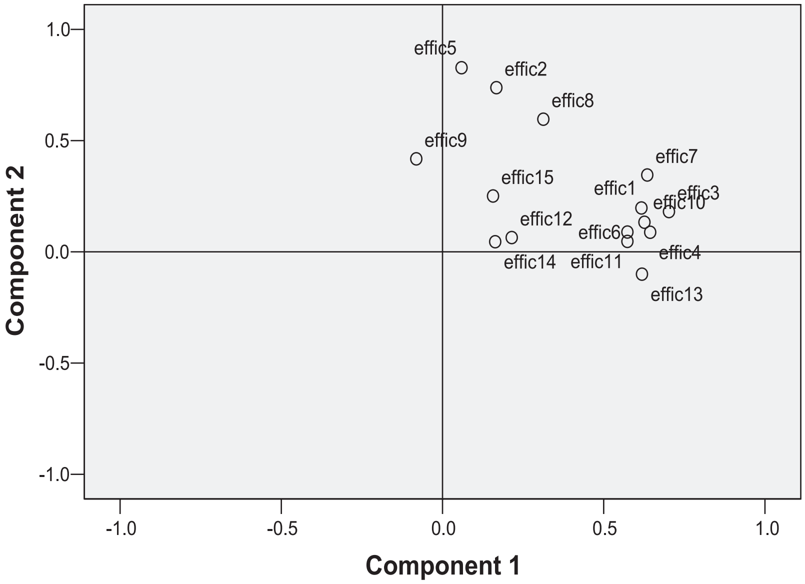
☑ Rotated solution □ Loading plot(s)

Maximum Iterations for Convergence: 25

For display, the **Rotated solution** is selected by default and represents the essence of what factor analysis is designed to do. The Output section includes several paragraphs about interpretation of a rotated solution. If you wish a visual display of factor structure after rotation, click on the **Loading plot(s)** option. The plot(s) option provides by default a 3-D graph of the first three factors. The chart is clever; you can drop lines to the floor and view it from a variety of perspectives, but no matter what manipulations you enact, one constant remains: the graph's meaning is almost unintelligible. Two-dimensional graphs are much easier to interpret. SPSS Factor Analysis options do not include this possibility. If, however, you wish to convert the 3-D graph to a much simpler 2-D graph, just open the chart by double-clicking on it. Once the chart-subject-to-editing appears, right click on the chart and a dialog box will open called **3-D Rotation**. Then rotate the third (Z) axis out of the picture. While editing for further clarity, be patient as you "battle with the beast."

These 2-D charts are quite interpretable, particularly if you refer to the values listed in the rotated component (or factor) matrix. Notice that on the chart (next page), the horizontal axis is Factor 1 (efficacy for empathic types of helping) and the vertical axis is Factor 2 (efficacy for informational types of helping). Note also that the eight data points at the extreme right feature high loadings (.6 to .8) on the efficacy for

**Screen 20.5** Factor Plot of Two Rotated Factors, Efficacy for Empathic Help (Component 1) with Efficacy for Informational Help (Component 2)

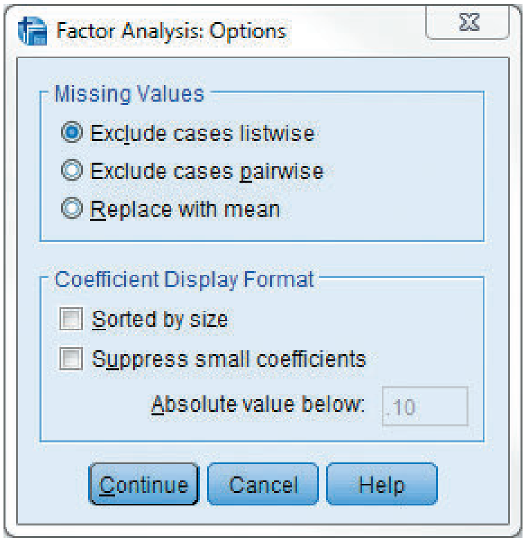


empathic-helping (Factor 1) and low loadings (−.1 to .4) on efficacy for informational helping (Factor 2).

Back to the main dialog window: A click on the **Scores** pushbutton opens a small dialog box that allows you to save certain scores as variables. This window is not shown but the **Display factor score coefficient matrix** (when selected) will display the component score coefficient matrix in the output.

Finally, a click on the **Options** button opens a dialog box (Screen 20.6, below) that allows two different options concerning the display of the rotated factor matrix. The **Sorted by size** selection is very handy. It sorts variables by the magnitude of their factor loadings by factor. Thus, if six variables load onto Factor 1, the factor loadings for those six variables will be listed from the largest to the smallest in the column titled Factor 1. The same will be true for variables loading highly on the second factor, the third factor, and so forth. The Output section demonstrates this feature. Then you can suppress factor loadings that are less than a particular value (default is .10) if you feel that such loadings are insignificant. Actually, you can do it no matter how you feel. Anyone able to conduct factor analysis has already dealt with **Missing Values** and would be insulted by the implication that she or he would resort to an automatic procedure at this stage of the process.

**Screen 20.6** The Factor Analysis: Options Window



Factor Analysis: Options

Missing Values
- ● Exclude cases listwise
- ○ Exclude cases pairwise
- ○ Replace with mean

Coefficient Display Format
- ☐ Sorted by size
- ☐ Suppress small coefficients

Absolute value below: .10

What follow are the two step-by-step sequences. The first (Step 5) is the simplest default factor analysis. The second (sequence Step 5a) includes many of the options described in the previous pages. The italicized text before each box will identify features of the procedure that follows.

*To conduct a factor analysis with the 15 efficacy items (***effic1** *to* **effic15***) with all the default options (plus the* **Varimax** *method of rotation), perform the following sequence. The starting point is Screen 20.1. Perform the four-step sequence (page 258) to arrive at this screen.*

| In Screen | Do This | Step 5 |
|---|---|---|
| 20.1 | effic1 → press Shift *and* effic15 → ➡ → Rotation | |
| 20.4 | Varimax → Continue | |
| 20.1 | OK | Back1 |

These five clicks perform a factor analysis that

1. calculates a correlation matrix of the 15 efficacy questions from the data file,
2. extracts 15 factors by the principal-components method,
3. selects as factors to be rotated all factors that have an eigenvalue greater than 1.0,
4. rotates the selected factors to a Varimax solution, and
5. prints out the factor transformation matrix.

*To conduct a factor analysis with the 15 efficacy items (***effic1** *to* **effic15***), request* **Univariate descriptives** *for the 15 variables, correlation* **Coefficients***, and the standard measures of multivariate normality (***KMO and Bartlett's test of sphiricity***); request the principal components method of extraction (default) and the* **Scree plot***; access the* **Varimax** *method of rotation and the available* **Loading plots** *of factor scores; and request that factor loadings be sorted by factor number and* **Sorted by size***, perform the following sequence of steps. The starting point is Screen 20.1. Carry out the four-step sequence (page 258) to arrive at this screen.*

| In Screen | Do This | Step 5a |
|---|---|---|
| 20.1 | effic1 → press Shift *and* effic15 → ➡ → Descriptives | |
| 20.2 | Univariate descriptives → Coefficients → KMO & Bartlett's test of sphiricity → Continue | |
| 20.1 | Extraction | |
| 20.3 | Scree plot → Continue | |
| 20.1 | Rotation | |
| 20.4 | Varimax → Loading Plots → Continue | |
| 20.1 | Options | |
| 20.6 | Sorted by size → Continue | |
| 20.1 | OK | Back1 |

# 20.6 Output

## 20.6.1 Factor Analysis

What follows is output from sequence Step 5a (previous page). The output presented below is rendered in a fairly simple, intuitive format, with variable names, not labels. See pages 24–25 (for PC) or 41–42 (for Mac) for switching between using variable names or labels. As in most of the advanced procedures, output is condensed and slightly reformatted.

**KMO and Bartlett's Test**

| Test | Test Statistic | df | Significance |
|---|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy | .871 | | |
| Bartlett's Test of Sphericity | 1321.696 | 105 | .000 |

| Term | Definition/Description |
|---|---|
| **KAISER-MAYER-OLKIN** | A measure of whether your distribution of values is adequate for conducting factor analysis. Kaiser himself (enjoying the letter m) designates levels as follows: A measure > .9 is marvelous, > .8 is meritorious, > .7 is middling, > .6 is mediocre, > .5 is miserable, and < .5 is unacceptable. In this case .871 is meritorious, almost marvelous. |
| **BARTLETT TEST OF SPHERICITY** | This is a measure of the multivariate normality of your set of distributions. It also tests whether the correlation matrix is an identity matrix (factor analysis would be meaningless with an identity matrix). A significance value < .05 indicates that that these data do NOT produce an identity matrix (or "differ significantly from identity") and are thus approximately multivariate normal and acceptable for factor analysis. |

**Communalities** **Total Variance Explained**

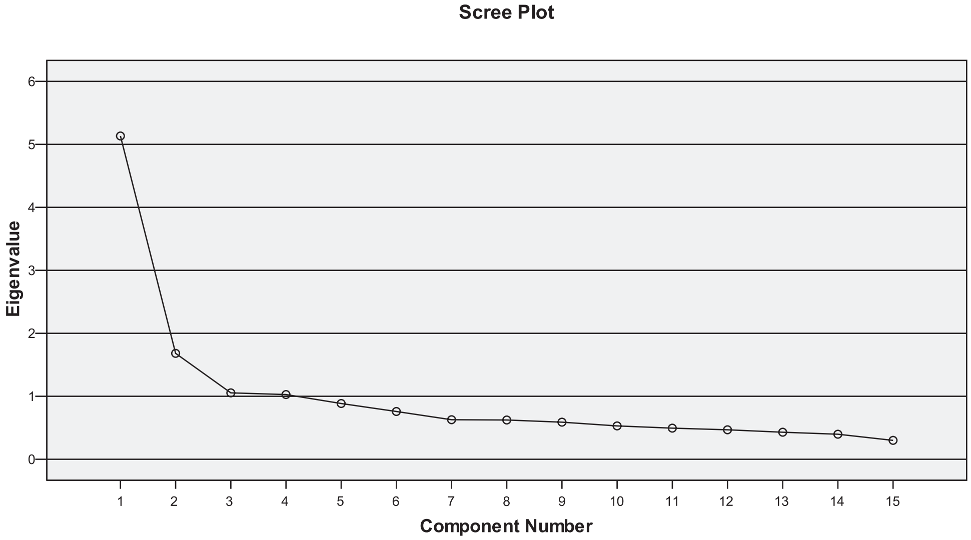
| | | | Initial Eigenvalues | | |
|---|---|---|---|---|---|
| Variable | Initial Communality | Component | Total | % of Variance | Cumulative % |
| effic1 | 1.00000 | 1 | 5.133 | 34.2 | 34.2 |
| effic2 | 1.00000 | 2 | 1.682 | 11.2 | 45.4 |
| effic3 | 1.00000 | 3 | 1.055 | 7.0 | 52.5 |
| effic4 | 1.00000 | 4 | 1.028 | 6.9 | 59.3 |
| effic5 | 1.00000 | 5 | .885 | 5.9 | 65.2 |
| effic6 | 1.00000 | 6 | .759 | 5.1 | 70.3 |
| effic7 | 1.00000 | 7 | .628 | 4.2 | 74.5 |
| effic8 | 1.00000 | 8 | .624 | 4.2 | 78.6 |
| effic9 | 1.00000 | 9 | .589 | 3.9 | 82.5 |
| effic10 | 1.00000 | 10 | .530 | 3.5 | 86.1 |
| effic11 | 1.00000 | 11 | .494 | 3.3 | 89.4 |
| effic12 | 1.00000 | 12 | .458 | 3.1 | 92.5 |
| effic13 | 1.00000 | 13 | .429 | 2.9 | 95.3 |
| effic14 | 1.00000 | 14 | .398 | 2.7 | 98.0 |
| effic15 | 1.00000 | 15 | .300 | 2.0 | 100.0 |

Extraction method: Principle Component Analysis.

The two columns to the left refer only to the variables and the communalities. The four columns to the right refer to the components or factors. Note that there are four factors with eigenvalues larger than 1.0 and they account for almost 60% of the total variance. The definitions below clarify others items shown in this output.

| Term | Definition/Description |
|---|---|
| **PRINCIPAL-COMPONENT ANALYSIS** | The default method of factor extraction used by SPSS. Other options are available and are listed on page 260. |
| **VARIABLE** | All 15 efficacy variables used in the factor analysis are listed here. |
| **COMMUNALITY** | The default procedure assigns each variable a communality of 1.00. Different communalities may be requested. See the SPSS manuals for details. |

| Term | Definition/Description |
|---|---|
| COMPONENT | The number of each component (or factor) extracted. Note that the first two columns provide information about the *variables*, and the last four provide information about the *factors*. |
| EIGENVALUE | The proportion of variance explained by each factor. |
| % OF VARIANCE | The percent of variance explained by each factor, the eigenvalue divided by the sum of the communalities (15 in this case). |
| CUMULATIVE % | The sum of each step in the previous column. |



This is called a **Scree plot**. It plots the eigenvalues on a bicoordinate plane. It derives its name from the *scree* that is deposited at the base of a landslide. The scree plot is sometimes used to select how many factors to rotate to a final solution. The traditional construct for interpretation is that the scree should be ignored and that only factors on the steep portion of the graph should be selected and rotated. The SPSS default is to select and rotate any factor with an eigenvalue greater than 1.0. Since the default procedure is followed in this example, four (4) factors are selected for rotation; based on the scree, we might instead select two or three factors.

SPSS will next print out the *un*rotated component structure. This is rarely of interest to the researcher and, to save space, is not included here. What follows is the 4 × 4 factor transformation matrix. If you multiply the factor transformation matrix (below) by the original (unrotated) 4 × 15 factor matrix, the result would be the *rotated* factor matrix.

**Component (Factor) Transformation Matrix**

| Component | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | .733 | .405 | .419 | .352 |
| 2 | −.441 | .736 | −.195 | .420 |
| 3 | −.113 | −.538 | .017 | .835 |
| 4 | −.506 | .066 | .859 | −.043 |

The *rotated* factor structure is displayed next. Note that due to the selection of the **Sorted by size** option, the factor loadings are sorted in two ways: (a) The highest factor loadings for each factor are selected and listed in separate blocks and (b) within each block, the factor loadings are sorted from largest to smallest. The numbers in each column are the factor loadings for each factor, roughly the equivalent of the correlation between a particular item and the factor.

To aid in the interpretation of the rotated factor matrix, next to each efficacy variable (**effic1** to **effic15**) we listed the three categories of efficacy the questions were intended to test:

| | |
|---|---|
| [Emot] | Efficacy for emotional types of helping |
| [Inf] | Efficacy for informational types of helping |
| [Instr] | Efficacy for instrumental types of helping |
| [----] | The open-ended question |

**Rotated Component Matrix**

| Variable | Component 1 | Component 2 | Component 3 | Component 4 |
|---|---|---|---|---|
| effic3 [Instr] | .702 | .181 | .022 | .153 |
| effic4 [Emot] | .644 | .088 | .302 | −.044 |
| effic7 [Emot] | .635 | .345 | −.090 | .205 |
| effic10 [Emot] | .626 | .133 | .472 | −.186 |
| effic13 [Emot] | .618 | −.100 | .506 | .071 |
| effic1 [Emot] | .617 | .197 | .162 | .295 |
| effic11 [Inf] | .573 | .047 | .274 | .125 |
| effic5 [Inf] | .059 | .828 | .081 | .128 |
| effic2 [Inf] | .167 | .738 | .024 | .234 |
| effic8 [Inf] | .313 | .597 | .226 | −.057 |
| effic14 [Emot] | .163 | .045 | .779 | .254 |
| effic15 [----] | .157 | .251 | .624 | .113 |
| effic12 [Instr] | .215 | .064 | .192 | .738 |
| effic9 [Instr] | .081 | .418 | .180 | .631 |
| effic6 [Instr] | .574 | .089 | −.009 | .604 |

The initial reaction of the researcher who conducted this analysis was "pretty good factor structure!" The first factor is composed primarily of variables that measure efficacy for *emotional* types of helping. One question from instrumental helping ("To what extent did you have the ability to listen carefully or to appraise your friend's situation?") and one from informational helping ("Did you believe you were capable of reducing tension and helping your friend get his/her mind off the problem?") were included in the first factor. It is not difficult to see why these two items might load onto the same factor as efficacy for emotional helping.

Factor 2 is composed entirely of the remaining three measures of efficacy for *informational* types of giving. Factor 4 is composed entirely of the remaining three measures of efficacy for *instrumental* types of helping. Factor 3 is a rather strange factor and would likely not be used. Included is the measure of efficacy for *other* types of helping—a measure that the majority of subjects did not answer. The other variable, **effic14**, is a somewhat strange measure that a number of subjects seemed confused about. It dealt with efficacy for helping the friend reduce self-blame. In many problem situations self-blame was not an issue.

This is the type of thinking a researcher does when attempting to interpret the results from a factor analysis. The present output seems to yield a fairly interpretable pattern of three types of efficacy: efficacy for emotional types of helping, efficacy for instrumental types of helping, and efficacy for informational types of helping. Factor 3, the strange one, would likely be dropped. Because the 2 variables in Factor 3 also have factor loadings on the other 3 factors, the researcher may omit these 2 variables and run the factor analysis again with just the 13 variables, to see if results differ.

# Chapter 21
# Cluster Analysis

DESCRIPTION OF any procedure is often most effectively accomplished by comparison to another procedure. To clarify an understanding of hierarchical cluster analysis, we will compare it to factor analysis. Cluster analysis is in some ways similar to factor analysis (Chapter 20), but it also differs in several important ways. If you are unfamiliar with factor analysis, please read through the introduction of Chapter 20 before attempting this chapter. The introductory paragraphs of this section will compare and contrast cluster analysis with factor analysis. Following this, the sequential steps of the cluster analysis procedure will be identified, along with a description of the data set used as an example to aid in the understanding of each step of the process.

## 21.1 Cluster Analysis and Factor Analysis Contrasted

Cluster analysis and factor analysis are similar in that both procedures take a larger number of cases or variables and reduce them to a smaller number of factors (or clusters) based on some sort of similarity that members within a group share. However, the statistical procedure underlying each type of analysis and the ways that output is interpreted are often quite different:

1. Factor analysis is used to reduce a larger number of *variables* to a smaller number of factors that describe these variables. Cluster analysis is more typically used to combine cases into groups. More specifically, cluster analysis is primarily designed to use variables as criteria for grouping (agglomerating) *subjects* or *cases* (not variables) into groups, based on each subject's (or case's) scores on a given set of variables. For instance: With a data set of 500 subjects measured on 15 different types of helping behavior (the variables), factor analysis might be used to extract factors that describe 3 or 4 categories or types of helping based on the 15 help variables. Cluster analysis is more likely to be used with a data set that includes, for instance, 21 cases (such as brands of DVD players) with perhaps 15 variables that identify characteristics of each of the 21 DVD players. Rather than the *variables* being clustered (as in factor analysis), the cases (DVD players here) would be clustered into groups that shared similar features with each other. The result might be three or four clusters of brands of DVD players that share similarities such as price, quality of picture, extra features, and reliability of the product.
2. Although cluster analysis is typically used for grouping cases, clustering of variables (rather than subjects or cases) is just as easy. Since cluster analysis is still more frequently used to cluster cases than variables, clustering cases will be demonstrated in the present example. However, in the Step by Step section we will illustrate the sequence of steps for clustering variables.
3. The statistical procedures involved are radically different for cluster analysis and factor analysis. Factor analysis analyzes all variables at each factor extraction step to calculate the variance that each variable contributes to that factor. Cluster

analysis calculates a *similarity* or a *distance* measure between each subject or case and every other subject or case and then it groups the two subjects/cases that have the greatest similarity or the least distance into a cluster of two. Then it computes the similarity or distance measures all over again and either combines the next two subjects/cases that are closest or (if the distance is smaller) combines the next case with the cluster of two already formed (yielding either two clusters of two cases each or one cluster of three cases). This process continues until all cases are grouped into one large cluster containing all cases. The researcher decides at which stage to stop the clustering process.

4. Reflecting the sentiments of an earlier paragraph, the **Hierarchical Cluster** command is more likely to be used in business, sociology, or political science (e.g., categories of the 25 best-selling brands of TVs; classification of 40 communities based on several demographic variables; groupings of the 30 most populous cities based on age, SES, and political affiliation) than for use in psychology. Psychologists are more often trying to find similarities between variables or underlying causes for phenomena than trying to divide their subjects into groups via some sort of mathematical construct; and when psychologists *are* trying to divide subjects into groups, discriminant analysis (Chapter 22) often accomplishes the process more efficiently than does cluster analysis. But these are only generalities, and there are many appropriate exceptions.
5. With the present ease of using cluster analysis to cluster variables into groups, it becomes immediately interesting to compare results of cluster analysis with factor analysis using the same data set. As with factor analysis, there are a number of variations to the way that the key components of cluster analysis (principally how distances/similarities are determined and how individual variables are clustered) may be accomplished. Since with both factor analysis and cluster analysis results can often be altered significantly by how one runs the procedure, researchers may shamelessly try them all and use whatever best supports their case. But, one should ideally use the process that best represents the nature of the data, considering carefully the rationale of whether a shared-variance construct (factor analysis) or a similarities-of-features procedure (cluster analysis) is better suited for a particular purpose.

## 21.2 Procedures for Conducting Cluster Analysis

Cluster analysis goes through a sequence of procedures to come up with a final solution. To assist understanding, before we describe these procedures we will introduce the example constructed for this analysis and use it to clarify each step. Cluster analysis would not be appropriate for either the **grades.sav** file or any of the helping files. In each data set the focus is on the nature of the variables, and there is little rationale for wanting to cluster subjects. We turn instead to a data set that is better designed for cluster analysis. This is the only chapter in which these data will be used. It consists of a modified analysis from the magazine *Consumer Reports* on the quality of the top 21 brands of DVD players. The file associated with this data set is called **dvd.sav**. While this began as factual information, we have "doctored" the data to help create a clear cluster structure and thus serve as a convincing example. Because of this, the brand names in the data set are also fictional.

**STEP 1** Select variables to be used as criteria for cluster formation. In the **dvd.sav** file, cluster analysis will be conducted based on the following variables (listed in the order shown in the first Output display): Price, picture quality (five measures), reception quality (three measures), audio quality (three measures), ease of programming (one measure), number of events (one measure), number of days for future programming (one measure), remote control features (three measures), and extras (three measures).

**STEP 2** Select the procedure for measuring distance or similarity between each case or cluster (initially, each case is a cluster of one, the 21 brands of DVD players). The SPSS default for this measure is called the *squared Euclidean distance*, and it is simply the sum of the squared differences for each variable for each case. For instance, Brand A may have scores of 2, 3, and 5 for the three audio ratings. Brand B may have ratings of 4, 3, and 2 for the same three measures. Squared Euclidean distance between these two brands would be $(2 - 4)^2 + (3 - 3)^2 + (5 - 2)^2 = 13$. In the actual analysis, these squared differences would be summed for all 21 variables for each brand, yielding a numeric measure of the distance between each pair of brands. SPSS provides for the use of measures other than squared Euclidean distance to determine distance or similarity between clusters. These are described in the SPSS manuals.

A question that may come to mind is, will the cluster procedure be valid if the scales of measurement for the variables are different? In the **dvd.sav** file, most measures are rated on a 5-point scale from *much poorer than average* (1) to *much better than average* (5). But the listed prices fluctuate between $200 and $525, and events, days, and extras are simply the actual numbers associated with those variables. The solution suggested by SPSS is to standardize all variables—that is, change each variable to a *z* score (with a mean of 0 and a standard deviation of 1). There are other standardization options but the *z* score is so widely used that it has the advantage of familiarity. This will give each variable equal metrics, and will give them equal weight as well. If all variables are already in the same metric, or if you wish to retain the weighting of the original values, then it is not necessary to change to *z* scores.

**STEP 3** Form clusters. There are two basic methods for forming clusters, agglomerative and divisive. For agglomerative hierarchical clustering, SPSS groups cases into progressively larger clusters until all cases are in one large cluster. In divisive hierarchical clustering, the opposite procedure takes place. All cases are grouped into one large cluster, and clusters are split at each step until the desired number of clusters is achieved. The agglomerative procedure is the SPSS default and will be presented here.

Within the agglomerative framework, there are several options of mathematical procedures for combining clusters. The default procedure is called the *between-group linkage* or the *average linkage within groups*. SPSS computes the smallest average distance between all group pairs and combines the two groups that are closest. Note that in the initial phase (when all clusters are individual cases), the average distance is simply the computed distance between pairs of cases. Only when actual clusters are formed does the term *average* distance apply. The procedure begins with as many clusters as there are cases (21 in the **dvd.sav** file). At Step 1, the two cases with the smallest distance between them are clustered. Then SPSS computes distances once more and combines the two that are next closest. After the second step you will have either 18 individual cases and 1 cluster of 3 cases or 17 individual cases and 2 clusters of 2 cases each. This process continues until all cases are grouped into one large cluster. Other methods of clustering cases are described in the SPSS Manuals.

**STEP 4** Interpreting results. Similar to factor analysis, interpretation and the number of clusters to accept as a final solution are largely a matter of the interpretation of the researcher. In the **dvd.sav** file a three-cluster solution seemed best. There appeared to be three primary qualities that differentiated among groups: The first group was highest in price ($M$ = $511.67), highest in picture quality ($M = 5.00$), and had the greatest number of additional features ($M = 10.8$). The second group contained DVD players that were moderate in price, medium to low on picture quality, and had a smaller number of additional features ($400.00, 3.00, and 7.8, respectively). The third group contained the budget models ($262.22, 2.78, and 3.0). The other variables did not seem to contribute to the clustering in any systematic way.

The order followed for the description of cluster analysis in the Step by Step section will be similar to that followed in the chapters that present the more complex statistical procedures. Sequence Steps 1–4 and 6–7 will be identical to other chapters. There will be three versions of Step 5. The first (Step 5) will be the simplest default procedure for conducting cluster analysis. The second (Step 5a) will include a number of additional available options to tailor your program to fit your particular needs or to produce additional desired output. The last presentation (Step 5b) will show how to conduct cluster analysis of variables. To illustrate clustering variables we will employ the same variables used in the factor analysis chapter (Chapter 20), the 15 self-efficacy questions from the **helping2.sav** file. Although this procedure will be shown in the Step by Step section, space restraints prevent reproducing it in the Output section.

The best resource on cluster analysis known to the authors is by Brian S. Everitt and colleagues (2001). Please see the reference section for additional information about this book.

## 21.3 Step by Step

### 21.3.1 Cluster Analysis

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | 🖱 ⊞ → 🖱 All Programs ▷ → 🖱 IBM SPSS Statistics → 🖱 IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | 🖱 🚀 → 🖱 ■ → 🖱 Σ | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| | Step 2 |
|---|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* | |

*Screens 1 and 2 (on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **dvd.sav** *file for further analyses:*

| In Screen | Do This | Step 3 |
|---|---|---|
| Front1 | 🖱 File → Open → 🖱 Data [*or* 🖱 📂] | |
| Front2 | type dvd.sav → 🖱 Open [*or* 🖱 🖱 dvd.sav] | Data |

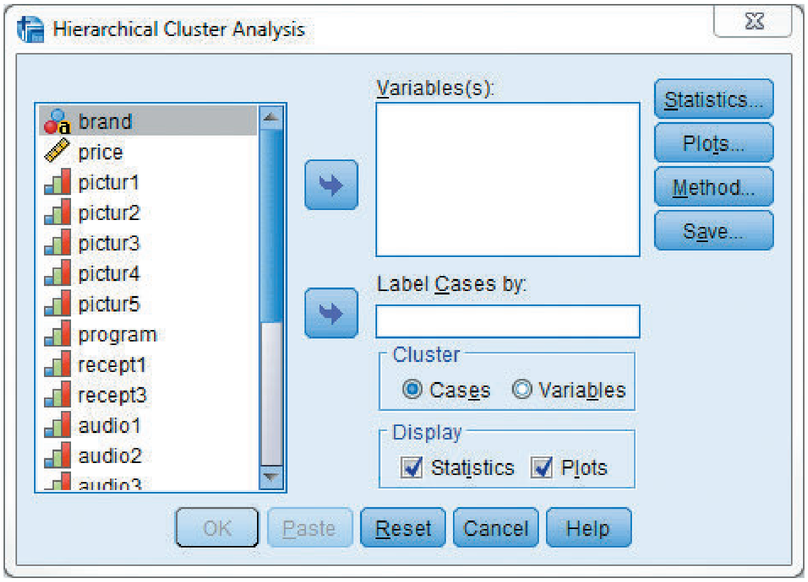
Whether first entering SPSS or returning from earlier operations the standard menu of commands across the top is required. As long as it is visible you may perform any analyses. It is not necessary for the data window to be visible.

*After completion of Step 3 a screen with the desired menu bar appears. When you click a command (from the menu bar), a series of options will appear (usually) below the selected command. With each new set of options, click the desired item. The sequence to access cluster analysis begins at any screen with the menu of commands visible:*

| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | **Analyze** → **Classify** → **Hierarchical Cluster** | 21.1 |

These three clicks open up the Hierarchical Cluster Analysis main dialog window (Screen 21.1, below). The use of this window will vary substantially depending on whether you choose to **Cluster Cases** or to **Cluster Variables**. If you cluster cases (as we do in our example in this chapter), then you will identify variables you wish to be considered in creating clusters for the cases (in our example, all variables except for **brand**). These will be pasted into the **Variable(s)** box. Then you need to specify how you wish your cases to be identified. This will usually be an ID number or some identifying name. In the example, the **brand** variable (it lists the 21 brand names) will be pasted into the lower box to identify our 21 cases.

Screen 21.1 The Hierarchical Cluster Analysis Main Dialog Window



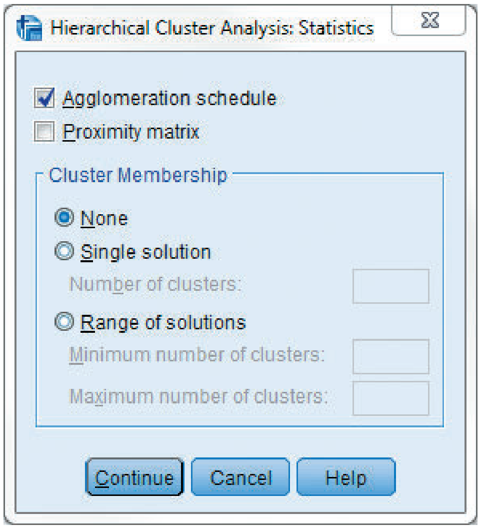
Cluster: ● Cases ○ Variables

Display: ☑ Statistics ☑ Plots

If instead you select the **Cluster Variables** option, you will paste the desired variables into the **Variable(s)** box, but the variable *names* will serve to identify the variables and the **Label Cases by** box will remain empty. Under **Display, Statistics** and **Plots** are selected by default. In most cases you will wish to retain both these selections. Four additional sets of options are represented by the four pushbuttons on the right side of the window. A description of frequently used options for each of these buttons follows.

A click on **Statistics** opens a new dialog window (Screen 21.2, following page). The **Agglomeration schedule** is selected by default and represents a standard feature of cluster analysis (see the Output section for display and explanation). The proximity matrix allows a look at distances between all cases and clusters. With a small file this might be useful, but with large files the number of comparisons grows geometrically, occupies many pages of output, and make this option impractical. In the **Cluster Membership** box, three options exist:

**Screen 21.2** The Hierarchical Cluster Analysis: Statistics Window



Hierarchical Cluster Analysis: Statistics

☑ Agglomeration schedule

□ Proximity matrix

Cluster Membership

● None

○ Single solution

Number of clusters:

○ Range of solutions

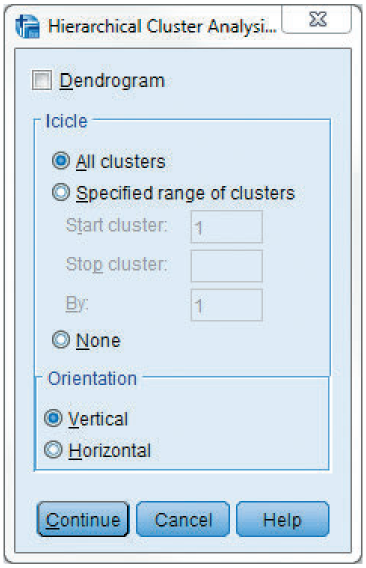
Minimum number of clusters:

Maximum number of clusters:

- **None**: This option actually lists *all* clusters since an icicle plot (a default procedure) will identify all possible solutions. What it does *not* do is identify cases included in each cluster for a *particular* solution.
- **Single solution**: Specify some number greater than 1 that indicates cluster membership for a specific number of clusters. For instance if you type 3 into the box indicating number of clusters, SPSS will print out a three-cluster solution.
- **Range of solutions**: If you wish to see several possible solutions, type the value for the smallest number of clusters you wish to consider in the first box and the value for the largest number of clusters you wish to see in the second box. For instance if you typed in 3 and 5, SPSS would show case membership for a three-cluster, a four-cluster, and a five-cluster solution.

The next available dialog box opens up options for inclusion or modifications of dendograms or icicle plots; plots that are particularly well suited to cluster analysis. Both **icicle plots** and **dendograms** are displayed and explained in the Output section. Please refer to those pages for visual reference if you wish. A click on the **Plots** pushbutton will open Screen 21.3, below.

**Screen 21.3** The Hierarchical Cluster Analysis: Plots Window



Hierarchical Cluster Analysi...

□ Dendrogram

Icicle

● All clusters

○ Specified range of clusters

Start cluster: 1

Stop cluster:

By: 1

○ None

Orientation

● Vertical

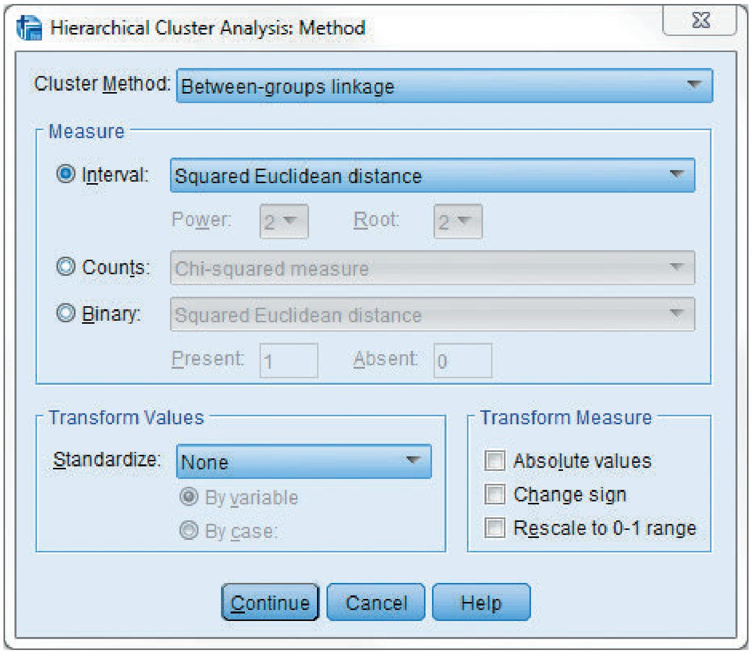
○ Horizontal

The **Dendogram** provides information similar to that covered in the icicle plot but features, in addition, a relative measure of the magnitude of differences between variables or clusters at each step of the process. An icicle plot of the entire clustering process is included by default.

Since early steps in this process have little influence on the final solution, many researchers prefer to show only a relatively narrow range of clustering solutions. This may be accomplished by a click on the **Specific range of clusters** option, followed by typing in the smallest number of clusters you wish to consider, the largest number of clusters you wish to see in a solution, and the interval between those values. For instance if the numbers you entered were 3, 5, and 1, you would see three-cluster, four-cluster, and five-cluster solutions. If the numbers you selected were 2, 10, and 2, SPSS would display solutions for 2, 4, 6, 8, and 10 clusters. A click on **None** eliminates any icicle plots from the output. The **Vertical** option is capable of displaying many more cases on a page (about 24) than the **Horizontal** option (about 14). You may select the horizontal option when there are too many variables or cases to fit across the top of a single page.

The **Method** pushbutton from Screen 21.1 opens the most involved of the cluster analysis windows, Screen 21.4 (below). There are a number of different options for the **Cluster Method**, the **Interval Measure**, and the **Transform Values Standardize**. For each one we will describe the most frequently used option and simply list the others. The manual *IBM SPSS Statistics 19 Guide to Data Analysis* provides descriptions of the others.

Screen 21.4 The Hierarchical Cluster Analysis: Method Window



Hierarchical Cluster Analysis: Method

Cluster Method: Between-groups linkage

Measure

● Interval: Squared Euclidean distance

Power: 2 Root: 2

○ Counts: Chi-squared measure

○ Binary: Squared Euclidean distance

Present: 1 Absent: 0

Transform Values

Standardize: None

● By variable

○ By case:

Transform Measure

□ Absolute values

□ Change sign

□ Rescale to 0-1 range

**Between-groups linkage** (selected in Screen 21.4, above) also called the *average linkage within groups* simply joins the variables or clusters that have the least distance between them at each successive stage of the analysis. A more complete description of this method may be found on page 269. Other options (visible after a click on the ▼ button) include:

- **Within-group linkage** or *average linkage within groups*
- **Nearest neighbor** or *single linkage*
- **Furthest neighbor** or *complete linkage*
- **Centroid clustering**
- **Ward's method**

**Squared Euclidean distance** is the default method of determining distances or similarities between cases or clusters of cases. This concept is discussed fairly extensively in Chapter 19. Briefly, it is the sum of squared differences between matching variables for each case. Other options include:

- **Cosine**: This is a similarity measure based on cosines of vectors of values.
- **Pearson correlation**: This is a similarity measure determined by correlations of vectors of values.
- **Chebychev**: This is a distance measure, the maximum absolute difference between values for items.

- **Block**: This is a distance measure.
- **Minkowski**: This is a distance measure.
- **Customized**: This allows the researcher to create unique criteria for clustering.

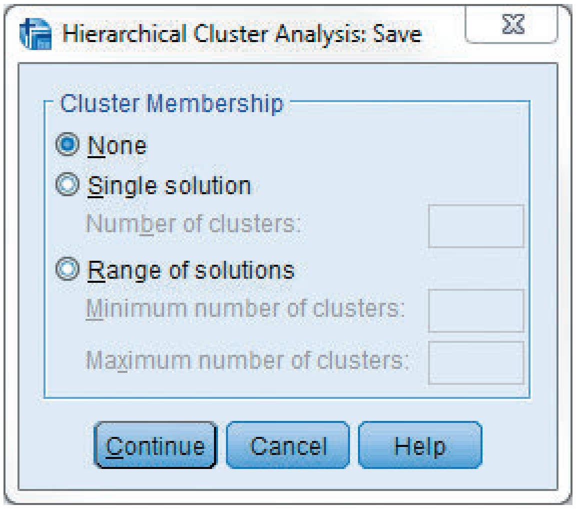
A procedure discussed in the introduction is accessed by selecting **Standardize** (under **Transform Values**). This option provides different methods for standardizing variables. The default for standardization is **None**, but when standardization does take place, **Z scores** is the most frequently used option. Of alternatives to the *z* score, the first three standardization methods listed below will produce identical results. The two that follow, in which the mean or the standard deviation is allowed to vary, may produce different results. Selection of one or another is based upon that which is most appropriate to the data set or most convenient for the researcher. In addition to **Z scores**, the other options include:

- **Range –1 to 1**: All variables are standardized to vary between –1 and 1.
- **Range 0 to 1**: All variables are standardized to vary between 0 and 1.
- **Maximum magnitude of 1**: All variables are standardized to have a maximum value of 1.
- **Mean of 1**: All variables are standardized to have a mean value of 1 (standard deviations may vary).
- **Standard deviation of 1**: All variables are standardized to have a standard deviation of 1 (mean values may vary).

The **Transform Measures** option (lower right corner of the window) simply takes the absolute value, reverses the sign of the original variable values, or rescales variable values on a 0 to 1 metric.

The final small dialog box available in the hierarchical cluster analysis procedure deals with saving new variables. The **Save** window is shown below (Screen 21.5) and allows three options. The default is **None**. If you choose to save new variables, they will appear in the form of a new variable in the last column of your data file and simply include a different coded number for each case. For instance, if you save as a new variable a three-cluster solution, each case will be coded 1, 2, or 3. This dialog box also allows you to save more than one solution as new variables. If you select the **Range of solutions** option and indicate from 3 through 5 clusters, three new variables will be saved: One variable that codes Cases 1, 2, and 3; a second that codes Cases 1, 2, 3, and 4; and a third that codes Cases 1, 2, 3, 4, and 5.

Screen 21.5 The Hierarchical Cluster Analysis: Save New Variables Window



Hierarchical Cluster Analysis: Save

Cluster Membership

● None

○ Single solution

Number of clusters:

○ Range of solutions

Minimum number of clusters:

Maximum number of clusters:

In the step-by-step sequences we will demonstrate three procedures: Step 5 accesses the simplest default cluster analysis, Step 5a provides a cluster analysis with a number of the options discussed earlier, and Step 5b a cluster analysis that clusters variables rather than cases. The italicized text before each box will identify the specifics of the analysis. Step 5a will be illustrated in the Output section.

*To conduct the simplest default hierarchical cluster analysis (but standardizing the metrics of all variables) in which you attempt to cluster the 21 brands of DVD players based on similarities of features, perform the following sequence of steps. The starting point is Screen 21.1. Carry out the four-step sequence (pages 270–271) to arrive at this screen.*

| In Screen | Do This | Step 5 |
|---|---|---|
| 21.1 | 🖱 **brand** → 🖱 *lower* ➡ → 🖱 **price** → press **Shift** *and* 🖱 **exras3** | |
| | *(highlighting all)* → 🖱 *upper* ➡ → 🖱 **Method** | |
| 21.4 | 🖱 ▼ *to the right of* **Standardize** → 🖱 **Z scores** → 🖱 **Continue** | |
| 21.1 | 🖱 **OK** | Back1 |

This procedure selects the brand name to identify different cases, selects all remaining variables as the basis on which clustering takes place, changes all variables to *z* scores to give each variable an equal metric and equal weight, determines distance between variables by the squared Euclidean distance, clusters variables based on between group linkage, and produces the agglomeration schedule and vertical icicle plot as output.

*To conduct a hierarchical cluster analysis of the 21 DVD-player brands based on the 20 variables that describe each* **brand**, *to request the* **Agglomeration Schedule**, *the* **Dendogram**, *and the* **Vertical Icicle** *plot in the output, change all variables to* **Z scores** *to yield equal metrics and equal weighting, select the* **Squared Euclidean distance** *as the method of determining distance between clusters and the* **Furthest neighbor** *method of clustering, and save a three-cluster solution as a new variable, perform the following sequence of steps. The starting point is Screen 21.1 (see page 271). Not demonstrated here, but strongly suggested: List all cases (page 277) with variables in the desired order for sake of reference during the cluster analysis procedure. The Output section includes this list of cases.*

| In Screen | Do This | Step 5a |
|---|---|---|
| 21.1 | 🖱 **brand** → 🖱 *lower* ➡ → 🖱 **price** → press **Shift** *and* 🖱 **exras3** | |
| | *(highlighting all)* → 🖱 *upper* ➡ → 🖱 **Plots** | |
| 21.3 | 🖱 **Dendogram** → 🖱 **Continue** | |
| 21.1 | 🖱 **Method** | |
| 21.4 | 🖱 ▼ *to the right of* **Cluster Method** → 🖱 **Furthest neighbor** | |
| | 🖱 ▼ *to the right of* **Standardize** → 🖱 **Z scores** → 🖱 **Continue** | |
| 21.1 | 🖱 **Save** | |
| 21.5 | 🖱 **Single solution** → type **3** *(In the box to the right)* → 🖱 **Continue** | |
| 21.1 | 🖱 **OK** | Back1 |

*To conduct a simplest default hierarchical cluster analysis in which variables (rather than cases) are clustered, for the 15 efficacy questions* (**effic1** *to* **effic15**) *from the* **helping2.sav** *file, using all standard defaults and without standardization (not needed since all variables have the same metric), perform the following sequence of steps. The starting point is Screen 21.1. Carry out the four-step sequence* (*pages 270–271—selecting* **helping2.sav**) *to arrive at this screen.*

| In Screen | Do This | Step 5b |
|---|---|---|
| 21.1 | **effic1** → press **Shift** *and* **effic15** *(highlighting all)* → *upper* ➡ → | |
| | **Variables** → **OK** | Back1 |

This simple procedure will conduct a cluster analysis of the 15 self-efficacy questions. A summary of variable information such as valid cases and missing values will be printed, followed by an Agglomeration Schedule and a vertical icicle plot to indicate the nature of the clustering process. These results would be interesting to compare with the factor analysis (previous chapter) using the same variables from the same data set.

Upon completion of any of Step 5, 5a, or 5b, the output screen will appear (Screen 1, inside back cover). All results from the just-completed analysis are included in the Output Navigator. Make use of the scroll bar arrows (▲▼▶◀) to view the results. Even when viewing output, the standard menu of commands is still listed across the top of the window. Further analyses may be conducted without returning to the data screen.

## 21.4 Printing Results

Results of the analysis (or analyses) that have just been conducted require a window that displays the standard commands (**File Edit Data Transform Analyze…**) across the top. A typical print procedure is shown below beginning with the standard output screen (Screen 1, inside back cover).

*To print results, from the Output screen perform the following sequence of steps:*

| In Screen | Do This | Step 6 |
|---|---|---|
| Back1 | *Select desired output or edit (see pages 16–22)* → **File** → **Print** | |
| 2.9 | *Consider and select desired print options offered, then* → **OK** | |

*To exit you may begin from any screen that shows the* **File** *command at the top.*

| In Screen | Do This | Step 7 |
|---|---|---|
| Menu | **File** → **Exit** | 2.1 |

Note: After clicking **Exit**, there will frequently be small windows that appear asking if you wish to save or change anything. Simply click each appropriate response.

## 21.5 Output

### 21.5.1 Cluster Analysis

The following output is produced by sequence Step 5a shown on the previous page. What is reproduced here is a fairly close duplicate (both content and format) of what SPSS actually produces. We begin with a list of the variables (the 21 brands of DVD

**Hierarchical Cluster Analysis: The Data File**

| | brand | price | pictur1 | pictur2 | pictur3 | pictur4 | pictur5 | program | recept1 | recept3 | audio1 | audio2 | audio3 | features | events | days | remote1 | remote2 | remote3 | extras1 | extras2 | extras3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | SONNY | 520 | 5 | 5 | 5 | 5 | 5 | 4 | 3 | 3 | 5 | 3 | 3 | 5 | 8 | 31 | 4 | 4 | 4 | 13 | 13 | 13 |
| 2 | ANGLER | 535 | 5 | 5 | 5 | 5 | 5 | 3 | 4 | 3 | 5 | 4 | 4 | 4 | 6 | 365 | 3 | 4 | 4 | 12 | 12 | 12 |
| 3 | MITTENSUB | 515 | 5 | 5 | 5 | 5 | 5 | 4 | 3 | 3 | 5 | 3 | 3 | 5 | 8 | 28 | 4 | 4 | 4 | 13 | 13 | 13 |
| 4 | SINGBO | 470 | 5 | 5 | 5 | 5 | 5 | 3 | 3 | 3 | 3 | 3 | 4 | 4 | 6 | 365 | 2 | 3 | 3 | 7 | 7 | 7 |
| 5 | WHACKACHY | 525 | 5 | 5 | 5 | 5 | 5 | 2 | 4 | 3 | 4 | 3 | 4 | 4 | 6 | 365 | 4 | 3 | 3 | 11 | 11 | 11 |
| 6 | SILVERMOON | 370 | 4 | 4 | 4 | 4 | 4 | 3 | 4 | 3 | 3 | 5 | 4 | 5 | 6 | 365 | 4 | 3 | 3 | 8 | 8 | 8 |
| 7 | EXPERTSEE | 430 | 4 | 4 | 4 | 4 | 4 | 5 | 3 | 4 | 4 | 5 | 4 | 4 | 8 | 365 | 4 | 4 | 4 | 8 | 8 | 8 |
| 8 | FROMSHEEBA | 505 | 5 | 5 | 5 | 5 | 5 | 3 | 4 | 2 | 4 | 4 | 4 | 4 | 8 | 365 | 3 | 3 | 3 | 9 | 9 | 9 |
| 9 | POTASONIC1 | 450 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 4 | 4 | 4 | 8 | 30 | 3 | 3 | 3 | 8 | 8 | 8 |
| 10 | CLIMAX | 365 | 4 | 4 | 4 | 4 | 4 | 2 | 3 | 3 | 4 | 4 | 4 | 3 | 8 | 365 | 4 | 2 | 2 | 4 | 4 | 4 |
| 11 | POTASONIC2 | 435 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 3 | 4 | 4 | 4 | 30 | 4 | 4 | 4 | 9 | 9 | 9 |
| 12 | MAGNESIA | 265 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 2 | 4 | 4 | 3 | 3 | 8 | 365 | 4 | 4 | 4 | 3 | 3 | 3 |
| 13 | DULL | 200 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 2 | 4 | 4 | 4 | 4 | 8 | 365 | 3 | 3 | 3 | 3 | 3 | 3 |
| 14 | BURSTINGSTAR | 380 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 4 | 4 | 4 | 4 | 4 | 30 | 4 | 4 | 4 | 6 | 6 | 6 |
| 15 | VCJ | 420 | 3 | 3 | 3 | 3 | 3 | 2 | 4 | 3 | 4 | 4 | 4 | 4 | 8 | 365 | 2 | 3 | 2 | 6 | 6 | 6 |
| 16 | ARC | 335 | 2 | 2 | 2 | 2 | 2 | 5 | 4 | 3 | 5 | 3 | 4 | 5 | 8 | 365 | 3 | 2 | 3 | 8 | 8 | 8 |
| 17 | MAGNETOX | 205 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 1 | 3 | 4 | 3 | 3 | 8 | 365 | 4 | 4 | 4 | 3 | 3 | 3 |
| 18 | RECALLAKING | 200 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 2 | 3 | 4 | 4 | 4 | 8 | 365 | 3 | 3 | 3 | 4 | 4 | 4 |
| 19 | PAULSANG | 205 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 2 | 4 | 4 | 4 | 3 | 6 | 365 | 3 | 4 | 2 | 2 | 2 | 2 |
| 20 | EG | 275 | 2 | 2 | 2 | 2 | 2 | 3 | 4 | 2 | 4 | 4 | 4 | 3 | 8 | 365 | 3 | 3 | 3 | 2 | 2 | 2 |
| 21 | RASES | 225 | 2 | 2 | 2 | 2 | 2 | 3 | 4 | 2 | 4 | 4 | 4 | 3 | 8 | 365 | 3 | 4 | 2 | 0 | 0 | 0 |

players) and the information in the data file about each one. This is followed by the agglomeration schedule, the vertical icicle plot, and the dendogram. Text between each section clarifies the meaning of the output.

This is simply the data file with brand names to the left and variable names along the top. The **Case Summaries** procedure (see Chapter 4) accesses this important output. Since the file with cluster analysis is usually fairly short, it is often a good idea to print it out because it makes an excellent reference as you attempt to interpret the clustering patterns. As mentioned in the introduction, this data file has been doctored (the brand names too) to create a more interpretable cluster pattern.

Following this listing, descriptive statistics for each variable will be printed out (mean, standard deviation, minimum, maximum, and *N*), followed by a listing of the new variable names after they have been changed to *z* scores (e.g., **price** becomes **zprice, pictur1** becomes **zpictur1, pictur2** becomes **zpictur2**, and so forth). Since these data are straightforward, we will save space by not reproducing them.

On the following page the agglomeration schedule is listed. The procedure followed by cluster analysis at Stage 1 is to cluster the two cases that have the smallest squared Euclidean distance between them. Then SPSS will recompute the distance measures between all single cases and clusters (there is only one cluster of two cases after the first step). Next, the 2 cases (or clusters) with the smallest distance will be combined, yielding either 2 clusters of 2 cases (with 17 cases unclustered) or 1 cluster of 3 cases (with 18 cases unclustered). This process continues until all cases are clustered into a single group. To clarify, we will explain Stages 1, 10, and 14 (following page).

**Agglomeration Schedule**
**Squared Euclidean Distance with Complete Linkage**

| Stage | Cluster Combined | | Coefficients | Stage Cluster First Appears | | Next Stage |
|---|---|---|---|---|---|---|
| | Cluster 1 | Cluster 2 | | Cluster 1 | Cluster 2 | |
| **1** | **1** | **3** | **.002** | **0** | **0** | **17** |
| 2 | 13 | 18 | 2.708 | 0 | 0 | 9 |
| 3 | 12 | 17 | 4.979 | 0 | 0 | 14 |
| 4 | 20 | 21 | 5.014 | 0 | 0 | 7 |
| 5 | 11 | 14 | 8.509 | 0 | 0 | 10 |
| 6 | 5 | 8 | 11.725 | 0 | 0 | 8 |
| 7 | 19 | 20 | 11.871 | 0 | 4 | 14 |
| 8 | 2 | 5 | 13.174 | 0 | 6 | 13 |
| 9 | 13 | 15 | 14.317 | 2 | 0 | 12 |
| **10** | **9** | **11** | **19.833** | **0** | **5** | **15** |
| 11 | 6 | 7 | 22.901 | 0 | 0 | 15 |
| 12 | 10 | 13 | 23.880 | 0 | 9 | 16 |
| 13 | 2 | 4 | 28.378 | 8 | 0 | 17 |
| **14** | **12** | **19** | **31.667** | **3** | **7** | **16** |
| 15 | 6 | 9 | 40.470 | 11 | 10 | 18 |
| 16 | 10 | 12 | 44.624 | 12 | 14 | 19 |
| 17 | 1 | 2 | 47.720 | 1 | 13 | 20 |
| 18 | 6 | 16 | 49.963 | 15 | 0 | 19 |
| 19 | 6 | 10 | 64.785 | 18 | 16 | 20 |
| 20 | 1 | 6 | 115.781 | 17 | 19 | 0 |

At **Stage 1**, Case 1 is clustered with Case 3. The squared Euclidean distance between these two cases is .002. Neither variable has been previously clustered (the two zeros under Clusters 1 and 2), and the next stage (when the cluster containing Case 1 combines with another case) is Stage 17. (Note that at Stage 17, Case 2 joins the Case 1 cluster.)

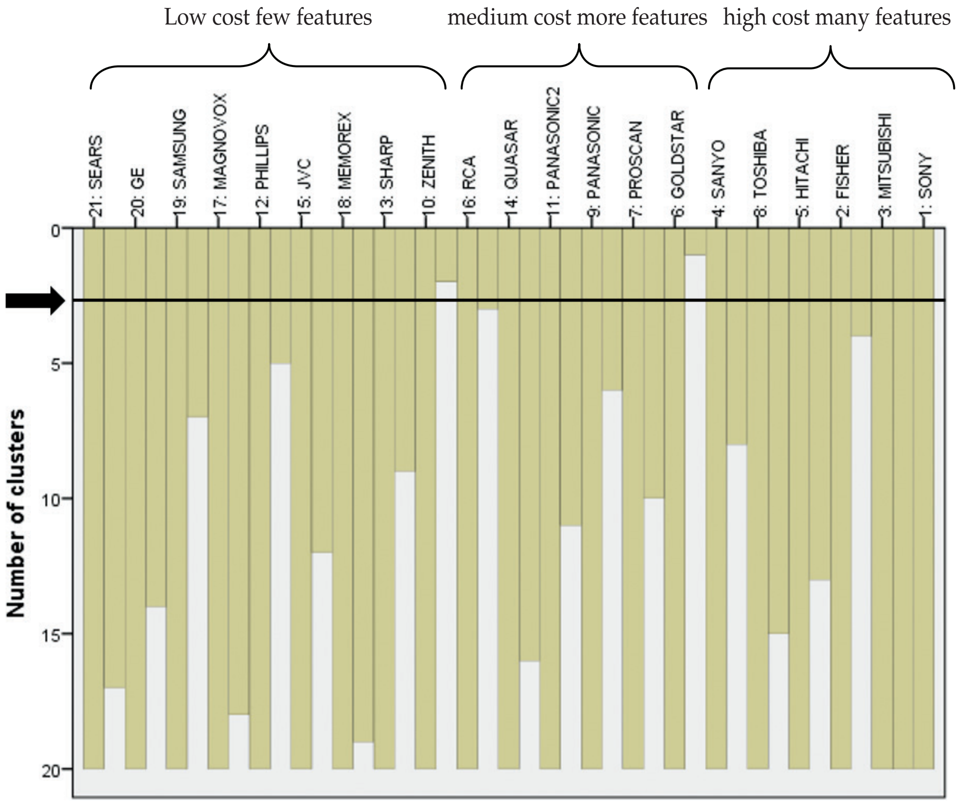
At **Stage 10**, Case 9 joins the Case 11 cluster (Case 11 was previously clustered with Case 14 back in Stage 5, thus creating a cluster of three cases: Cases 9, 11, and 14). The squared Euclidean distance between Case 9 and the Case 11 cluster is 19.833. Case 9 has not been previously clustered (the zero under Cluster 1), and Case 11 was previously clustered at Stage 5. The next stage (when the cluster containing Case 9 clusters) is Stage 15 (when it combines with the Case 6 cluster).

At **Stage 14**, the clusters containing Cases 12 and 19 are joined. Case 12 had been previously clustered with Case 17, and Case 19 had been previously clustered with Cases 20 and 21, thus forming a cluster of five cases (Cases 12, 17, 19, 20, and 21). The squared Euclidean distance between the two joined clusters is 31.667. Case 12 was previously joined at Stage 3 with Case 17. Case 19 was previously joined at Stage 7 with the Case 20 cluster. The next stage when the Case 12 cluster will combine with another case/cluster is Stage 16 (when it joins with the Case 10 cluster).

The icicle plot (that follows) displays the same information graphically. This graph is identical to SPSS output but a heavy horizontal line, arrow, and labeling contributes to the clarity of interpretation.

This graph displays the same information as the agglomeration schedule table except that the value of the distance measure is not shown. The numbers to the left indicate the number of clusters at each level. For instance, the line opposite the "20" has 20 clusters, 19 individual cases, and Brands 1 and 3 joined into a single cluster. Notice the heavy horizontal line and the arrow to the left of the "3." At this stage there are three clusters, and the experimenters determined that three clusters was the most meaningful solution. As you may recall from the introduction, the cluster containing Brands 1, 2, 3, 4, 5, and 8 are DVD players characterized by high price, high picture quality, and many features. The second cluster, Brands 6, 7, 9, 11, 14, and 16, are DVD players characterized by medium-range price, lower picture quality, and fewer features. The final cluster, Brands 10, 12, 13, 15, 17, 18, 19, 20, and 21, are budget models with low price, medium picture quality, and few features.

**Vertical Icicle Plot using Complete Linkage**

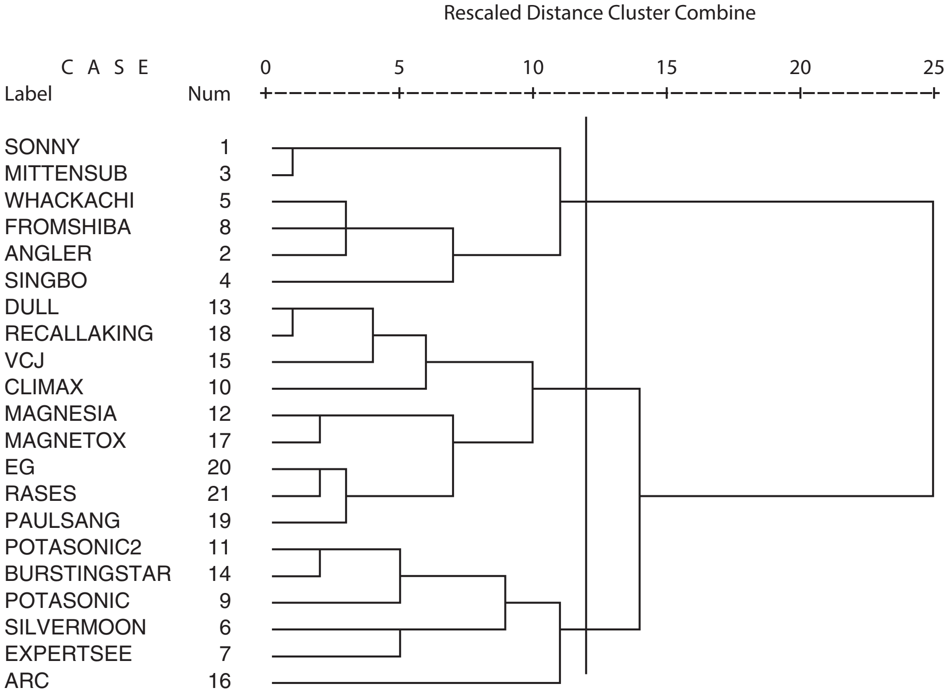


The dendogram, a tree-type display of the clustering process, is the final output display. In this case, the output is essentially identical to the SPSS output.

**Dendogram using Complete Linkage**

```
* * * * HIERARCHICAL CLUSTER ANALYSIS * * * *

Dendrogram using Complete Linkage
```



All three of the visual displays of the clustering procedure (the agglomeration table, the icicle plot, and the dendogram) provide slightly different information about the process. In addition to the branching-type nature of the dendogram, which allows the researcher to trace backward or forward to any individual case or clusters at any level, it adds an additional feature that the icicle plot does not. The 0 to 25 scale along the top of the chart gives an idea of how great the distance was between cases or groups that are clustered at a particular step. The distance measure has been rescaled from 0 to 25, with 0 representing no distance and 25 rating the greatest distance. While it is difficult to interpret distance in the early clustering phases (the extreme left of the graph), as you move to the right relative distances become more apparent.

Note the vertical line (provided by the authors) that designates the three-cluster solution. A similar vertical line placed in different positions on the graph will reveal the number of clusters at any stage by the number of horizontal lines that are crossed. To find the membership of a particular cluster, simply trace backward down the branches (or is it roots?) to the name and case number.

# Chapter 22
# Discriminant Analysis

DISCRIMINANT ANALYSIS is used primarily to predict membership in two or more mutually exclusive groups. The procedure for predicting membership is initially to analyze pertinent variables where the group membership is already known. For instance, a top research university wishes to predict whether applicants will complete a Ph.D. program successfully. Such a university will have many years of records of entry characteristics of applicants and additional information about whether they completed the Ph.D. and how long it took them. They thus have sufficient information to use discriminant analysis. They can identify two discrete groups (those who completed the Ph.D. and those who did not) and make use of entry information such as GPAs, GRE scores, letters of recommendation, and additional biographical information. From these variables it is possible to create a formula that maximally differentiates between the two groups. This formula (if it discriminates successfully) could be used to analyze the likelihood of success for future applicants. There are many circumstances where it is desirable to be able to predict with some degree of certainty outcomes based on measurable criteria: Is an applicant for a particular job likely to be successful or not? Can we identify whether mentally ill patients are suffering from schizophrenia, bipolar mood disorder, or psychosis? If a prisoner is paroled, is he or she likely to return to crime or become a productive citizen? What factors might influence whether a person is at risk to suffer a heart attack or not? The elements common to all five of these examples are that (1) group membership is known for many individuals and (2) large volumes of information are available to create formulas that might predict future outcomes better than current instruments.

By way of comparison, discriminant analysis is similar to cluster analysis (Chapter 21) in that the researcher is attempting to divide individuals or cases (not *variables*) into discrete groups. It differs from cluster analysis in the procedure for creating the groups. Discriminant analysis creates a regression equation (Chapters 15 and 16) that makes use of a dependent (criterion) variable that is discrete rather than continuous. Based on preexisting data in which group membership is already known, a regression equation can be computed that maximally discriminates between the two groups. The regression equation can then be used to help predict group membership in future instances. Discriminant analysis is a complex procedure and the **Discriminant** command contains many options and features that extend beyond the scope of this book. We will therefore, throughout this chapter, refer you to the SPSS Manuals (consult reference section for specifics) for additional options when appropriate.

Like many of the more advanced procedures, **Discriminant** follows a set sequence of steps to achieve a final solution. Some of these steps are more experimenter-driven, and others are more data-driven. These steps will be presented here in the introduction. To assist in this explanation, we will introduce yet another example. Then material in the Step by Step and Output sections will provide additional detail.

## 22.1 The Example: Admission into a Graduate Program

Neither the **grades.sav** nor the **helping.sav** files provide appropriate data sets to demonstrate the discriminant function. While it is possible to use discriminant analysis to predict, for instance, subject's gender or class standing, it is much easier, in the words of psychologist Gordon Allport, to "just ask them." What we have chosen is a topic where discriminant analysis might actually be used: Criteria for acceptance into a graduate program. Every year, selectors miss-guess and select students who are unsuccessful in their efforts to finish the degree. A wealth of information is collected about each applicant prior to acceptance, and department records indicate whether that student was successful in completing the course. Our example uses the information collected prior to acceptance to predict successful completion of a graduate program. The file is called **graduate.sav** and consists of 50 students admitted into the program between 7 and 11 years ago. The dependent variable is **category** (1 = finished the Ph.D., 2 = did not finish), and 17 predictor variables are utilized to predict category membership in one of these two groups:

- **gender**: 1 = female, 2 = male
- **age**: age in years at time of application
- **marital**: 1 = married, 2 = single
- **gpa**: overall undergraduate GPA
- **areagpa**: GPA in their area of specialty
- **grearea**: score on the major-area section of the GRE
- **grequant**: score on the quantitative section of the GRE
- **greverbal**: score on the verbal section of the GRE
- **letter1**: first of the three recommendation letters (rated 1 = weak through 9 = strong)
- **letter2**: second of the three recommendation letters (same scale)
- **letter3**: third of the three recommendation letters
- **motive**: applicant's level of motivation (1 = low through 9 = high)
- **stable**: applicant's emotional stability (same scale for this and all that follow)
- **resource**: financial resources and support system in place
- **interact**: applicant's ability to interact comfortably with peers and superiors
- **hostile**: applicant's level of inner hostility
- **impress**: impression of selectors who conducted an interview

Repeating once more before we move on, the dependent variable is called **category** and has two levels: 1 = those who finished the Ph.D. and 2 = those who did not finish the Ph.D. Finally, although the study has excellent face validity, the data are all fictional—created by the authors to demonstrate the discriminant function.

## 22.2 The Steps Used in Discriminant Analysis

**STEP 1** Selection of variables. Theoretical or conceptual concerns, prior knowledge of the researcher, and some bivariate analyses are frequently used in the initial stage. With the **graduate.sav** file, the number of predictor variables is small enough (17) that it would be acceptable to enter all of them into the regression equation. If, however, you had several hundred variables (e.g., number of questions on many personality tests), this would not be feasible due to conceptual reasons (e.g., collinearity of variables, loss of degrees of freedom) and due to practical reasons (e.g., limitations of available memory). A common first step is to compute a **correlation matrix** of predictor variables. This correlation matrix has a special meaning when applied to discriminant analysis: It is called the *pooled within-group correlation matrix* and involves the average correlation for the two or more correlation matrices for each variable pair. Also available are covariance matrices for separate groups, for pooled within-groups, and for the entire sample. Another popular option is to calculate a number of **t tests** between the two groups for each variable or to conduct **one-way ANOVA**s if you are discriminating into more than two groups. Because the goal of analysis is to find the best discriminant function possible, it is quite normal to try several procedures and several different sets of variables to achieve best results. In the present sample, in a series of *t*-tests significant differences were found between the two levels of **category** (those who finished the Ph.D. and those who did not) for **gpa, grequant, letter1, letter2, letter3, motive, age, interact**, and **hostile**. In the analyses described below, we will demonstrate first a forced-entry procedure that includes all variables that show significant bivariate relationships and then a stepwise procedure that accesses all 17 independent variables.

**STEP 2** Procedure. Two entry procedures are presented in this chapter. The default procedure is designated **Enter independents together** and, in regression vernacular, is a *forced entry* process. The researcher indicates which variables will be entered, and *all* designated variables are entered simultaneously. A second procedure, called **Wilks' lambda**, is a stepwise operation that is based on minimizing the Wilks' lambda (λ) after each new variable has been entered into the regression equation. As in stepwise multiple regression analysis, there is a criterion for entry into the regression equation ($F > 3.84$ is default) and also a criterion for removal from the equation once a variable has been entered if its contribution to the equation drops below a designated level ($F < 2.71$ is default). Wilks' λ is the ratio of within-groups sum of squares to the total sum of squares. It is designed (in this setting) to indicate whether a particular variable contributes significantly to explaining additional variance in the dependent (or criterion) variable. There is an *F* and *p* value associated with Wilks' λ that indicates the level of significance. A more complete description is given in the Output section.

So, which of these two methods usually produces better results? As disturbing as it might seem, the computer (in a stepwise procedure with all variables entered) often does better than when the researcher preselects variables. However, there are frequently conceptual or practical reasons for not allowing the computer to make all of the decisions. With the present data set, when all variables were entered using the **Wilks' lambda** procedure (stepwise selection of variables), a regression equation was produced that misclassified only 3 of 50 cases. Using the **Wilks' lambda** procedure but entering only the variables that showed bivariate differences, 4 of 50 cases were misclassified. Using the **Enter independents together** procedure and entering all 17 variables, 4 of 50 were misclassified, and using the **Enter independents together** procedure with only the 9 variables that showed significant bivariate differences, 5 of 50 cases were misclassified. There are several other selection procedures offered by SPSS, including using canonical correlation to measure the effect size influence of the regression equation on the group membership, and an extensive discussion of these selection methods is provided in the SPSS Manuals.

**STEP 3** Interpretation and use. The rationale behind discriminant analysis is to make use of existing data pertaining to group membership and relevant predictor variables to create a formula that will accurately predict group membership, using the same variables with a new set of subjects. The formula created by the **Discriminant** command will generally not be as accurate in predicting new cases as it was on the original data. With the present data set, **Discriminant** classified 94% of the individuals correctly when predicting group membership for the same 50 subjects that were used to create the regression formula; but it is unlikely to do that well on subjects where membership is unknown. *IBM SPSS Statistics Guide to Data Analysis* describes the *jackknife* procedure to identify a more likely percentage of correct classifications in applications when group membership is unknown. Over time, the discriminant formula can be improved and refined as additional data become available.

The best text known by the authors on discriminant analysis is by Geoffrey McLachlan (2004). Please see the reference section for additional information.

# 22.3 Step by Step

## 22.3.1 Discriminant Analysis

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | → All Programs ▷ → IBM SPSS Statistics → IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | → → | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| | Step 2 |
|---|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* | |

*Screens 1 and 2 (on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **graduate.sav** *file for further analyses:*

| In Screen | Do This | Step 3 |
|---|---|---|
| Front1 | File → Open → Data [*or* ] | |
| Front2 | type `graduate.sav` → Open [*or* graduate.sav] | Data |

Whether first entering SPSS or returning from earlier operations the standard menu of commands across the top is required. As long as it is visible you may perform any analyses. It is not necessary for the data window to be visible.

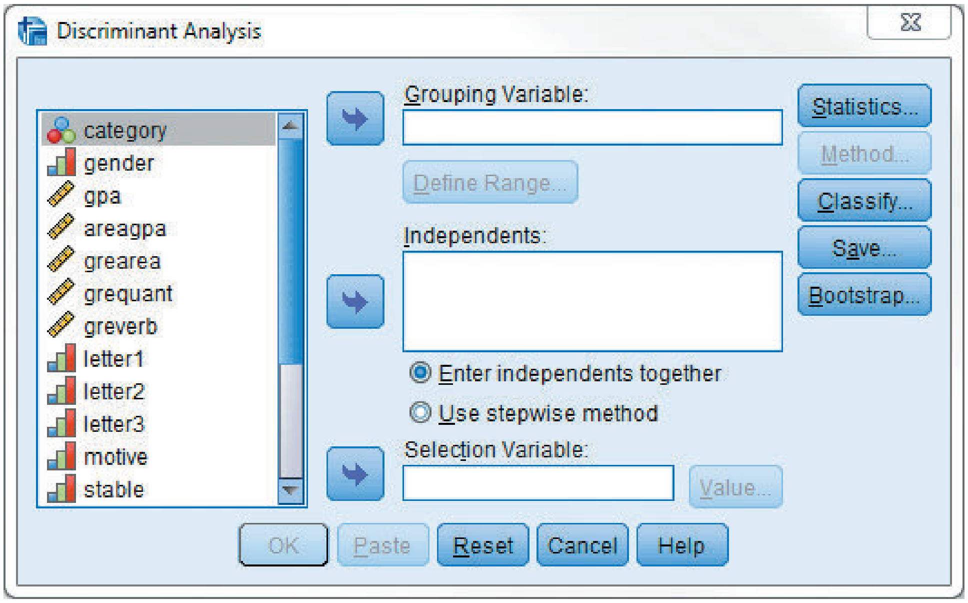
*After completion of Step 3 a screen with the desired menu bar appears. When you click a command (from the menu bar), a series of options will appear (usually) below the selected command. With each new set of options, click the desired item. The sequence to access discriminant analysis begins at any screen with the menu of commands visible:*

| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | Analyze → Classify → Discriminant | 22.1 |

The Step 4 sequence opens the main dialog window for discriminant analysis (Screen 22.1, below). In addition to the traditional list of variables to the left and the five function pushbuttons to the right, this screen features:

1. a **Grouping Variable** box for entry of a single categorical dependent variable,
2. an **Independents** box where one may designate as many variables as desired to assist in predicting the levels of the dependent variable,
3. two options (**Enter independents together** and **Use stepwise method**) that identify the entry procedure for selecting variables, and
4. four pushbuttons on the right side of the window that identify a number of different additional selections.

**Screen 22.1** The Discriminant Analysis Main Dialog Window



For two of the pushbuttons (**Save** and **Value** to the right of **Selection Variable**) we describe their function but do not display the associated dialog box.

The **Selection Variable** option operates like the **If** procedure described in Chapter 4. If, for instance, you paste **gender** into the Selection Variable window, the **Value** button would brighten up. A click on this button opens a tiny dialog window that allows you to select which level of the variable (in this case, 1 = women, 2 = men) you wish to select for the procedure. The analysis will then be conducted with the selected subjects or cases. This selection may be reversed later if you wish.

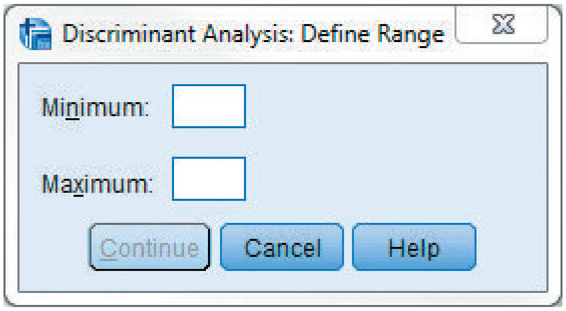
The **Save** option allows you to save as new variables:

- **Predicted group membership**
- **Discriminant scores**
- **Probabilities of group membership**

All three of these constructs will be discussed in some detail throughout this chapter.

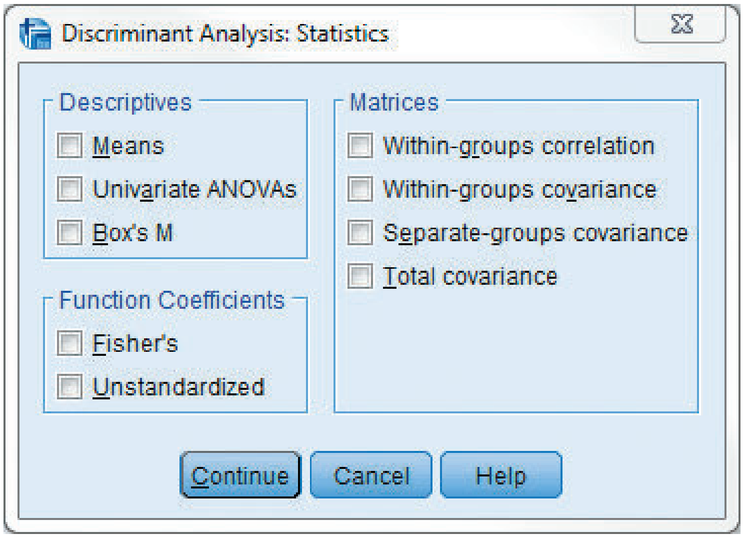
The final item concerning the main dialog window for discriminant analysis is the coding of the dependent variable. Although discriminant analysis is used most frequently to categorize into one of two groups, it is also possible to create discriminant functions that group subjects into three or more groups. When you enter the **Grouping variable** (**category** in this case), the **Define range** pushbutton becomes bright and before you can conduct the analysis you must define the lowest and highest coded value for the dependent variable in the small dialog box that opens (Screen 22.2, following page). In this case, **category** has only two levels and you would enter 1 and 2 in the two boxes then click **Continue**. If you have a variable with more than two levels you may specify numbers that represent less than the full range of values.

**Screen 22.2** The Discriminant Analysis: Define Range Window



A click on the **Statistics** pushbutton opens a new window with a number of frequently desired options. This dialog box is shown as Screen 22.3 (below). Because all items in the **Statistics** box are frequently used, each will be described in the list that follows. Most of these items are used to analyze the bivariate nature of the variables prior to beginning the discriminant process.

**Screen 22.3** Discriminant Analysis: Statistics Window

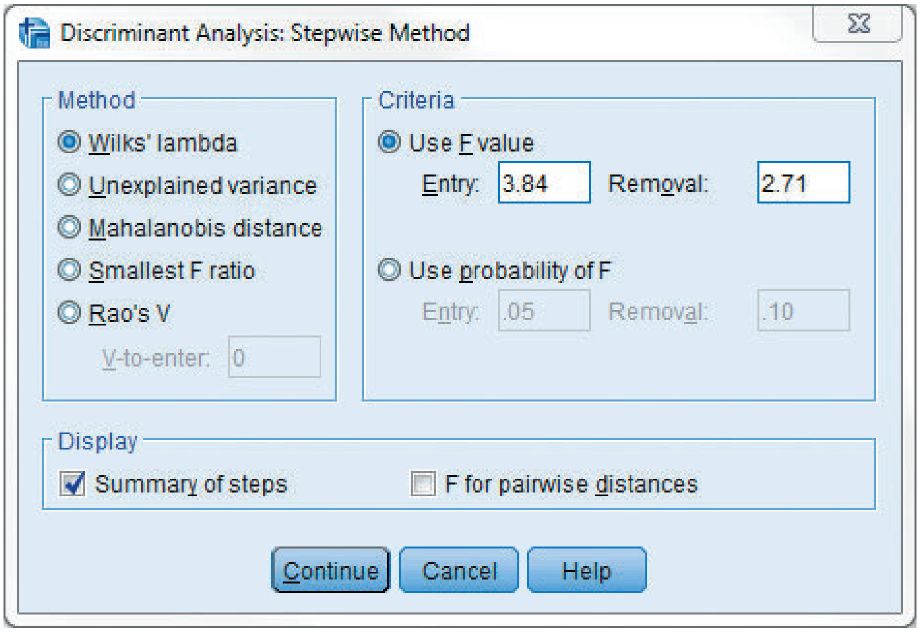


- **Means:** The means and standard deviations for each variable for each group (the two levels of **category** in this case) and for the entire sample.
- **Univariate ANOVAs:** If there are two levels of the dependent variable, these will be *t* tests (using an *F* rather than a *t* statistic). This compares the mean values for each group for each variable to see if there are significant univariate differences between means.
- **Box's M:** The test for equality of the covariances matrices for each level of the dependent variable. This is a test of the multivariate normality of your data.
- **Fisher's Function Coefficients:** These are the canonical discriminant function coefficients designed to discriminate maximally between levels of the dependent variable.
- **Unstandardized Function Coefficients:** The unstandardized coefficients of the discriminant equation based on the raw scores of discriminating variables.

- **Within-groups correlation:** A matrix composed of the means of each corresponding value within the two (or more) matrices for each level of the dependent variable.
- **Within-groups covariance**: Identical to the previous entry except a covariance matrix is produced instead of a correlation matrix.
- **Separate-groups covariance**: A separate covariance matrix for members of each level of the dependent variable.
- **Total covariance**: A covariance matrix for the entire sample.

A click on the **Method** button opens the **Stepwise Method** dialog box (Screen 22.4, below). This option is available only if you have previously selected the **Use stepwise method** option from the main dialog box (Screen 22.1). This window provides opportunity to select the **Method** used in creating the discriminant equation, the **Criteria** for entry of variables into the regression equation, and two **Display** options. The **Wilks' lambda** is the only method we will describe here. The **Wilks' lambda** method is a stepwise procedure that operates by minimizing the Wilks' lambda value after each new variable has been entered into the regression equation. A more complete description of this may be found on page 283 .

Screen 22.4 Discriminant Analysis: Stepwise Method Window



Discriminant Analysis: Stepwise Method

Method
● Wilks' lambda
○ Unexplained variance
○ Mahalanobis distance
○ Smallest F ratio
○ Rao's V
V-to-enter: 0

Criteria
● Use F value
Entry: 3.84 Removal: 2.71
○ Use probability of F
Entry: .05 Removal: .10

Display
☑ Summary of steps
□ F for pairwise distances

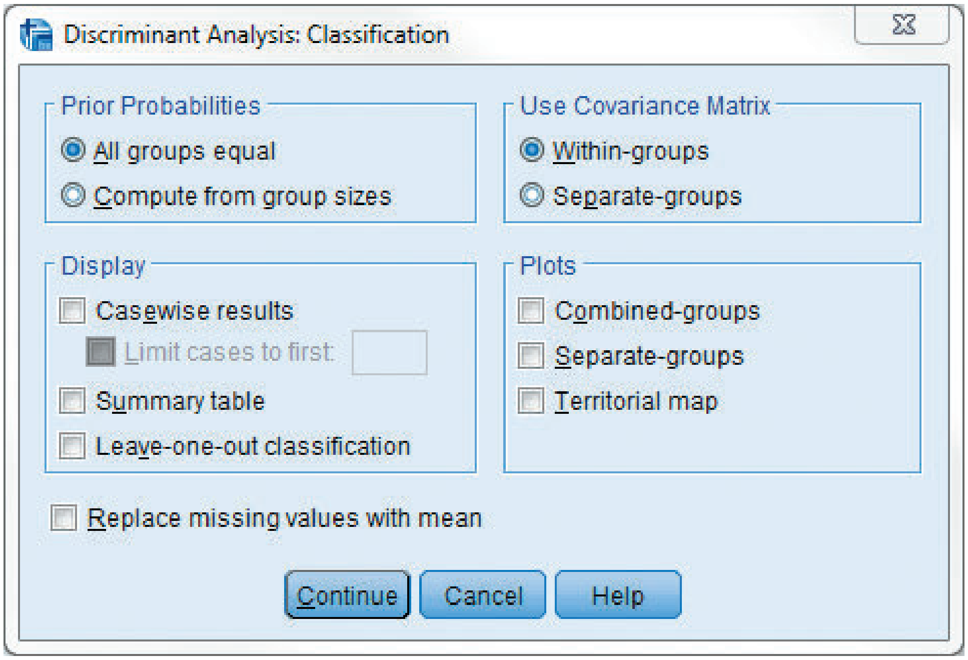
The **Criteria** for entry sets default *F* values of 3.84 to enter a new variable and 2.71 to remove an already entered variable. These *F* values represent significance levels of approximately .05 and .10, respectively. In discriminant analysis, researchers often set more generous entry criteria and, selection of *F* values as low as 1.15 to enter and 1.0 to remove are not uncommon since the goal is to maximize accurately in predicting groups with the available data.

Under the **Display** section, the **Summary of steps** option is selected by default, is central to discriminant analysis output, and under normal circumstances is always retained. The **F for pairwise distances** relates only to the **Mahalanobis distance** method.

The final table considered here deals with the classification and display of discriminant results. A click of the **Classify** pushbutton opens the window shown as Screen 22.5. Most options in this box are of interest to the researcher and each will be described in the list that follows:

- **All groups equal:** Probability of membership into two or more groups created by the dependent variable are assumed to be equal.

Screen 22.5 The Discriminant Analysis: Classification Window



Discriminant Analysis: Classification

Prior Probabilities
- ● All groups equal
- ○ Compute from group sizes

Use Covariance Matrix
- ● Within-groups
- ○ Separate-groups

Display
- □ Casewise results
- □ Limit cases to first:
- □ Summary table
- □ Leave-one-out classification

Plots
- □ Combined-groups
- □ Separate-groups
- □ Territorial map

□ Replace missing values with mean

- **Compute from group sizes:** If groups are of different sizes, the groups are weighted by the proportion of the number of cases in each group.
- **Combined-groups plot:** The histogram (if two groups) or scatterplot (if more than two groups) includes all groups on a single plot.
- **Separate-groups plots:** This option creates as many plots as there are groups and displays only one group on each plot.
- **Territorial map:** This chart shows centroids and boundaries in a graphic format; it is used only if there are three or more levels of the dependent variable.
- **Within-groups covariance matrix:** This is the default and will classify cases based on the pooled within-groups covariance matrix.
- **Separate-groups covariance matrices:** Classifies cases based on separate covariance matrices for each group.
- **Casewise results:** This is a useful output option if your file is not too large. It lists each case with information concerning the actual group, the group classification, probabilities for inclusion in the group, and the discriminant scores. This output is displayed in the Output section.
- **Summary table:** Another handy output device that sums up the number and percent of correct and incorrect classifications for each group. This is also displayed in the Output section.
- **Leave-one-out classification:** Each case in the analysis is classified by the functions derived from all cases other than that case. This is also known as the U method.
- **Replace Missing Values with mean:** As a high-quality researcher, you have, of course, dealt with missing values long before this.

In this chapter there will be a single Step 5 sequence. Yes, it is possible to create a simplest-default version of discriminant analysis, but anyone who has read this far would want to consider carefully a number of available options. The italicized text prior to the Step by Step box describes the included options. Then the Output section in narrative, definitions, and visual displays does an adequate job of describing the meaning of the results.

*To conduct a discriminant analysis that predicts membership into two groups based on the dependent variable* **category** *and creating the discriminant equation with inclusion of 17 independent variables (from* **age** *to* **stable***) selected by a* **stepwise** *procedure based on the minimization of* **Wilks' lambda** *at each step with an F-to-enter of 1.15 and an F-to-remove of 1.00; further, selecting* **Means, Box's M***, and* **Univariate ANOVAs***, to gain a fuller understanding of the univariate nature of independent variables; and selecting for output the* **Combined-groups plot***, the* **Results for each case***, and the* **Summary table***, perform the following sequence of steps. The starting point is Screen 22.1. Carry out the four-step sequence (page 284) to arrive at this screen.*

| In Screen | Do This | Step 5 |
|---|---|---|
| 22.1 | ✳ **category** → ✳ *upper* ➡ → ✳ **Define Range** | |
| 22.2 | **type** 1 → **press** **TAB** → **type** 2 → ✳ **Continue** | |
| 22.1 | ✳ **gender** → **press** **Shift** *and* ✳ **impress** *(highlight all 17)* → ✳ *middle* ➡ | |
| | → ✳ **Use stepwise method** → ✳ **Statistics** | |
| 22.3 | ✳ **Means** → ✳ **Univariate ANOVAs** → ✳ **Box's M** → ✳ **Continue** | |
| 22.1 | ✳ **Method** | |
| 22.4 | **press** **TAB** *twice* → **type** 1.15 → **press** **TAB** *once* → **type** 1.00 → | |
| | → ✳ **Continue** | |
| 22.1 | ✳ **Classify** | |
| 22.5 | ✳ **Combined groups** → ✳ **Casewise results** → ✳ **Summary table** → | |
| | ✳ **Continue** | |
| 22.1 | ✳ **OK** | **Back1** |

# 22.4 Output

## 22.4.1 Discriminant Analysis

The following output is produced by sequence Step 5 (above). As with factor analysis, SPSS's efforts to create a coherent and concise output format provide considerable opportunity for future growth. Their output occupies 18 bewildering pages. We present, below, results that we feel are most relevant to the discriminant process. We apologize that you may have to do a substantial amount of hunting to find the corresponding information in the SPSS output.

**Number of Cases by Group**

| Category | Label | Unweighted Cases |
|---|---|---|
| 1 | Finished Ph.D. | 25 |
| 2 | Did not finish Ph.D. | 25 |
| Total | | 50 |

The small table shown above reminds the researcher of the categories, the labels, and the number of unweighted cases. Prior versions of SPSS printed this table at the top of the output. SPSS 21 no longer does, but we retain it here as a good starting point. The table below (actually four tables in the SPSS output, which have been combined) identifies basic preliminary *uni* variate information (the means for the two levels and the overall mean for each variable); Wilks' lambda, *F*, and significance values contribute *bi*variate information about the differences between means for each variable. The *F* and *Signif* values identify for which variables the two groups differ significantly. This is the type of information that the researcher considers before running a discriminant analysis. Definitions of terms follow.

**Group means, Wilks' Lambda (U-statistic) and Univariate F-ratio**

| Variable | Finish Ph.D. Mean | Not Finish Mean | Total Mean | Wilks' Lambda | F | df1/df2 | Signif. |
|---|---|---|---|---|---|---|---|
| gender | 1.240 | 1.480 | 1.360 | .938 | 3.200 | 1 / 48 | .080 |
| gpa | 3.630 | 3.390 | 3.510 | .795 | 12.360 | 1 / 48 | .001 |
| areagpa | 3.822 | 3.734 | 3.778 | .951 | 2.493 | 1 / 48 | .121 |
| grearea | 655.600 | 648.800 | 652.200 | .998 | .113 | 1 / 48 | .738 |
| grequant | 724.000 | 646.800 | 685.400 | .628 | 28.393 | 1 / 48 | .000 |
| greverb | 643.200 | 620.000 | 631.200 | .974 | 1.283 | 1 / 48 | .263 |
| lettter1 | 7.720 | 6.160 | 6.940 | .650 | 25.889 | 1 / 48 | .000 |
| letter2 | 7.640 | 6.360 | 7.000 | .756 | 15.476 | 1 / 48 | .000 |
| letter3 | 7.960 | 6.160 | 7.060 | .534 | 41.969 | 1 / 48 | .000 |
| motive | 8.360 | 7.280 | 7.820 | .679 | 22.722 | 1 / 48 | .000 |
| stable | 6.400 | 6.360 | 6.380 | 1.000 | .007 | 1 / 48 | .934 |
| resource | 5.920 | 5.640 | 5.780 | .993 | .337 | 1 / 48 | .564 |
| marital | 1.640 | 1.560 | 1.600 | .993 | .322 | 1 / 48 | .573 |
| age | 29.960 | 25.120 | 27.540 | .768 | 14.526 | 1 / 48 | .000 |
| interact | 7.000 | 6.160 | 6.580 | .904 | 5.079 | 1 / 48 | .029 |
| hostile | 2.120 | 3.080 | 2.600 | .787 | 13.017 | 1 / 48 | .001 |
| impress | 7.280 | 6.880 | 7.080 | .972 | 1.378 | 1 / 48 | .246 |

| Term | Definition/Description |
|---|---|
| **VARIABLE** | Names of the independent variables. |
| **FINISH PH.D. MEAN** | Mean value for each variable for the first level of category, those who finished the Ph.D. |
| **NOT FINISH MEAN** | Mean value for each variable for the second level of category, those who did not finish. |
| **TOTAL MEAN** | The mean value for each variable for all subjects. Since there are equal numbers of subjects in each level of **category**, this number is simply the average of the other two means. |
| **WILKS' LAMBDA** | The ratio of the within-groups sum of squares to the total sum of squares. This is the proportion of the total variance in the discriminant scores *not* explained by differences among groups. A lambda of 1.00 occurs when observed group means are equal (all the variance is explained by factors *other than* difference between these means), while a small lambda occurs when within-groups variability is small compared to the total variability. A small lambda indicates that group means appear to differ. The associated significance values indicate whether the difference is significant. |
| **F** | *F* values are the same as those calculated in a one-way analysis of variance. This is also the square of the *t* value calculated from an independent samples *t* test. |
| **df1/df2** | Degrees of freedom for each entered variable (1) and degrees of freedom for the entire analysis (50 − 1 − 1 = 48). |
| **SIGNIFICANCE** | The *p* value: Likelihood that the observed *F* value could occur by chance. |

Next is specification information concerning the discriminant procedure about to take place.

**Stepwise Variable Selection: Selection Rule: Minimize Wilks' Lambda**

| Max. Number of Steps | Min. Tolerance Level | Min. Partial F to Enter | Max. Partial F to Remove |
|---|---|---|---|
| 34 | .001 | 1.150 | 1.000 |

Prior probability for each group is .5000 (Note: this information is present in SPSS output but not as a table)

| Term | Definition/Description |
|---|---|
| **STEPWISE VARIABLE SELECTION** | This procedure enters variables into the discriminant equation, one at a time, based on a designated criterion for inclusion (*F* ≥ 3.84 is default); but will drop variables from the equation if the inclusion requirement drops below the designated level when other variables have been entered. |
| **SELECTION RULE** | The procedure selected here is to minimize Wilks' lambda at each step. |
| **MAXIMUM NUMBER OF STEPS** | Two times the number of variables designated in the ANALYSIS subcommand line (2 × 17 = 34). This number reflects that it is possible to enter a maximum of 17 variables and to remove a maximum of 17 variables. |
| **MINIMUM TOLERANCE LEVEL** | The tolerance level is a measure of linear dependency between one variable and the others. If a tolerance is less than .001, this indicates a high level of linear dependency and SPSS will not enter that variable into the equation. |
| **MAXIMUM F TO ENTER, MINIMUM F TO REMOVE** | This indicates that any variable with an *F* ratio greater than 1.15 (indicating some influence on the dependent variable) will enter the equation (in the order of magnitude of *F*) and any variable that has an *F* ratio drop below 1.00 after entry into the equation will be removed from the equation. |
| **PRIOR PROBABILITY FOR EACH GROUP** | The .5000 value indicates that groups are weighted equally. |

What follows is the output that deals with the regression analysis. This output displays the variables that were included in the discriminant equation with associated statistical data; the variables that did not meet the requirements for entry with associated statistical data; and the order in which variables were entered (or removed), along with Wilks' λ, significance, and variable labels.

**Variables in the Analysis after Step 11**

| Variable | Tolerance | F to remove | Wilks' Lambda | Label |
|---|---|---|---|---|
| letter3 | .703 | 3.813 | .315 | letter-of-recommendation number 3 |
| motive | .762 | 4.026 | .317 | rating of student motivation |
| age | .744 | 5.268 | .326 | age at time of entry into program |
| gender | .736 | 5.816 | .330 | women = 1, men = 2 |
| impress | .878 | 2.851 | .308 | selector's overall impression of candidate |
| resource | .845 | 2.428 | .305 | personal and financial resources |
| greverb | .005 | 5.035 | .324 | score on verbal GRE |
| grearea | .005 | 4.632 | .321 | score on area GRE |
| letter2 | .358 | 1.955 | .302 | letter-of-recommendation number 2 |

**Variables not in the Analysis after Step 11**

| Variable | Tolerance | Min Tolerance | F to enter | Wilks' λ | Label |
|---|---|---|---|---|---|
| gpa | .747 | .004 | .104 | .287 | overall undergraduate GPA |
| areagpa | .747 | .004 | .027 | .288 | GPA in major field |
| grequant | .548 | .004 | .000 | .288 | score on quantitative GRE |
| letter1 | .692 | .004 | .058 | .287 | first letter of recommendation |
| stable | .817 | .004 | .088 | .287 | emotional stability |
| marital | .914 | .005 | .046 | .287 | marital status |
| interact | .694 | .005 | .746 | .282 | ability at interaction with others |
| hostile | .765 | .005 | .001 | .288 | level of hostility |

**Variables Entered/Removed**

| Step | Action | | Wilks' Lambda | df1/df2/df3 | Exact F | df1/df2 | Sig. | Label |
|---|---|---|---|---|---|---|---|---|
| | Entered | Removed | | | | | | |
| 1 | letter3 | | .534 | 1 / 1 / 48 | 41.97 | 1 / 48 | .000 | third letter of recommendation |
| 2 | motive | | .451 | 2 / 1 / 48 | 28.64 | 2 / 47 | .000 | students motivation |
| 3 | letter1 | | .415 | 3 / 1 / 48 | 21.60 | 3 / 46 | .000 | first letter of recommendation |
| 4 | age | | .391 | 4 / 1 / 48 | 17.50 | 4 / 45 | .000 | age in years at entry |
| 5 | gender | | .363 | 5 / 1 / 48 | 15.42 | 5 / 44 | .000 | 1 = women 2 = men |
| 6 | impress | | .332 | 6 / 1 / 48 | 14.42 | 6 / 43 | .000 | rating of selector's impression |
| 7 | resource | | .319 | 7 / 1 / 48 | 12.83 | 7 / 42 | .000 | financial/personal resources |
| 8 | greverb | | .310 | 8 / 1 / 48 | 11.42 | 8 / 41 | .000 | GRE score on verbal |
| 9 | grearea | | .398 | 9 / 1 / 48 | 10.49 | 9 / 40 | .000 | GRE score on major area |
| 10 | | letter1 | .302 | 8 / 1 / 48 | 11.86 | 8 / 41 | .000 | first letter of recommendation |
| 11 | letter2 | | .288 | 9 / 1 / 48 | 11.00 | 9 / 40 | .000 | second letter of recommendation |

These three charts indicate which variables were included in the final discriminant function and which ones were not. Notice that all variables in the analyses after 11 steps have higher than the acceptable tolerance level (.001) and have *F* values greater than 1.15. Variables not in the equation all have acceptable tolerance levels, but the *F*-to-enter value is less than 1.05 for each of them. The third table gives a summary of the step-by-step procedure. Notice that in a total of 11 steps 9 variables were entered, but 1 variable (**letter1**) was dropped at Step 10 due to an *F* value falling below 1.00. Then at Step 11 the final variable was added (**letter2**).

This result raises an important experimental concern. When the values for the three letters were originally entered, no attention was given to which was stronger than which. Thus there is no particular rationale to drop **letter1** in favor of **letter2**, as suggested by these results. The appropriate response for the researcher would be to go back into the raw data file and reorder letters from strongest to weakest for each subject (e.g., **letter1** is strongest, **letter2** is next strongest, and **letter3** is weakest) and update the data file. Then, when a discriminant analysis is conducted, if one of the letters has greater influence than another, it may have meaning. It might indicate, for instance, that **letter1** (the strongest letter) has a significant influence but that **letter3** (the weakest) does not contribute significantly to the discriminating process. While this is representative of the thinking the researcher might do, remember that these data are fictional and thus do not reflect an objective reality.

Most of the terms used here are defined on the previous pages. Definitions of terms unique to this section follow:

| Term | Definition/Description |
|---|---|
| **NUMBER OF VARIABLES IN** | Indicates the numbers of variables in the discriminant equation at each step. |
| **SIG** | This is a measure of multivariate significance, not the significance of each new variable's unique contribution to explaining the variance. |

Next is the test for multivariate normality of the data.

**Box's Test of Equality of Covariance Matrices**

| Group | Label | Rank | Log Determinant | Box's M | Approx F | df1 | df2 | signif |
|---|---|---|---|---|---|---|---|---|
| 1 | Finished Ph.D. | 9 | 11.340 | | | | | |
| 2 | Not finished Ph.D. | 9 | 12.638 | | | | | |
| Pooled within-groups | | 9 | 13.705 | 82.381 | 1.461 | 45 | 7569.06 | .024 |

| Term | Definition/Description |
|---|---|
| **RANK** | Rank or size of the covariance matrix. The **9** indicates that this is a 9 × 9 matrix, the number of variables in the discriminant equation. |
| **LOG DETERMINANT** | Natural log of the determinant of each of the two (the two levels of the dependent variable, **CATEGORY**) covariance matrices. |
| **POOLED WITHIN-GROUPS COVARIANCE MATRIX** | A matrix composed of the means of each corresponding value within the two 9 × 9 matrices of the two levels of **category**. |
| **BOX'S M** | Based on the similarities of determinants of the covariance matrices for the two groups. It is a measure of multivariate normality. |
| **APPROXIMATE F** | A transformation that tests whether the determinants from the two levels of the dependent variable differ significantly from each other. It is conceptually similar to the *F* ratio in ANOVA, in which the between-groups variability is compared to the within-groups variability. |
| **SIGNIFICANCE** | A significance value of .024 suggests that data do differ significantly from multivariate normal. However, a value less than .05 does not automatically disqualify the data from the analysis. It has been found that even when multivariate normality is violated, the discriminant function can still often perform surprisingly well. Since this value is low, it would be well to look at the univariate normality of some of the included variables. For instance, we know that the **gender** variable is not normally distributed, but inclusion of **gender** improves the discriminating function of the equation. |

What follows is information concerning the canonical discriminant function, correlations between each of the discriminating variables and the canonical discriminant function, the standardized discriminant function coefficients (these are the coefficients in the discriminant equation), and the group centroids.

**Structure Matrix**
***Pooled-within-groups correlations (ordered by size of correlation)***

| | |
|---|---|
| letter3 | .594 |
| grequant | .489 |
| motive | .437 |
| letter1 | .429 |
| letter2 | .361 |
| age | .350 |
| hostile | –.326 |
| gpa | .270 |
| areagpa | .172 |
| gender | –.164 |
| impress | .108 |
| greverb | .104 |
| interact | .070 |
| stable | .059 |
| resource | .053 |
| grearea | .031 |
| marital | .013 |

**Standardized Canonical Discriminant Function Coefficients**

| | |
|---|---|
| gender | –.492 |
| grearea | –5.645 |
| greverb | 5.821 |
| letter2 | –.427 |
| letter3 | .417 |
| motive | .410 |
| resource | .308 |
| age | .469 |
| impress | .326 |

**Functions at Group Centroids**
***(Canonical Discriminant Function at Group Means)***

| group | Function 1 |
|---|---|
| 1 | 1.541 |
| 2 | –1.541 |

**Eignevalues and Wilks' Lambda**

| | | % of | | Canonical | Test of | Wilks' | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Function** | **Eigenvalue** | **Variance** | **Cum %** | **Correlation** | **function** | **Lambda** | $\chi^2$ | **df** | **Sig** |
| 1 | 2.475 | 100.00 | 100.00 | .844 | 1 | .288 | 54.18 | 9 | .000 |

The terms used in the tables on the previous page are defined here:

| Term | Definition/Description |
|---|---|
| **STANDARDIZED CANONICAL DISCRIMINANT FUNCTION COEFFICIENTS** | This is the list of coefficients of the discriminant equation. The standardized scores are similar to Betas in multiple regression and are useful to give an idea of the magnitude of each effect. Each subject's standardized discriminant score would be computed by entering his or her standardized values for each of the nine variables in the equation. (Note that variables are standardized based on the pooled within-group standard deviations—the square root of the relevant ANOVA's mean square error—rather than the simple standard deviation for the variable.) The discriminant equation follows:<br>$D = -.49(\textbf{gender}) + -5.65(\textbf{grearea}) + 5.82(\textbf{greverb}) + -.43(\textbf{letter2}) + .42(\textbf{letter3}) + .41(\textbf{motive}) + .31(\textbf{resource}) + .47(\textbf{age}) + .33(\textbf{impress})$ |
| **FUNCTION TEST OF FUNCTION** | The one (1) designates information about the *only* discriminant function created with two levels of the criterion variable. If there were three levels of the criterion variable, there would be information listed about two discriminant functions. |
| **WILKS' LAMBDA** | Ratio of the within-groups sum of squares to total sum of squares (more complete definition earlier in this chapter, page 283). |
| **CHI-SQUARE ($\chi^2$)** | A measure of whether the two levels of the function significantly differ from each other based on the discriminant function. A high chi-square value indicates that the function discriminates well. |
| **DF** | Degrees of freedom is equal to the number of variables used in the discriminant function (nine). |
| **SIG.** | *p* value associated with the chi-square function. |
| **EIGENVALUE** | Between-groups sums of squares divided by within-groups sums of squares.<br>A large eigenvalue is associated with a strong function. |
| **% OF VARIANCE, CUM %** | The function always accounts for 100% of the variance. |
| **CANONICAL CORRELATION** | The canonical correlation is a correlation between the discriminant scores and the levels of the dependent variable. Please refer to the chart on the following page to further clarify. Note the scores in the extreme right column. Those are discriminant scores. They are determined by substituting into the discriminant equation the relevant variable measures for each subject. There are 50 subjects; thus there will be 50 discriminant scores. There are also 50 codings of the dependent variable (**category**) that show 25 subjects coded 1 (finished the Ph.D.) and 25 subjects coded 2 (didn't finish). The canonical correlation is the correlation between those two sets of numbers, and is often used as an effect size measure to determine how strong the relationship between the discriminant scores and the actual groups that subjects are in. A high correlation indicates a function that discriminates well. The present correlation of .844 is extremely high (1.00 is perfect). |
| **POOLED-WITHIN-GROUPS-CORRELATIONS** | "Pooled within group" differs from "values for the entire (total) group" in that the pooled values are the average (mean) of the group correlations. If the *N*s are equal (as they are here), then this would be the same as the value for the entire group. The list of 17 values is the correlations between each variable of interest and the discriminant scores. For instance, the first correlation listed (**letter3** .59437) is the correlation between the 50 ratings for **letter3** and the 50 discriminant scores (extreme right column on the chart that follows). |
| **GROUP CENTROIDS** | The average discriminant score for subjects in the two groups. More specifically, the discriminant score for each group when the variable means (rather than individual values for each subject) are entered into the discriminant equation. Note that the two scores are equal in absolute value but have opposite signs. The dividing line between group membership is zero (0). |

The following chart is titled "Casewise Statistics" and gives information about group membership for each subject, probability of group membership, and discriminant scores. Only 16 of the 50 cases are shown to conserve space. In the third column, the asterisks identify cases that were misclassified.

**Casewise Statistics**

| Case # | Actual Group | Highest Group | | | Second-Highest Group | | Discriminant Score |
|---|---|---|---|---|---|---|---|
| | | Predicted Group | P(D>d\| G = g) | P(G = g\| D = d) | 2nd highest Group | P(G = g\| D = d) | |
| 1 | 1 | 1 | .988 | .992 | 2 | .008 | 1.557 |
| 2 | 1 | 1 | .479 | .999 | 2 | .001 | 2.249 |
| 3 | 1 | 1 | .622 | .998 | 2 | .002 | 2.035 |
| 4 | 1 | 1 | .721 | .975 | 2 | .025 | 1.184 |
| 5 | 2 | 1** | .624 | .962 | 2 | .038 | 1.052 |
| 6 | 1 | 1 | .472 | .999 | 2 | .001 | 2.261 |
| – | – | – | – | – | – | – | – |
| – | – | – | – | – | – | – | – |
| 37 | 2 | 2 | .681 | .970 | 1 | .0308 | −1.130 |
| 38 | 2 | 2 | .712 | .974 | 1 | .026 | −1.172 |
| 39 | 1 | 2** | .314 | .839 | 1 | .161 | −.535 |
| 40 | 2 | 2 | .639 | .998 | 1 | .002 | −2.010 |
| – | – | – | – | – | – | – | – |
| – | – | – | – | – | – | – | – |
| 45 | 2 | 2 | .420 | .906 | 1 | .094 | −.736 |
| 46 | 2 | 2 | .330 | 1.000 | 1 | .000 | −2.516 |
| 47 | 1 | 2** | .976 | .992 | 1 | .008 | −1.572 |
| 48 | 2 | 2 | .276 | 1.000 | 1 | .000 | −2.631 |
| 49 | 2 | 2 | .820 | .996 | 1 | .004 | −1.769 |
| 50 | 2 | 2 | .353 | 1.000 | 1 | .000 | −2.471 |

**Classification Results**

| Actual Group | Number of Cases | Predicted Group Membership | |
|---|---|---|---|
| | | Category 1 | Category 2 |
| Category 1 — Finished Ph.D. | 25 | 23 (92%) | 2 (08%) |
| Category 2 — Not Finished Ph.D. | 25 | 1 (04%) | 24 (96%) |

Once again, the definitions of terms is the entire commentary on the two tables shown above.

| Term | Definition/Description |
|---|---|
| **ACTUAL GROUP** | Indicates the actual group membership of that subject or case. |
| **PREDICTED GROUP** | Indicates the group the discriminant function assigned this subject or case to. Asterisks (**) indicate a misclassification. |
| **P(D > d \| G = g)** | Given the discriminant value for that case (*D*), what is the probability of belonging to that group (*G*)? |
| **P(G = g \| D = d)** | Given that this case belongs to a given group (*G*), how likely is the observed discriminant score (*D*)? |
| **2ND GROUP** | What is the second most likely assignment for a particular case? Since there are only two levels in the present data set, the second most likely group will always be the "other one." |
| **DISCRIMINANT SCORES** | Actual discriminant scores for each subject, based on substitution of variable values into the discriminant formula (presented earlier). |
| **CLASSIFICATION RESULTS** | Simple summary of number and percent of subjects classified correctly and incorrectly. |

# Chapter 23
# General Linear Models: MANOVA and MANCOVA

THIS IS the first chapter describing a procedure that uses several dependent variables concurrently within the same analysis: Multivariate Analysis of Variance (MANOVA) and Covariance (MANCOVA). The **General Linear Model** procedure is used to perform MANOVA in SPSS and is one of the more complex commands in SPSS. It can be used, in fact, to compute multivariate linear regression, as well as MANOVA and MANCOVA. This chapter describes how to perform MANOVA and MANCOVA, and Chapter 24 goes on to illustrate multivariate analysis of variance using within-subjects designs and repeated-measures designs. Because the procedures are so complex, we will restrict our discussion here to the most frequently used options. The procedures described in this chapter are an extension of ANOVA; if you are unfamiliar with ANOVA, you should ground yourself firmly in those operations (see Chapters 12–14) before attempting to conduct a MANOVA.

As described in earlier chapters, an independent-samples *t* test indicates whether there is a difference between two separate groups on a particular dependent variable. This simple case (one independent variable with two levels and one continuous dependent variable) was then extended to examine the effects of

- an independent variable with more than two levels (one-way ANOVA),
- multiple independent variables (two- and three-way ANOVA), and
- including the effects of covariates (ANCOVA).

Throughout this progression from very simple to quite complex independent variable(s), the dependent variable has remained a single continuous variable. The good news is that (at least until the next chapter), the independent variables and covariates involved in MANOVA and MANCOVA procedures will not get more complicated.

There are times, however, when it may be important to determine the effects of one or more independent variables on *several* dependent variables simultaneously.

For example, it might be interesting to examine the differences between men and women (**gender**, an independent variable with two levels) on both previous GPAs (**gpa**) and scores on the final exam (**final**). One popular approach to examining these gender differences is to do two separate *t* tests or a one-way ANOVA (remember, because $t^2 = F$, the tests are equivalent). This approach has the advantage of conceptual clarity and ease of interpretation; however, it does have disadvantages. In particular, when several separate tests are performed (in this case, two: one for each dependent variable), the experimentwise (or family-wise) error increases—that is, the chance that one or more of your findings may be due to chance increases. Furthermore, when dependent variables are correlated (and previous GPA and score on a final exam usually are) with each other, doing separate tests may not give the most accurate picture of the data.

In response to these and other problems associated with doing multiple *t* tests or multiple ANOVAs, Hotelling's $T^2$ was developed to replace the *t* test, and MANOVA—Multivariate Analysis of Variance—was developed to replace ANOVA. SPSS performs these tests using the **General Linear Model—Multivariate** procedure. This procedure can also analyze covariates, allowing the computation of Multivariate Analysis of Covariance (MANCOVA). These tests examine whether there are differences among the dependent variables *simultaneously*; one test does it all. Further analyses allow you to examine the *pattern* of changes in the dependent variables, either by conducting a series of univariate *F* tests or by using other post hoc comparisons.

Please note that whenever you are using multiple dependent variables, it is important to be certain that the dependent variables do not exhibit linear dependency on each other. For example, it would be incorrect to analyze class percent (**percent**) and total points received (**total**) together because the percent for the class depends on the total points received.

Just as in univariate analysis of variance (ANOVA), MANOVA produces an *F* statistic to determine whether there are significant differences among the groups. MANOVA is designed to test for interactions as well as for main effects. Since more than one dependent variable is present, however, multivariate *F* statistics involve matrix algebra and examine the differences between all of the dependent variables simultaneously.

We once more use the **grades.sav** file to test the effects of students' section (**section**) as well as whether they are upper or lower division (**lowup**) to determine the influence on students' scores on the five quizzes (**quiz1** to **quiz5**). In addition to this example, further analysis will be performed using previous GPA (**gpa**) as a covariate; that is, the effects of **gpa** on the dependent variables are removed before the MANOVA itself is performed. Although this is a moderately complex example (with two independent variables, one covariate, and five dependent variables), this procedure provides an example by which, in your own analyses, it is easy to increase or decrease the number of variables. So, if you have more or fewer independent or dependent variables than our example, you may simply use as many variables as you need. At the minimum, you need two dependent variables (if you have fewer dependent variables, it's not MANOVA, right?) and one independent variable with two levels. There is no theoretical maximum number of independent and dependent variables in your analysis, but in reality the sample size will limit you to a few variables at a time. Covariates are optional.

## 23.1 Step by Step

### 23.1.1 General Linear Model: MANOVA and MANCOVA

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | ✳ ⊞ → ✳ All Programs ▷ → ✳ IBM SPSS Statistics → ✳ IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | ✳ 🚀 → ✳ ■ → ✳ Σ | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| | Step 2 |
|---|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* | |

*Screens 1 and 2 (on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **grades.sav** *file for further analyses:*

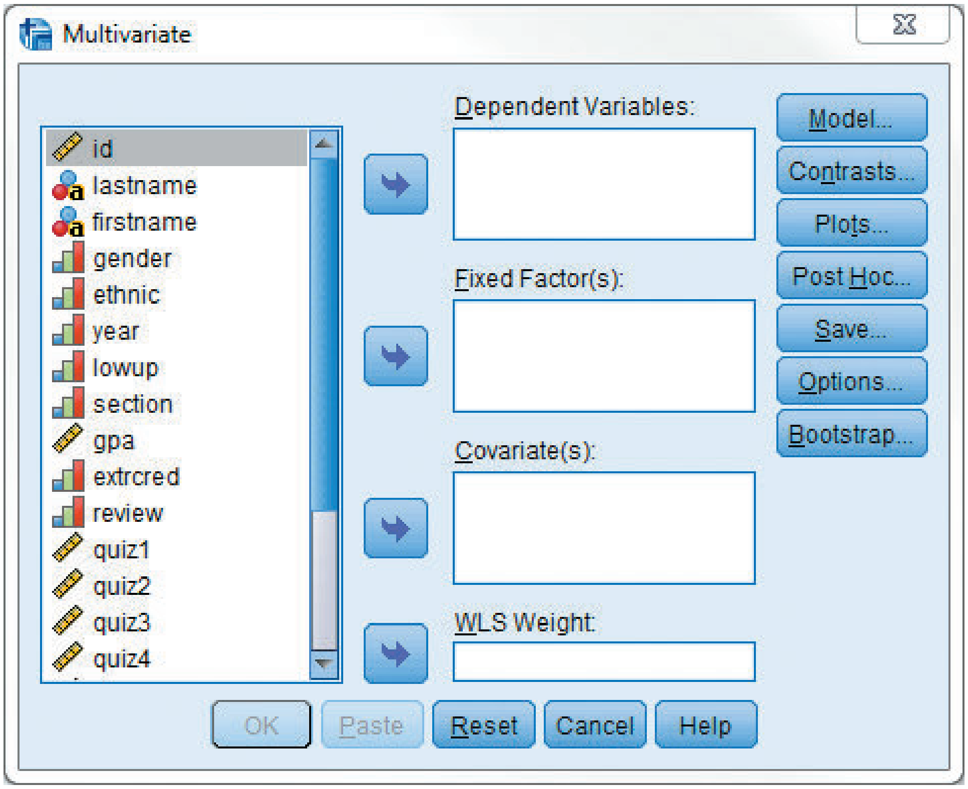
| In Screen | Do This | Step 3 |
|---|---|---|
| Front1 | ✱ **File** → **Open** → ✱ **Data** [*or* ✱ 📂] | |
| Front2 | **type** `grades.sav` → ✱ **Open** [*or* ✱ ✱ **grades.sav**] | Data |

*After completion of Step 3 a screen with the desired menu bar appears. When you click a command (from the menu bar), a series of options will appear (usually) below the selected command. With each new set of options, click the desired item. The sequence to access General Linear Models begins at any screen with the menu of commands visible:*

| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | ✱ **Analyze** → **General Linear Model** → ✱ **Multivariate** | 23.1 |

After this sequence step is performed, a window opens (Screen 23.1, below) in which you specify the dependent variables and independent variables (called fixed factors in this screen). The box to the left of the window contains all variables in your data file. Any variables that you wish to treat as dependent variables in the analysis should be moved into the **Dependent Variables** box by clicking on the top ➡. In our example, we will examine two independent variables and five dependent variables in the **grades.sav** file. As dependent variables, we will examine the five quizzes. For the independent variables, or fixed factors, we examine the three classes (**section**) and whether the student is upper- or lower-divisions (**lowup**).

**Screen 23.1** The General Linear Models—Multivariate Window (for MANOVA & MANCOVA)

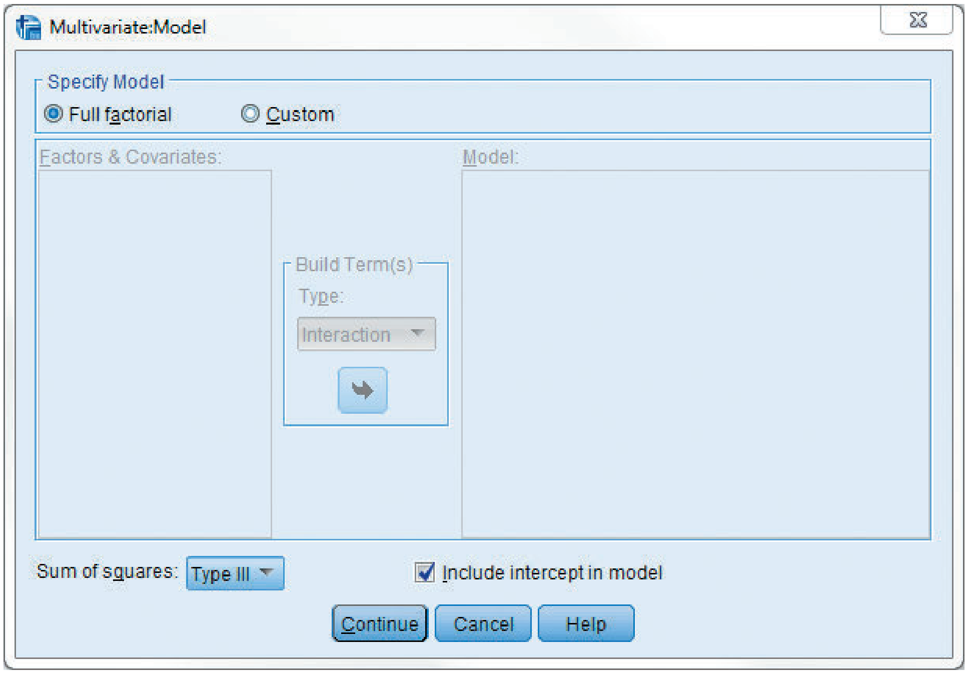


Covariates may also be entered in Screen 23.1, if desired. If you wish to perform a MANOVA, without any covariates, then you should leave the **Covariate(s)** box empty. If, however, you wish to factor out the effect of one or more variables from the MANOVA, then you should enter these variables in the **Covariate(s)** box. Do this by selecting each variable from the box on the left of the window, and clicking on the ➡ to the left of the **Covariate(s)** box.

After the covariates (if desired) are specified, there are four more buttons in the main dialog box that may need to be selected: **Model, Plots, Post Hoc**, and **Options**. (The **Contrasts** and **Save** buttons are more advanced than this introductory chapter will consider.) Each of these important selections will be described.

In most cases, a full factorial model is desired: This produces a model with all main effects and all interactions included, and is the default. At times, you may wish to test only certain main effects and interactions. To accomplish this, use the **Model** button (from Screen 23.1), which calls up Screen 23.2 (below). To select which main effects and interactions you wish to test, select the **Custom** button, and move any main effects and interactions desired in the model from the **Factor & Covariates** box to the left of the window, to the **Model** box to the right of the window, by clicking the ➡ button.

Screen 23.2 Model selection dialog window for Multivariate General Linear Models



Multivariate:Model

Specify Model

● Full factorial ○ Custom

Factors & Covariates:

Model:

Build Term(s)

Type: Interaction

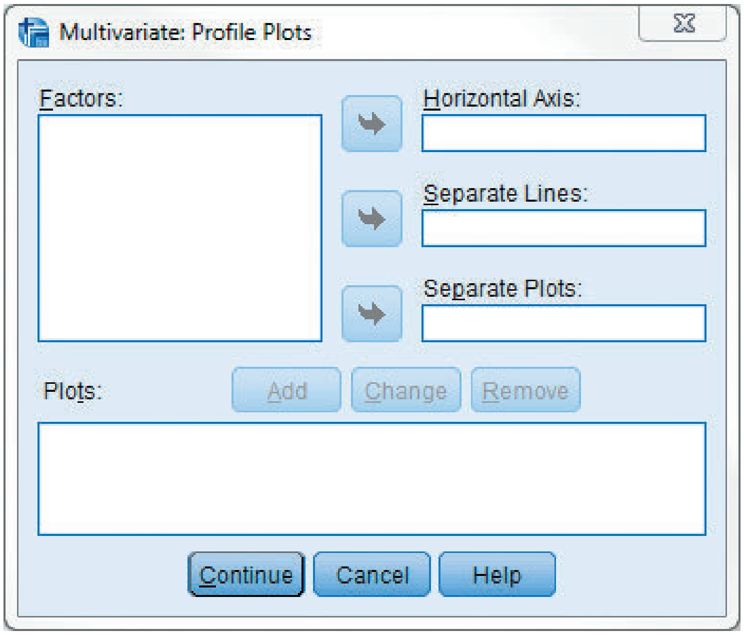
Sum of squares: Type III

☑ Include intercept in model

The Model dialog box also allows you to select the type of **Sum of Squares**. You may select Types I, II, III, or IV: Type III is the default and is appropriate for most situations. Type IV is usually more appropriate if you have missing cells in your design (that is, some cells in your MANOVA model that do not have any participants).

When you are interpreting MANOVA or MANCOVA results, it is often useful to see plots of the means of the dependent variables as determined by the different levels of the factors. To do this, select the **Plots** button from Screen 23.1; this calls up another dialog box, shown in Screen 23.3 (following page). This dialog allows you to specify how to plot each factor. Separate plots will be produced for each dependent variable in your analysis. Note that if your analysis includes one or more covariates, the plots produced will not include the *actual* means of your data, but instead will include the *estimate* of the means *adjusted* for the covariate(s). In that case, it is often useful to produce a set of profile **Plots** without the covariate.

Screen 23.3 Profile Plot selection dialog box for Multivariate General Linear Models



On the left side of Screen 23.3 are listed the **Factors** in your model. To the right are three boxes; into each of these boxes you can move one of the **Factors** by clicking on the ➡ button. These boxes let you specify which factor you want to be drawn with separate categories across the **Horizontal Axis**, which factor you want to have drawn with **Separate Lines** (as many separate lines as there are levels of your factor), and which factor (if any) you want to be used to produce several **Separate Plots** (one plot for each level of that factor).

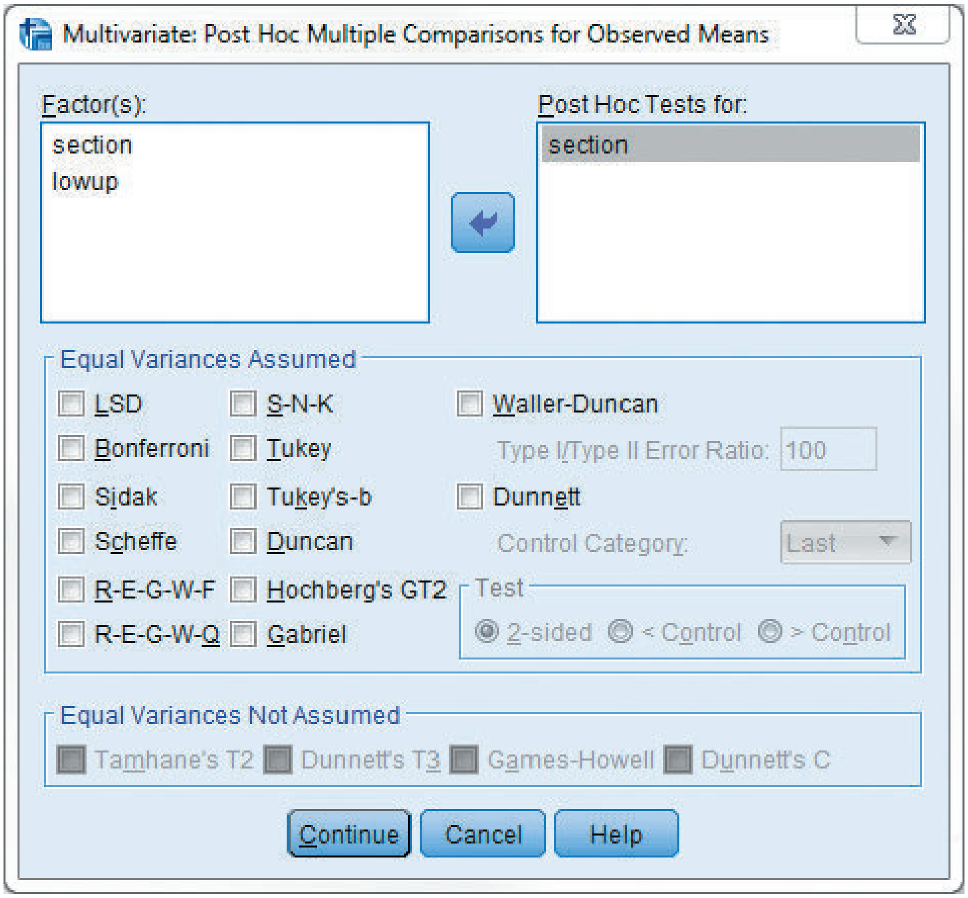
You must specify a **Horizontal Axis**, but **Separate Lines** and **Separate Plots** are optional. If you want **Separate Plots**, however, you must specify **Separate Lines** first. Once you have specified a plot (or series of plots), click **Add** to add it to the list of plots to be produced at the bottom of the dialog. Click **Continue** once you are finished specifying which plots you wish.

In addition to using plots to help you interpret any significant main effects and interactions you may find, **Post Hoc Multiple Comparisons** are often used to determine the specific meaning of main effects or interactions. *Post hoc* tests are used to determine which levels of a variable are significantly different from other levels of that variable; these tests are done for each factor that you specify, and are produced for each dependent variable that is in your analysis.

Because including covariates goes beyond using only dependent and independent variables, but instead examines the effect of the independent variables on the dependent variables above and beyond the effects of the covariates on the dependent variables, *post hoc* tests are not available for analyses that include a covariate.

To select *post hoc* tests, click on the **Post Hoc** button (on Screen 23.1). SPSS produces a truly dizzying array of *post hoc* tests (Screen 23.4) from which you may select whichever of the 18 tests you want. Before selecting tests, you need to specify which **Factor(s)** you want to perform **Post Hoc Tests for**, by selecting one or more variables in the **Factor(s)** box and moving them to the **Post Hoc Tests** for box by clicking on the ➡ button.

**Screen 23.4** Post Hoc Multiple Comparisons for Observed Means dialog box for Multivariate General Linear Models



Factor(s):
section
lowup

Post Hoc Tests for:
section

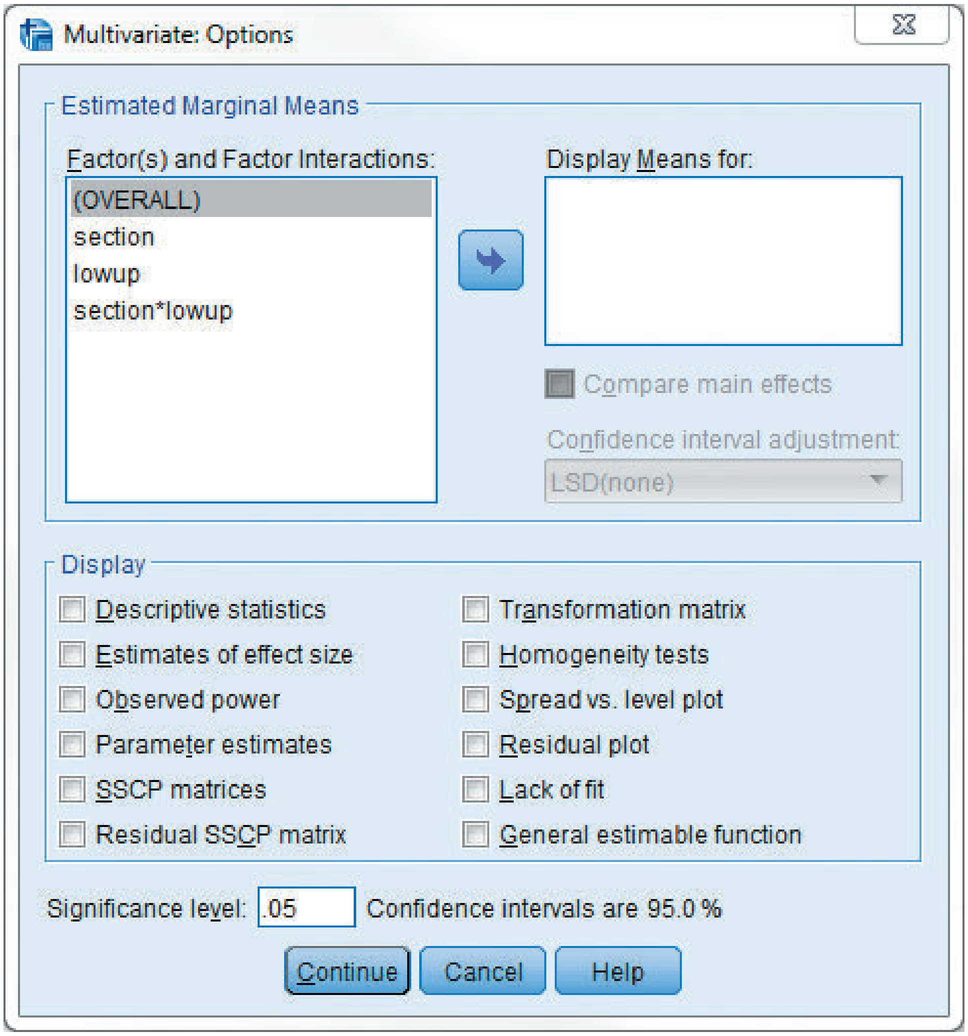
Multivariate: Post Hoc Multiple Comparisons for Observed Means
Equal Variances Assumed
□ LSD
□ S-N-K
□ Waller-Duncan
□ Bonferroni
□ Tukey
Type I/Type II Error Ratio: 100
□ Sidak
□ Tukey's-b
□ Dunnett
□ Scheffe
□ Duncan
Control Category: Last
□ R-E-G-W-F
□ Hochberg's GT2
Test
□ R-E-G-W-Q
□ Gabriel
● 2-sided ○ < Control ○ > Control
Equal Variances Not Assumed
□ Tamhane's T2 □ Dunnett's T3 □ Games-Howell □ Dunnett's C
Continue
Cancel
Help

Once you have selected one or more factors, you may select which of the *post hoc* tests you wish to perform. We don't recommend that you run all 18; in fact, many of these tests are very rarely used. Our discussion will focus on several of the most frequently used *post hoc* tests:

- **LSD**: LSD stands for "least significant difference," and all this test does is to perform a series of *t* tests on all possible combinations of the independent variable on each dependent variable. This is one of the more "liberal" tests that you may choose: Because it does not correct for experimentwise error, you may find significant differences due to chance alone. For example, if you are examining the effects of an independent variable with 5 levels, the LSD test will perform 10 separate *t* tests (1 for each possible combination of the levels of the factor). With an alpha of .05, each individual test has a 5% chance of being significant purely due to chance; because 10 tests are done, however, the chance is quite high that at least 1 of these tests will be significant due to chance. This problem is known as experimentwise or family-wise error.
- **Bonferroni**: This test is similar to the **LSD** test, but the Bonferroni test adjusts for experimentwise error by dividing the alpha value by the total number of tests performed. It is therefore a more "conservative" test than the least significant difference test.
- **Scheffé**: The Scheffé test is still more conservative than the Bonferroni test, and uses *F* tests (rather than the *t* tests in the least significant difference and Bonferroni tests). This is a fairly popular test.
- **Tukey**: This test uses yet a different statistic, the Studentized range statistic, to test for significant differences between groups. This test is often appropriate when the factor you are examining has many levels.

It should be noted that the four *post hoc* tests described here, as well as the majority of the *post hoc* tests available, assume that the variances of your cells are all equal. The four tests listed at the bottom of Screen 23.4 do not make this assumption and should be considered if your cells do not have equal variances. To determine whether or not your cells have equal variances, use the **Homogeneity tests** (described in the **Options** section after Screen 23.5).

**Screen 23.5** Options dialog box for Multivariate General Linear Models



Multivariate: Options

Estimated Marginal Means

Factor(s) and Factor Interactions:
(OVERALL)
section
lowup
section*lowup

Display Means for:

□ Compare main effects

Confidence interval adjustment: LSD(none)

Display

□ Descriptive statistics
□ Estimates of effect size
□ Observed power
□ Parameter estimates
□ SSCP matrices
□ Residual SSCP matrix
□ Transformation matrix
□ Homogeneity tests
□ Spread vs. level plot
□ Residual plot
□ Lack of fit
□ General estimable function

Significance level: .05 Confidence intervals are 95.0 %

Continue Cancel Help

In the **Options** dialog box, the default has nothing selected (no means, descriptive statistics, estimates of effect size, or homogeneity tests). If this is what you wish, then you do not need to click **Options**. If you do wish to choose these or other options, then click on **Options** and Screen 23.5 appears. Many selections are included here that will not be described due to space limitations; the most critical and frequently used ones follow:

- **Estimated Marginal Means**: By selecting each factor in the left box, and clicking on the ➡ button, means will be produced for each dependent variable at each level of that factor. If you do not have a covariate in your analysis, then the means produced will be the actual means of your data; if you have one or more covariates, then means will be adjusted for the effect of your covariate(s). In the **Factor(s) and Factor Interactions** box, **(OVERALL)** refers to the grand mean across all cells. Checking the **Compare main effects** box produces a series of *post hoc* tests examining the differences between cells in each factor. You may choose an LSD or Bonferroni comparison (described earlier), as well as a Sidak test (more conservative than the Bonferroni test). If you want other tests, use the **Post Hoc** option (described earlier) when you are not using covariates.

- **Descriptive Statistics:** Produces means and standard deviations for each dependent variable in each cell.
- **Estimates of effect size:** Produces eta squared ($\eta^2$), the effect-size measure. This indicates how much of the total variance is explained by the independent variables.
- **Observed power:** Allows you to see the probability of finding a significant effect, given your sample size and effect size.
- **Parameter estimates:** Produces parameter estimates (and significance values) for each factor and covariate in the model. This is particularly useful when your model has one or more covariates.
- **Homogeneity tests:** Examines whether the variance-covariance matrices are the same in all cells.
- **Significance level** and **Confidence interval:** Lets you specify the alpha value that you want to use for the analysis. You specify the significance level (the default is .05), and SPSS calculates the confidence intervals based on that significance level.

As you can see in Screen 23.5, many other options are available in SPSS. The SPSS manuals describe them fully, but they are beyond the scope of this book. Also, some background information useful in interpreting **Residual SSCP matrix**es and **Residual plot**s is found in Chapter 28 of this book.

*The following steps compute a multivariate analysis of variance (MANOVA) with the five quizzes (**quiz1** to **quiz5**) as the dependent variables, and with the section number (**section**) and status (lower- or upper-division; **lowup**) as the independent variables. Please perform whichever of Steps 1–4 (pages 297–298) are necessary to arrive at Screen 23.1.*

| In Screen | Do This | Step 5 |
|---|---|---|
| 23.1 | ✳ **quiz1** → *hold down* **Shift** *and* ✳ **quiz5** → ✳ *top* ➡ → ✳ **section** → | |
| | ✳ *second* ➡ → ✳ **lowup** → ✳ *second* ➡ → ✳ **OK** | Back1 |

*If you want to include one or more covariates in the analysis, they may be specified in the* **Covariate(s)** *box. In the following variation of Step 5, we include* **gpa** *as a covariate.*

| In Screen | Do This | Step 5a |
|---|---|---|
| 23.1 | ✳ **quiz1** → *hold down* **Shift** *and* ✳ **quiz5** → ✳ *top* ➡ → | |
| | ✳ **section** → ✳ *second* ➡ → ✳ **lowup** → ✳ *second* ➡ → | |
| | ✳ **gpa** → ✳ *third* ➡ → ✳ **OK** | Back1 |

*The following variation of Step 5 includes a variety of useful supplementary statistics and graphs. Please note that, because* post hoc *tests are included, a covariate is not included. If you wish to include a covariate, you may do so, but you will not be able to select* post hoc *tests, and any graphs requested will produce means adjusted for the covariate, rather than the actual means of the data. Also note that post hoc tests are only performed on* **section**, *because post hoc tests are not very useful when applied to a factor with only two levels (such as* **lowup**).

| In Screen | Do This | Step 5b |
|---|---|---|
| 23.1 | ✱ **quiz1** → *hold down* **Shift** *and* ✱ **quiz5** → ✱ *top* ➥ | |
| | → ✱ **section** → ✱ *second from the top* ➥ | |
| | → ✱ **lowup** → ✱ *second from the top* ➥ → ✱ **Plots** | |
| 23.3 | ✱ **section** → ✱ *top* ➥ → ✱ **lowup** → ✱ *second from the top* ➥ | |
| | → ✱ **Add** → ✱ **Continue** | |
| 23.1 | ✱ **Post Hoc** | |
| 23.4 | ✱ **section** → ✱ ➥ → ✱ **Bonferroni** → ✱ **Tukey** → ✱ **Continue** | |
| 23.1 | ✱ **Options** | |
| 23.5 | ✱ **section** → ✱ ➥ → ✱ **lowup** → ✱ ➥ | |
| | → ✱ **Descriptive statistics** → ✱ **Estimates of effect size** | |
| | → ✱ **Observed power** → ✱ **Parameter estimates** | |
| | → ✱ **Homogeneity tests** → ✱ **Continue** | |
| 23.1 | ✱ **OK** | Back1 |

Upon completion of Step 5, 5a, or 5b, the output screen will appear (Screen 1, inside back cover). All results from the just-completed analysis are included in the Output Navigator. Make use of the scroll bar arrows (▲▼▶◀) to view the results. Even when viewing output, the standard menu of commands is still listed across the top of the window. Further analyses may be conducted without returning to the data screen.

## 23.2 Printing Results

Results of the analysis (or analyses) that have just been conducted require a window that displays the standard commands (**File Edit Data Transform Analyze...**) across the top. A typical print procedure is shown in the following page beginning with the standard output screen (Screen 1, inside back cover).

*To print results, from the Output screen perform the following sequence of steps:*

| In Screen | Do This | Step 6 |
|---|---|---|
| Back1 | *Select desired output or edit (see pages 16–22)* → ✳ **File** → ✳ **Print** | |
| 2.9 | *Consider and select desired print options offered, then* → ✳ **OK** | |

*To exit you may begin from any screen that shows the* **File** *command at the top.*

| In Screen | Do This | Step 7 |
|---|---|---|
| Menu | ✳ **File** → ✳ **Exit** | 2.1 |

Note: After clicking **Exit**, there will frequently be small windows that appear asking if you wish to save or change anything. Simply click each appropriate response.

# 23.3 Output

## 23.3.1 General Linear Models: Multivariate Analysis of Variance and Covariance (MANOVA and MANCOVA)

The following is partial output from sequence Steps 5a and 5b on (pages 303–304). We abbreviated the following output, because the complete printout takes 23 pages. Some of the sections of the report may not be present unless you specifically request them. Because of this, we list each section of the report separately, along with its interpretation. The sections labeled "General Interpretation" will be present in all reports. Also, because *post hoc* tests are only available if no covariates are included, sections dealing with *post hoc* tests and covariates do not both appear within a single output.

## 23.3.2 Between-Subjects Factors (General Interpretation)

**Between-Subjects Factors**

| | | Value Label | N |
|---|---|---|---|
| section | 1 | | 33 |
| | 2 | | 39 |
| | 3 | | 33 |
| lowup | 1 | lower | 22 |
| | 2 | upper | 83 |

This output simply lists each independent variable, along with the levels of each (with value labels) and the sample size (*N*).

## 23.3.3 Descriptive Statistics

**Descriptive Statistics**

| | section | lowup | Mean | Std. Deviation | N |
|---|---|---|---|---|---|
| quiz1 | 1 | Lower | 9.43 | .976 | 7 |
| | | Upper | 7.85 | 2.493 | 26 |
| | | Total | 8.18 | 2.338 | 33 |
| | 2 | Lower | 5.45 | 2.115 | 11 |

(*output continues*)

For each cell in the model, means, standard deviations, and sample size (*N*) are displayed. Output is given for each dependent variable in the model.

## 23.3.4 Box's Test of Equality of Covariance Matrices

| | |
|---|---|
| Box's M | 56.108 |
| F | 1.089 |
| df1 | 45 |
| df2 | 6109.950 |
| Sig. | .317 |

This statistic tests whether the covariance matrices for the dependent variables are significantly different. In our example, they are not ($p > .05$); if they did differ significantly, then we might believe that the covariance matrices are different. (That would be bad, because the equality of covariance matrices is an assumption of MANOVA.) This test is very sensitive, so just because it detects differences between the variance-covariance matrices does not necessarily mean that the *F* values are invalid.

**Multivariate Tests (General Interpretation)**

| Effect | | Value | F | Hypothesis df | Error df | Sig. | Partial Eta Squared | Observed Power |
|---|---|---|---|---|---|---|---|---|
| Intercept | Pillai's Trace | 0.948 | 347.556 | 5.000 | 95.000 | .000 | .948 | 1.000 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| section | Pillai's Trace | 0.127 | 1.299 | 10.000 | 192.000 | .233 | .063 | .657 |
| | Wilks' Lambda | 0.876 | 1.302 | 10.000 | 190.000 | .232 | .064 | .658 |
| | Hotelling's Trace | 0.139 | 1.305 | 10.000 | 188.000 | .231 | .065 | .659 |
| | Roy's Largest Root | 0.112 | 2.155 | 5.000 | 96.000 | .065 | .101 | .686 |
| lowup | Pillai's Trace | 0.031 | 0.598 | 5.000 | 95.000 | .702 | .031 | .210 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| section* lowup | Pillai's Trace | 0.234 | 2.540 | 10.000 | 192.000 | .007 | .117 | .949 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

Information presented in this section includes tests of each main effect and interaction possible in your design. The intercept refers to the remaining variance (usually the error variance). In this example, we can see that there is a significant interaction between **section** and **lowup** ($F$ (10, 192) = 2.54, $p < .01$). Note that if a covariate is included, it will be included in this table as well; in our example, our covariate **gpa** is not significantly related to the quiz scores.

| Term | Definition/Description |
|---|---|
| **VALUE** | Test names and values indicate several methods of testing for differences between the dependent variables due to the independent variables. Pillai's method is considered to be one of the more robust tests. |
| **F** | Estimate of the *F* value. |
| **HYPOTHESIS DF** | (Number of DV's − 1) × (Levels of $IV_1$ − 1) × (Levels of $IV_2$ − 1) × … For example, the degrees of freedom for the **section** × **lowup** effect is: (5 DV's − 1) × (3 levels of **section** −1) × (2 levels of **lowup** − 1) = 4 × 2 × 1 = 10 |
| **ERROR DF SIG.** | Calculated differently for different tests. *p* value (level of significance) for the *F*. |
| **PARTIAL ETA SQUARED** | The effect-size measure. This indicates how much of the total variance is explained by each main effect or interaction. In this case, for example, the **section** by **lowup** interaction accounts for 11.7% of the variance in the five quizzes, once the variance accounted for by the other main effects and interactions have been removed. |
| **OBSERVED POWER** | The probability that a result will be significant in a sample drawn from a population with an effect size equal to the effect size of your sample, and a sample size equal to the sample size of your sample. For example, if the effect size of **section** for the population of all classes was equal to .063 (the effect size in this sample), there is a 65.7% chance of finding that effect to be significant in a sample of 105 students (the sample size analyzed here). |

## 23.3.5 Levine's Test of Equality of Error Variances

**Levene's Test of Equality of Error Variances[a]**

| | F | df1 | df2 | Sig. |
|---|---|---|---|---|
| quizl | 2.560 | 5 | 99 | .032 |
| quiz2 | 1.101 | 5 | 99 | .365 |
| quiz3 | 1.780 | 5 | 99 | .124 |
| quiz4 | 2.287 | 5 | 99 | .052 |
| quiz5 | .912 | 5 | 99 | .477 |

This test examines the assumption that the variance of each dependent variable is the same as the variance of all other dependent variables. Levene's test does this by doing an ANOVA on the *differences* between each case and the mean for that variable, rather than for the *value* of that variable itself. In our example, **quiz1** is significant, so we should interpret our results with caution (but the *F* isn't large, so we don't need to panic just yet).

## 23.3.6 Tests of Between-Subjects Effects (General Interpretation)

In addition to the multivariate tests that the general linear model procedure performs, it also does simple *univariate F* tests on each of the dependent variables. Although this procedure does not have the main advantage of MANOVA—examining all of the dependent variables simultaneously—it is often helpful in interpreting results from MANOVA.

| Source | Dependent Variable | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared | Observed Power |
|---|---|---|---|---|---|---|---|---|
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| lowup | quiz1 | 0.071 | 1 | 0.071 | 0.013 | .909 | .000 | .051 |
| | quiz2 | 0.611 | 1 | 0.611 | 0.239 | .626 | .002 | .077 |
| | quiz3 | 0.504 | 1 | 0.504 | 0.106 | .746 | .001 | .062 |
| | quiz4 | 2.653 | 1 | 2.653 | 0.541 | .464 | .005 | .113 |
| | quiz5 | 0.314 | 1 | 0.314 | 0.113 | .737 | .001 | .063 |
| section* lowup | quiz1 | 73.236 | 2 | 36.618 | 6.796 | .002 | .121 | .912 |
| | quiz2 | 10.356 | 2 | 5.178 | 2.027 | .137 | .039 | .409 |
| | quiz3 | 45.117 | 2 | 22.559 | 4.744 | .011 | .087 | .780 |
| | quiz4 | 37.712 | 2 | 18.856 | 3.842 | .025 | .072 | .685 |
| | quiz5 | 39.870 | 2 | 19.935 | 7.213 | .001 | .127 | .928 |
| Error | quiz1 | 533.462 | 99 | 5.389 | | | | |
| | quiz2 | 252.945 | 99 | 2.555 | | | | |
| | quiz3 | 470.733 | 99 | 4.755 | | | | |
| | quiz4 | 485.846 | 99 | 4.908 | | | | |
| | quiz5 | 273.593 | 99 | 2.764 | | | | |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

In this abbreviated output, we list ANOVA results for **lowup** and **section × lowup**; we can see that there is a significant univariate interaction of **section** by **lowup** on **quiz1, quiz3, quiz4**, and **quiz5**. Note that it is possible to have one or more significant univariate tests on an effect without the multivariate effect being significant, or for the multivariate test to be significant on an effect without any of the univariate tests reaching significance. If your analysis includes any covariates, they will be listed here as well, and ANCOVA will be performed on each dependent variable.

| Term | Definition/Description |
|---|---|
| **SUM OF SQUARES** | For each main effect and interaction, the between-groups sum of squares; the sum of squared deviations between the grand mean and each group mean, weighted (multiplied) by the number of subjects in each group. For the error term, the within-groups sum of squares; the sum of squared deviations between the mean for each group and the observed values of each subject within that group. |
| **DF** | Degrees of freedom. For main effects and interactions, DF = (Levels of $IV_1 - 1$) × (Levels of $IV_2 - 1$) × . . . ; for the error term, DF = *N* − DF for main effects and interactions − 1. |
| **MEAN SQUARE** | Sum of squares for that main effect or interaction (or for the error term) divided by degrees of freedom. |
| **F** | Hypothesis mean square divided by the error mean square. |
| **SIG.** | *p* value (level of significance) for the *F*. |
| **PARTIAL ETA SQUARED** | The univariate effect-size measure reported for each dependent variable. This effect size indicates how much of the error variance (total variance minus the other main effects and interaction effects) is explained by the independent variable. |
| **OBSERVED POWER** | The probability that a result will be significant in a sample drawn from a population with an effect size equal to the effect size of your sample, and a sample size equal to the sample size of your sample. |

## 23.3.7 Parameter Estimates

**Parameter Estimates**

| | | | | | | 95% Confidence Interval | | |
|---|---|---|---|---|---|---|---|---|
| **Dep Variable** | **Parameter** | **B** | **Std. Error** | **t** | **Sig.** | **Lower Bound** | **Upper Bound** | **Partial Eta Squared** |
| quiz1 | Intercept | 4.647 | .883 | 5.262 | .000 | 2.895 | 6.400 | .220 |
| | gpa | .796 | .303 | 2.627 | .010 | .195 | 1.398 | .066 |
| | [section = 1] | .807 | .623 | 1.295 | .198 | −.430 | 2.044 | .017 |

*(table continues)*

Note that the values (above) from sequence 5b include the influence of the covariate **gpa**. Although we have only shown a portion of the total listing of parameter estimates, many more parameters will be listed to fully describe the underlying General Linear Model. Because the independent variables are typically interpreted through examination of the main effects and interactions, rather than the parameter estimates, we limit our description to the covariate (**gpa**). The table above shows only the effects of **gpa** on **quiz1**; the full table includes all dependent variables.

| Term | Definition/Description |
|---|---|
| **B** | The coefficient for the covariate in the model. |
| **STANDARD ERROR** | A measure of the stability of the *B* values. It is the standard deviation of the *B* value given a large number of samples drawn from the same population. |
| **t** | *B* divided by the standard error of *B*. |
| **SIG.** | Significance of *t*; the probability that these *t* values could occur by chance; the probability that *B* is not significantly different from zero. Because this *B* is significant, we know that gpa did have a significant effect on quiz1. |
| **LOWER AND UPPER 95% CONFIDENCE LIMITS** | Based on the *B* and the standard error, these values indicate that there is (in this example) a 95% chance that *B* is between .195 and 1.398. |
| **PARTIAL ETA SQUARED** | The effect-size measure. This indicates how much of the total variance is explained by the independent variable when removing the effect of the other independent variables and interactions between them. |

## 23.3.8 Estimated Marginal Means

**2. Lower or upper division**

| Dependent Variable | Lower or Upper Division | Mean | Std. Error | 95% Confidence Interval | |
|---|---|---|---|---|---|
| | | | | Lower Bound | Upper Bound |
| **quiz1** | lower | 7.628 | .538 | 6.560 | 8.696 |
| | upper | 7.560 | .255 | 7.053 | 8.066 |
| **quiz2** | lower | | | | |

(*table continues for the other four quizzes*)

For each of the dependent variables, marginal means and standard errors are given for each level of the independent variables. Standard error is the standard deviation divided by the square root of *N*.

## 23.3.9 Post Hoc Tests

**Multiple Comparison**

| Dependent Variable | | (I) Section | (J) Section | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Lower Bound | Upper Bound |
| quiz1 | Tukey HSD | 1 | 2 | .80 | .549 | .319 | –.51 | 2.10 |
| | | | 3 | 1.33 | .571 | .056 | –.03 | 2.69 |
| | | 2 | 1 | –.80 | .549 | .319 | –2.10 | .51 |
| | | | 3 | .54 | .549 | .593 | –.77 | 1.84 |
| | | 3 | 1 | –1.33 | .571 | .056 | –2.69 | .03 |
| | | | 2 | –.54 | .549 | .593 | –1.84 | .77 |
| | Bonferroni | 1 | 2 | .80 | .549 | .449 | –.54 | 2.13 |
| | | | 3 | 1.33 | .571 | .065 | –.06 | 2.73 |
| | | 2 | 1 | –.80 | .549 | .449 | –2.13 | .54 |
| | | | 3 | .54 | .549 | .994 | –.80 | 1.87 |
| | | 3 | 1 | –1.33 | .571 | .065 | –2.73 | .06 |
| | | | 2 | –.54 | .549 | .994 | –1.87 | .80 |

For each dependent variable, *post hoc* tests are computed. There are 18 *post hoc* tests you may select from; we show here only two of the most popular tests, the Tukey's HSD and the Bonferroni. Pairwise comparisons are computed for all combinations of levels of the independent variable (in this example, sections 1 and 2, sections 1 and 3, sections 2 and 3, etc.). SPSS displays the difference between the two means, the standard error of that difference, as well as whether that difference is significant and the 95% confidence interval of the difference. In this example, there are no significant differences between the different sections, but both the Tukey test and the Bonferroni suggest that sections 1 and 3 nearly reach significance at the .05 level.

# Exercises

Answers to selected exercises are downloadable at **www.spss-step-by-step.net.**

1. Using the **grade.sav** file, compute and interpret a MANOVA examining the effect of whether or not students completed the extra credit project on the total points for the class and the previous GPA.
2. Using the **grades.sav** file, compute and interpret a MANOVA examining the effects of **section** and **lowup** on **total** and **GPA**.
3. Why would it be a bad idea to compute a MANOVA examining the effects of **section** and **lowup** on **total** and **percent**?
4. A researcher wishes to examine the effects of high- or low-stress situations on a test of cognitive performance and self-esteem levels. Participants are also divided into those with high- or low-coping skills. The data are shown after question 5 (ignore the last column for now). Perform and interpret a MANOVA examining the effects of stress level and coping skills on both cognitive performance and self-esteem level.
5. Coping skills may be correlated with immune response. Include immune response levels (listed below) in the MANOVA performed for question 4. What do these results mean? In what way are they different than the results in question 4? Why?

| Stress Level | Coping Skills | Cognitive Performance | Self-Esteem | Immune Response |
|---|---|---|---|---|
| High | High | 6 | 19 | 21 |
| Low | High | 5 | 18 | 21 |
| High | High | 5 | 14 | 22 |
| High | Low | 3 | 8 | 15 |
| Low | High | 7 | 20 | 22 |
| High | Low | 4 | 8 | 17 |
| High | High | 6 | 15 | 28 |
| High | Low | 5 | 7 | 19 |
| Low | Low | 5 | 20 | 16 |
| Low | Low | 5 | 17 | 18 |

## Chapter 24

# G.L.M.: Repeated-Measures MANOVA

THE PREVIOUS chapter discussed designs with more than one dependent variable, and multiple independent variables with two or more levels each (MANOVA and MANCOVA). These analyses have all involved between-subjects designs, in which each subject is tested in only one level of the independent variable(s). There may be times, however, when a within-subjects design is more appropriate. Each subject is tested for more than one level of the independent variable or variables in a within-subjects or repeated-measures design.

This chapter will describe three different within-subjects procedures that the **General Linear Model—Repeated Measures** procedure may compute. The first is a completely within-subjects design, in which each subject experiences every experimental condition and produces values for each cell in the design. Mixed-design analyses are then described, in which one or more of the independent variables are within subjects, and one or more are between subjects. This example also includes a covariate. Finally, doubly multivariate designs are discussed. Doubly multivariate designs are similar to standard within-subjects designs, except that there are multiple dependent variables tested within subjects.

The **grades.sav** file is again the example. In this chapter, however—in order to demonstrate a within-subjects design—the meaning of **quiz1** through **quiz4** will be (perhaps somewhat capriciously) redefined. In particular, instead of being the scores on four different quizzes, **quiz1** through **quiz4** will refer to scores on the same quiz taken under four different conditions (or, alternatively, equivalent tests taken under four different conditions). These four conditions are based on a 2 (colors of paper) by 2 (colors of ink) within-subjects design. Students are given the same quiz (or equivalent quizzes) on either blue or red paper, printed in either green or black ink. This design produces the following combinations of paper colors and ink colors, with **quiz1** through **quiz4** assigned to each of the cells as noted:

| | **Ink Color** | |
|---|---|---|
| **Paper Color** | **Green** | **Black** |
| **Blue** | Blue paper, green ink<br>**quiz1** | Blue paper, black ink<br>**quiz2** |
| **Red** | Red paper, green ink<br>**quiz3** | Red paper, black ink<br>**quiz4** |

Because this chapter is an extension of the previous chapter, some of the advanced options and output refer you to the previous chapter. Also, if you don't have a basic understanding of between-subjects MANOVA or MANCOVA, you should read the previous chapter before trying to interpret the output from this chapter.

# 24.1 Step by Step

## 24.1.1 General Linear Models: Within-Subjects and Repeated-Measures MANOVA

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | 🖱 ⊞ → 🖱 All Programs ▷ → 🖱 IBM SPSS Statistics → 🖱 IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | 🖱 🚀 → 🖱 ■ → 🖱 Σα | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| | Step 2 |
|---|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* | |

*Screens 1 and 2 (on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **grades.sav** *file for further analyses:*

| In Screen | Do This | Step 3 |
|---|---|---|
| Front1 | 🖱 **File** → **Open** → 🖱 **Data** [*or* 🖱 📂] | |
| Front2 | **type** `grades.sav` → 🖱 **Open** [*or* 🖱 🖱 **grades.sav**] | Data |

Whether first entering SPSS or returning from earlier operations the standard menu of commands across the top is required. As long as it is visible you may perform any analyses. It is not necessary for the data window to be visible.

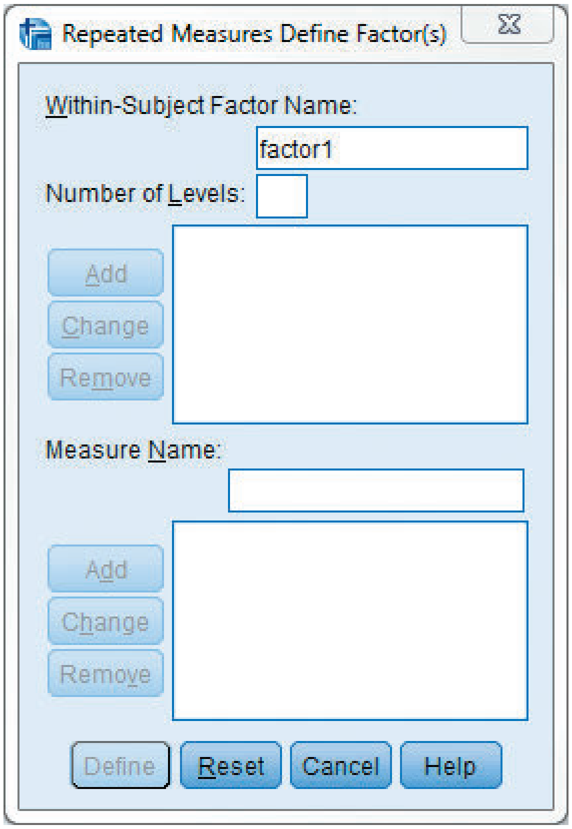
*After completion of Step 3 a screen with the desired menu bar appears. When you click a command (from the menu bar), a series of options will appear (usually) below the selected command. With each new set of options, click the desired item. The sequence to access general linear models begins at any screen with the menu of commands visible:*

| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | 🖱 **Analyze** → **General Linear Model** → 🖱 **Repeated Measures** | 24.1 |

After this sequence step is performed, a window opens (Screen 24.1, following page) in which you specify the names of the within-subjects independent variables. The names that you type in the **Within-Subject Factor Name** box are *not* names of variables that currently exist in your data file; they are *new* variable names that will only be used by the **Repeated Measures** command.

In our example, we define two independent variables as **papercol** (the color of the paper on which the quiz was printed) and **inkcolor** (the color of the ink with which the quiz was printed). To specify each within-subjects variable name, first type the name of the variable in the **Within-Subject Factor Name** box. Then, type the number of levels of that independent variable in the **Number of Levels** box, and click the **Add** button. If you wish to do doubly multivariate ANOVA, in which there are multiple dependent variables for each level of the within-subjects design, then you will use the lower portion of the dialog box. In the **Measure Name** box, you can then type the name of each dependent variable used in the analysis and click on the **Add** button when finished.

Screen 24.1 General Linear Model—Repeated Measures Define Factor(s) Dialog Window



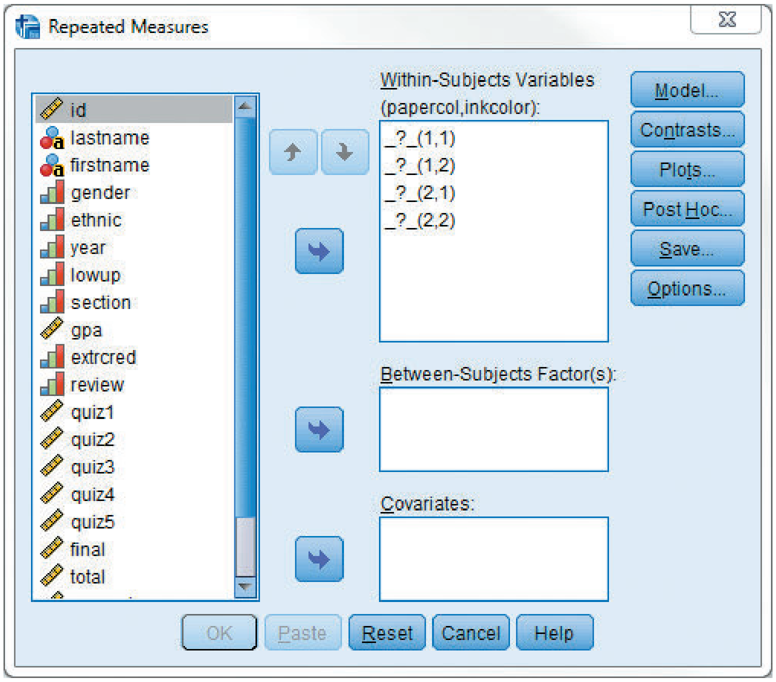
After you have specified the **Within-Subject Factor** name or names, and (optionally) entered the dependent variable names for a doubly multivariate design, click on the **Define** button to specify the rest of your analysis.

Screen 24.2 now appears (following page). The box to the left of the window contains all variables in your data file. The **Within-Subjects Variables** box, in the top center of the window, lets you specify what variable in your data file represents each level of each within-subjects variable in the analysis. In this case, we have defined **papercol** and **inkcolor** (paper color and ink color), with two levels each, so we need to specify four variables from the box on the left, and match them up with the cells specified in the **Within-Subjects Variables** box. The numbers in parenthesis in that box refer to the cells in the design. By way of explanation, observe the chart that follows:

| Quiz number | Cell | Paper color | Ink color |
|---|---|---|---|
| 1 | (1,1) | blue = 1 | green = 1 |
| 2 | (1,2) | blue = 1 | black = 2 |
| 3 | (2,1) | red = 2 | green = 1 |
| 4 | (2,2) | red = 2 | black = 2 |

Note that the ink color changes more rapidly than paper color as we move from **quiz1** to **quiz4**. Here we demonstrate with two variables with two levels each, but the same rationale applies with a greater number of variables that have more than two levels. If your design is doubly multivariate, with more than one dependent measure for each cell in the within-subjects design, then those multiple dependent measures will be listed in the **Within-Subjects Variables** box.

Screen 24.2 General Linear Model—Repeated Measures Dialog Window



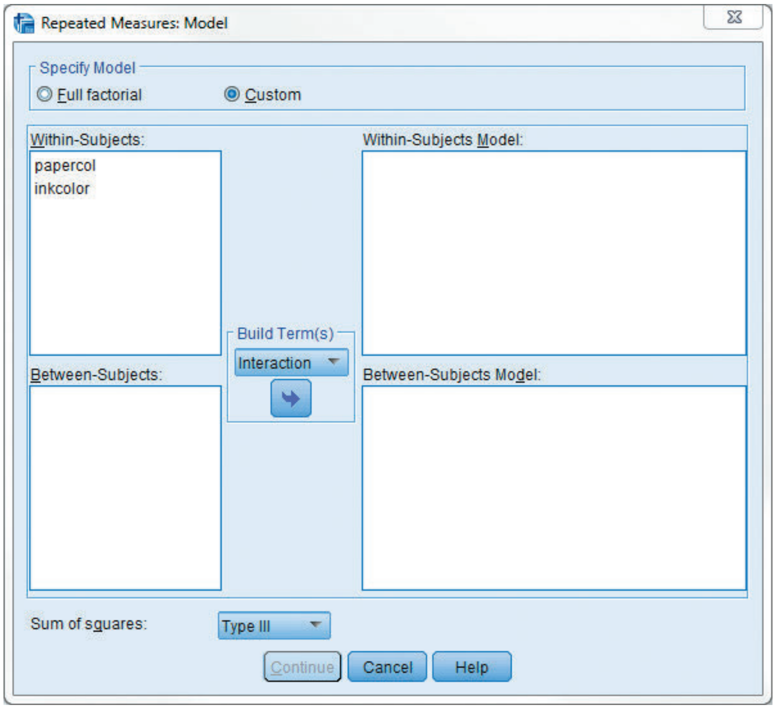
For each variable in the within-subjects analysis, select the variable in the box to the left of the window, and click on the top ➡. By default, SPSS assumes that the first variable you select will go in the top cell in the **Within-Subjects Variables** box [cell (1,1) in our example], the second variable you select will go in the second cell in the **Within-Subjects Variables** box [(1,2) in our example], and so forth. If a variable does not go in the proper cell in the **Within-subject Variables** box, then click on the up and down arrow buttons (↑↓) to move the variable up and down the cell list.

Once the within-subjects dependent variables have been specified, define any between-subjects variables or covariates in your model. Select any variables in the variable list box to the left of Screen 24.2 that you wish to treat as between-subjects variables or covariates, and click on the appropriate ➡ button to move it into the **Between-Subjects Factor(s)** box in the middle of the window or the **Covariates** box near the bottom of the window.

Six buttons to the right of Screen 24.2 may be selected: **Model, Contrasts, Plots, Post Hoc, Save**, and **Options**. The **Options** button produces a large array of selections and is identical to Screen 23.5 in the previous chapter; see that description (page 302) for those choices. The **Plots** button functions identically to the **Plots** button in the previous chapter (see page 300), but now you can plot both within-subjects and between-subjects factors. The **Post Hoc** button also works similarly to the way it worked in the previous chapter (page 301), but in this case it is important to remember that you can only select *post hoc* tests for between-subjects factors. The **Contrasts** and **Save** buttons are beyond the scope of this book, and will not be discussed.

The **Model** button calls up Screen 24.3 (following page). In most cases, a **Full Factorial** model is desired: This produces a model with all main effects and all interactions included. At times, you may wish to test only certain main effects and interactions. In this case, select the **Custom** button, and move any main effects (chosen by selecting one variable) and interactions (chosen by selecting two or more variables) desired in the model from the left boxes to the right boxes, by clicking the ➡ button. Variables and interactions moved from the **Within-Subjects** box at the left go to the **Within-Subjects Model** box on the right, and variables and interactions moved from the **Between-Subjects** box on the left go to the **Between-Subjects Model** on the right; you cannot specify interactions between within-subjects and between-subjects variables.

**Screen 24.3** General Linear Model—Repeated Measures Model Specification Dialog Window



Once you have specified the model, click **Continue**, and then click the **OK** to calculate the within-subjects or repeated-measures MANOVA or MANCOVA.

Three different versions of sequence Step 5 will be included in this chapter; the first demonstrates within-subjects MANOVA, the second includes both a within- and a between-subjects variable along with a covariate, and the third a doubly multivariate design.

*This step instructs SPSS to perform a 2 (color of paper) x 2 (color of ink) within-subjects analysis of variance, with* **quiz1** *through* **quiz4** *referring to the cells of the model as presented in the table in the introductory section of this chapter.*

| In Screen | Do This | Step 5 |
|---|---|---|
| 24.1 | ✳ **factor1** → *use* **DELETE** *and* **BACKSPACE** *keys to delete text from the box* | |
| | → **type** `papercol` → **press** **TAB** → **type** 2 → ✳ **Add** | |
| | **type** `inkcolor` → **press** **TAB** → **type** 2 → ✳ **Add** → ✳ **Define** | |
| 24.2 | ✳ **quiz1** → *hold down* **Shift** *and* ✳ **quiz4** → ✳ *top* ➡ → ✳ **Options** | |
| 23.5 | ✳ **Descriptive statistics** → ✳ **Estimates of effect size** → ✳ **Continue** | |
| 24.2 | ✳ **OK** | **Back1** |

*The following step illustrates a mixed design, in which some of the factors are within-subjects and some are between-subjects. In this case a 2 (ink color) x 2 (gender of subject) MANOVA is performed, using* **gpa** *as a covariate.*

| In Screen | Do This | Step 5a |
|---|---|---|
| 24.1 | 🖱 **factor1** → *use* **DELETE** *and* **BACKSPACE** *keys to delete text from the box* | |
| | → type `inkcolor` → press **TAB** → type `2` → 🖱 **Add** → 🖱 **Define** | |
| 24.2 | 🖱 **quiz1** → *hold down* **Shift** *and* 🖱 **quiz2** → 🖱 *top* ➡ | |
| | 🖱 **gender** → 🖱 *middle* ➡ → 🖱 **gpa** → 🖱 *bottom* ➡ → | |
| | 🖱 **Options** | |
| 23.5 | 🖱 **Descriptive statistics** → 🖱 **Estimates of effect size** → 🖱 **Continue** | |
| 24.2 | 🖱 **OK** | Back1 |

*The following step illustrates a doubly multivariate design, in which there are two or more dependent variables measured in different levels of one or more within-subjects factors. In this example,***quiz1** *and* **quiz2** *are short-answer quizzes, and* **quiz3** *and* **quiz4** *are multiple-choice quizzes.* **quiz1** *and* **quiz3** *were taken in the green-ink color condition, and* **quiz2** *and* **quiz4** *were taken in the black-ink color condition.*

| In Screen | Do This | Step 5b |
|---|---|---|
| 24.1 | 🖱 **factor1** → *use* **DELETE** *and* **BACKSPACE** *keys to delete text from the box* → | |
| | type `inkcolor` → press **TAB** → type `2` → 🖱 **Add** → 🖱 *in the* | |
| | **Measure Name** *box* type `sa` → 🖱 **Add** → type `mc` → 🖱 **Add** → | |
| | 🖱 **Define** | |
| 24.2 | 🖱 **quiz1** → *hold down* **Shift** *and* 🖱 **quiz4** → 🖱 *top* ➡ → 🖱 **Options** | |
| 23.5 | 🖱 **Descriptive statistics** → 🖱 **Estimates of effect size** → 🖱 **Continue** | |
| 24.2 | 🖱 **OK** | Back1 |

Upon completion of Step 5, 5a, or 5b, the output screen will appear (Screen 1, inside back cover). All results from the just-completed analysis are included in the Output Navigator. Make use of the scroll bar arrows (▲▼▶◀) to view the results. Even when viewing output, the standard menu of commands is still listed across the top of the window. Further analyses may be conducted without returning to the data screen.

# 24.2 Printing Results

Results of the analysis (or analyses) that have just been conducted require a window that displays the standard commands (**File Edit Data Transform Analyze. . .**) across the top. A typical print procedure is shown below beginning with the standard output screen (Screen 1, inside back cover).

*To print results, from the Output screen perform the following sequence of steps:*

| In Screen | Do This | Step 6 |
|---|---|---|
| Back1 | *Select desired output or edit (see pages 16–22)* → ✳ **File** → ✳ **Print** | |
| 2.9 | *Consider and select desired print options offered, then* → ✳ **OK** | |

*To exit you may begin from any screen that shows the* File *command at the top.*

| In Screen | Do This | Step 7 |
|---|---|---|
| Menu | ✳ **File** → ✳ **Exit** | 2.1 |

Note: After clicking **Exit**, there will frequently be small windows that appear asking if you wish to save or change anything. Simply click each appropriate response.

# 24.3 Output

## 24.3.1 General Linear Models—Within-Subjects and Repeated-Measures MANOVA

Most of the output from the General Linear Models procedure using within-subjects designs is similar to the output for between-subjects designs (Chapter 23), so in order to conserve space, we present only representative portions of the complete output here. We have included all of the *essential* output and focus on material that is unique to within-subjects designs. Also, because there are three example analyses done in this chapter, each subheading will be followed by a note indicating which of the three types of analyses are applicable to that output section, as well as which step (Step 5, 5a, or 5b) is used to produce the sample output.

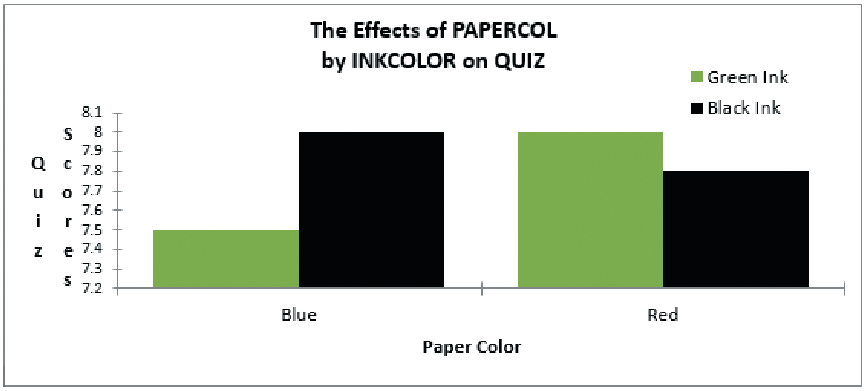
## 24.3.2 Multivariate Tests

- Applies to: Step 5 (Within-Subjects Designs), Step 5a (Mixed Designs), Step 5b (Doubly Multivariate Designs).
- Example from: Step 5 (Within-Subjects Designs).

**Multivariate Tests**

| Effect | | Value | F | Hypothesis df | Error df | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|---|
| papercol | Pillai's Trace | 0.027 | 2.866 | 1 | 104 | .093 | .027 |
| | Wilks' Lambda | 0.973 | 2.866 | 1 | 104 | .093 | .027 |
| | Hotelling's Trace | 0.028 | 2.866 | 1 | 104 | .093 | .027 |
| | Roy's Largest Root | 0.028 | 2.866 | 1 | 104 | .093 | .027 |
| inkcolor | Pillai's Trace | 0.020 | 2.124 | 1 | 104 | .148 | .020 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| papercol*<br>inkcolor | Pillai's Trace | 0.081 | 9.160 | 1 | 104 | .003 | .081 |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

SPSS will present information similar to that given here for each main effect and interaction possible in your design. In our example, the interaction effect of paper color by ink color is significant (*p* = .003), so the means of the cells must be examined to interpret the interaction effect, either by examining the means themselves or by performing additional *t* tests. In this case the means demonstrate an interaction shown in the figure (below): Black ink produces higher quiz scores than green ink, when the quiz is printed on blue paper; however, green ink produces higher quiz scores than black ink when the quiz is printed on red paper.



If you are analyzing a mixed design, the interaction effects of the within-subjects and between-subjects variables will also be produced. If you have covariates in the design, the interaction effects between your covariates and the within-subjects variables will also be given. Interpreting these interactions is tricky: You have to look for differences in the regression coefficients of your covariates for different levels of your within-subject variables.

| Term | Definition/Description |
|---|---|
| VALUE | Test names and values indicate several methods of testing for differences between the dependent variables due to the independent variables. Pillai's method is considered to be one of the more robust tests. |
| F | Estimate of the *F* value. |
| HYPOTHESIS DF | (Levels of IV$_1$ − 1) × (Levels of IV$_2$ − 1) × . . . For doubly multivariate designs, this is multiplied by (Number of DV's − 1). |
| ERROR DF | Calculated differently for different tests. |
| SIG. | *p* value (level of significance) for the *F*. |
| PARTIAL ETA SQUARED | The effect-size measure. This indicates how much of the variance in the dependent variable not accounted for by the other independent variables and interactions is explained by each main effect or interaction. In this case, for example, the papercol by inkcolor interaction accounts for 8.1% of the variance in the four quizzes (excluding the variance accounted for by the main effects of papercol and inkcol). |

## 24.3.3 Tests of Within-Subjects Contrasts

- Applies to: Step 5 (Within-Subjects Designs), Step 5a (Mixed Designs).
- Example from: Step 5 (Within-Subjects Designs).

**Tests of Within-Subjects Contrasts**

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|
| papercol | 2.917 | 1 | 2.917 | 2.866 | .093 | .027 |
| Error(papercol) | 105.833 | 104 | 1.018 | | | |
| inkcolor | 2.917 | 1 | 2.917 | 2.124 | .148 | .020 |
| Error(inkcolor) | 142.833 | 104 | 1.373 | | | |
| papercol*inkcolor | 12.688 | 1 | 12.688 | 9.160 | .003 | .081 |
| Error(papercol*inkcolor) | 144.062 | 104 | 1.385 | | | |

Most of this output is redundant with the Multivariate Tests described on the previous page. In this table, however, *F* values are calculated through using sum of squares and mean squares. Because most of the values in this output have already been described, only those values that are not presented earlier are defined here.

| Term | Definition/Description |
|---|---|
| **SUM OF SQUARES FOR MAIN EFFECTS AND INTERACTIONS** | The between-groups sum of squares; the sum of squared deviations between the grand mean and each group mean, weighted (multiplied) by the number of subjects in each group. |
| **SUM OF SQUARES FOR ERROR** | The within-groups sum of squares; the sum of squared deviations between the mean for each group and the observed values of each subject within that group. |
| **MEAN SQUARE FOR MAIN EFFECTS AND INTERACTIONS** | Hypothesis sum of squares divided by hypothesis degrees of freedom. Since there is only one degree of freedom for the hypothesis, this value is the same as hypothesis sum of squares. |
| **MEAN SQUARE FOR ERROR** | Error sum of squares divided by error degrees of freedom. |

## 24.3.4 Univariate Tests

- Applies to: Step 5b (Doubly Multivariate Designs).
- Example from: Step 5b (Doubly Multivariate Designs).

**Univariate Tests**

| Source | Measure | | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|---|---|
| Inkcolor | sa | Sphericity Assumed | 13.886 | 1 | 13 886 | 8.247 | .005 | .073 |
| | | Greenhouse-Geisser | 13.886 | 1.000 | 13.886 | 8.247 | .005 | .073 |
| | | Huynh-Feldt | 13.886 | 1.000 | 13.886 | 8.247 | .005 | .073 |
| | | Lower-bound | 13.886 | 1.000 | 13.886 | 8.247 | .005 | .073 |
| | mc | Sphericity Assumed | 1.719 | 1 | 1.719 | 1.599 | .209 | .015 |
| | | Greenhouse-Geisser | 1.719 | 1.000 | 1.719 | 1.599 | .209 | .015 |
| | | Huynh-Feldt | 1.719 | 1.000 | 1.719 | 1.599 | .209 | .015 |
| | | Lower-bound | 1.719 | 1.000 | 1.719 | 1.599 | .209 | .015 |
| Error (inkcolor) | sa | Sphericity Assumed | 175.114 | 104.00 | 1.684 | | | |
| | | Greenhouse-Geisser | 175.114 | 104.00 | 1.684 | | | |
| | | Huynh-Feldt | 175.114 | 104.00 | 1.684 | | | |
| | | Lower-bound | 175.114 | 104.00 | 1.684 | | | |
| | mc | Sphericity Assumed | 111.781 | 104 | 1.075 | | | |
| | | Greenhouse-Geisser | 111.781 | 104.00 | 1.075 | | | |
| | | Huynh-Feldt | 111.781 | 104.00 | 1.075 | | | |
| | | Lower-bound | 111.781 | 104.00 | 1.075 | | | |

For any doubly multivariate design (with more than one dependent variable), univariate *F* tests are performed for each main effect and interaction. In this example, there is a significant effect of ink color for essay exams, but not for multiple choice exams. Several different methods of calculating each test are used (Sphericity Assumed, Greenhouse-Geisser, etc.), but it is very rare for them to produce different results.

| Term | Definition/Description |
|---|---|
| **SUM OF SQUARES** | For each main effect and interaction, the between-groups sum of squares; for the error term, the within-groups sum of squares. |
| **DF** | Degrees of freedom. For main effects and interactions, DF = (Levels of $IV_1 - 1$) × (Levels of $IV_2 - 1$) × . . .; for the error term, DF = *N* − DF for main effects and interactions − 1. |
| **MEAN SQUARE** | Sum of squares for that main effect or interaction (or for the error term) divided by degrees of freedom. |
| **F** | Hypothesis mean square divided by the error mean square. |
| **SIG.** | *p* value (level of significance) for the *F*. |

### 24.3.5 Tests of Between-Subjects Effects

- Applies to: Step 5a (Mixed Designs).
- Example from: Step 5a (Mixed Designs).

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|
| Intercept | 489.802 | 1 | 489.802 | 73.411 | .000 | .419 |
| gpa | 53.083 | 1 | 53.033 | 7.956 | .006 | .072 |
| gender | .736 | 1 | .736 | .110 | .740 | .001 |
| Error | 680.548 | 102 | 6.672 | | | |

The between-subjects effects are listed together, along with the effects of any covariate. Most of this output is redundant with the Multivariate Tests, described on pages 317–318. In this table, however, as in the Tests of Within-Subjects Effects table, *F* values are calculated through using sum of squares and mean squares. All of the statistics reported in this table have been defined earlier in either the Multivariate Tests or the Tests of Within-Subjects Effects.

# Exercises

Answers to selected exercises are downloadable at **www.spss-step-by-step.net**.

1. Imagine that in the **grades.sav** file, the five quiz scores are actually the same quiz taken under different circumstances. Perform repeated-measures ANOVA on the five quiz scores. What do these results mean?

2. To the analysis in exercise 1, add whether or not students completed the extra credit project (**extrcred**) as a between-subjects variable. What do these results mean?

3. A researcher puts participants in a highly stressful situation (say, performing repeated-measures MANCOVA) and measures their cognitive performance. He then puts them in a low-stress situation (say, lying on the beach on a pleasant day). Participant scores on the test of cognitive performance are reported below. Perform and interpret a within-subjects ANOVA on these data.

| **Case Number:** | 1 | 2 | 3 | 4 | 5 | 5 | 6 | 7 | 8 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **High Stress:** | 76 | 89 | 86 | 85 | 62 | 63 | 85 | 115 | 87 | 85 |
| **Low Stress:** | 91 | 92 | 127 | 92 | 75 | 56 | 82 | 150 | 118 | 114 |

4. The researcher also collects data from the same participants on their coping ability. They scored (in case number order) 25, 9, 59, 16, 23, 10, 6, 43, 44, and 34. Perform and interpret a within-subjects ANCOVA on these data.

5. The researcher just discovered some more data. . .in this case, physical dexterity performance in the high-stress and low-stress situations (listed below, in the same case number order as in the previous two exercises). Perform and interpret a 2 (stress level: high, low) by 2 (kind of performance: cognitive, dexterity) ANCOVA on these data.

   Physical dexterity values:

| **Case Number:** | 1 | 2 | 3 | 4 | 5 | 5 | 6 | 7 | 8 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **High Stress:** | 91 | 109 | 94 | 99 | 73 | 76 | 94 | 136 | 109 | 94 |
| **Low Stress:** | 79 | 68 | 135 | 103 | 79 | 46 | 77 | 173 | 111 | 109 |

# Chapter 25
# Logistic Regression

LOGISTIC REGRESSION is an extension of multiple regression (Chapter 16) in which the dependent variable is not a continuous variable. In logistic regression, the dependent variable may have only two values. Usually these values refer to membership-nonmembership, inclusion-noninclusion, or yes-no.

Because logistic regression is an extension of multiple regression, we will assume that you are familiar with the fundamentals of multiple regression analysis: namely, that several variables are regressed onto another variable, using forward, backward, or other selection processes and criteria. These basic concepts are the same for logistic regression as for multiple regression. However, the meaning of the regression equation is somewhat different in logistic regression. In a standard regression equation, a number of weights are used with the predictor variables to predict a value of the criterion or dependent variable. In logistic regression the value that is being predicted represents a probability, and it varies between 0 and 1. In addition to this, it is possible to use a categorical predictor variable, using an indicator-variable coding scheme. This is described in the Step by Step and Output sections in more detail, but it essentially breaks up a single categorical predictor variable into a series of variables, each coded as 1 or 0 indicating whether or not the subjects are in a particular category.

At this point, the basic mathematics of logistic regression will be summarized, using the example that will be applied in this chapter. The example utilizes another helping file named **helping3.sav**. This is a file of real data (*N* = 537) that deals with issues similar to the **helping1.sav** file presented earlier, but variable names are different in several instances. It is described in greater detail in the Data Files appendix (page 364). In the **helping1.sav** file used in Chapter 16, the amount of help given to a friend was predicted by feelings of (a) sympathy aroused in the helper in response to the friend's need (**sympathy**); (b) feelings of anger or irritation aroused in the helper by the friend's need (**anger**); and (c) self-efficacy of the helper in relation to the friend's need (**efficacy**). In this case, however, instead of predicting the *amount* of help given to a friend, our model will predict a different dependent variable: whether the friend thought that the help given was useful or not. This is coded as a yes-or-no (dichotomous) variable. It should be noted that, although the rest of the data in the **helping3.sav** file are real, this categorical dependent was created for this example. Specifically, anyone who gave more than the average amount of help was coded "helpful" and anyone who gave less than the average amount of help was coded "unhelpful."

## Mathematics of Logistic Regression

If you want to understand logistic regression, then probabilities, odds, and the logarithm of the odds must be understood. Probabilities are simply the likelihood that something will happen; a probability of .20 of rain means that there is a 20% chance of rain. In the technical sense used here, odds are the ratio of the probability that an event will occur divided by the probability that an event will not occur. If there is a 20% chance of rain, then there is an 80% chance of no rain; the odds, then, are:

$$\text{Odds} = \frac{\text{prob(rain)}}{\text{prob(no rain)}} = \frac{.20}{.80} = \frac{1}{4} = .25$$

Although probabilities vary between 0 and 1, odds may be greater than 1: For instance, an 80% chance of rain has odds of .80/.20 = 4.0. A 50% chance of rain has odds of 1.

A key concept in logistic regression analysis is a construct known as a *logit*. A logit is the natural logarithm (*ln*) of the odds. If there is a 20% chance of rain, then there is a logit of:

$$\mathit{ln}\ (.25) = -1.386\ldots$$

In the example used in Chapter 16, the regression equation looked something like this:

$$\text{HELP} \ = \text{B}_0 + \text{B}_1 \times \text{SYMPATHY} + \text{B}_2 \times \text{ANGER} + \text{B}_3 \times \text{EFFICACY}$$

The amount of helping was a function of a constant, plus coefficients times the amount of sympathy, anger, and efficacy. In the example used in this chapter, in which we examine whether or not the subjects' help was useful instead of how much they helped, the regression equation might look something like this. Please note that we have now switched to the variable names used in the **helping3.sav** file: sympathy (**sympatht**), anger (**angert**), and efficacy (**effict**).

$$\mathit{In}\left[\frac{\text{prob(helping)}}{\text{prob(not helping)}}\right] = \text{B}_0 + \text{B}_1(\textbf{sympatht}) + \text{B}_2(\textbf{angert}) + \text{B}_3(\textbf{effict})$$

In this equation, the log of the odds of helping is a function of a constant, plus a series of weighted averages of sympathy, anger, and efficacy. If you wish to think in terms of the odds-of-helping or the probability-of-helping instead of the log-odds-of-helping, this equation may be converted to the following:

$$\frac{\text{prob(helping)}}{\text{prob(not helping)}} = e^{\text{B}_0} \times e^{\text{B}_1(\textbf{sympatht})} \times e^{\text{B}_2(\textbf{angert})} \times e^{\text{B}_3(\textbf{effict})}$$

or

$$\text{prob(helping)} = \frac{1}{1 + e^{-\text{B}_0} \times e^{-\text{B}_1(\textbf{sympath})} \times e^{-\text{B}_2(\textbf{anger})} \times e^{-\text{B}_3(\textbf{effict})}}$$

This equation is probably not very intuitive to most people (several of my students would certainly agree with this!); it takes a lot of experience before interpreting logistic regression equations becomes intuitive. Because of this as well as other problems in selecting an appropriate model (since model selection involves both mathematical and theoretical considerations), you should use extreme caution in interpreting logistic regression models.

Due to this complexity we refer you to three sources if you wish to gain a greater understanding of logistic regression: The SPSS Manuals cover logistic regression in much greater detail than we do here, and then there are textbooks by McLachlan (2004) and Wickens (1989), both of which are quite good.

# 25.1 Step by Step

## 25.1.1 Logistic Regression

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | → All Programs ▷ → IBM SPSS Statistics → IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | ✱ 🚀 → ✱ ■ → ✱ Σ | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| | Step 2 |
|---|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* | |

*Screens 1 and 2 (on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **helping3.sav** *file for further analyses:*

| In Screen | Do This | | Step 3 |
|---|---|---|---|
| Front1 | ✱ **File** → **Open** → ✱ **Data** | [*or* ✱ 📂] | |
| Front2 | **type** `helping3.sav` → ✱ **Open** | [*or* ✱ ✱ **helping3.sav**] | Data |

Whether first entering SPSS or returning from earlier operations the standard menu of commands across the top is required. As long as it is visible you may perform any analyses. It is not necessary for the data window to be visible.

*After completion of Step 3 a screen with the desired menu bar appears. When you click a command (from the menu bar), a series of options will appear (usually) below the selected command. With each new set of options, click the desired item. The sequence to access logistic regression begins at any screen with the menu of commands visible:*

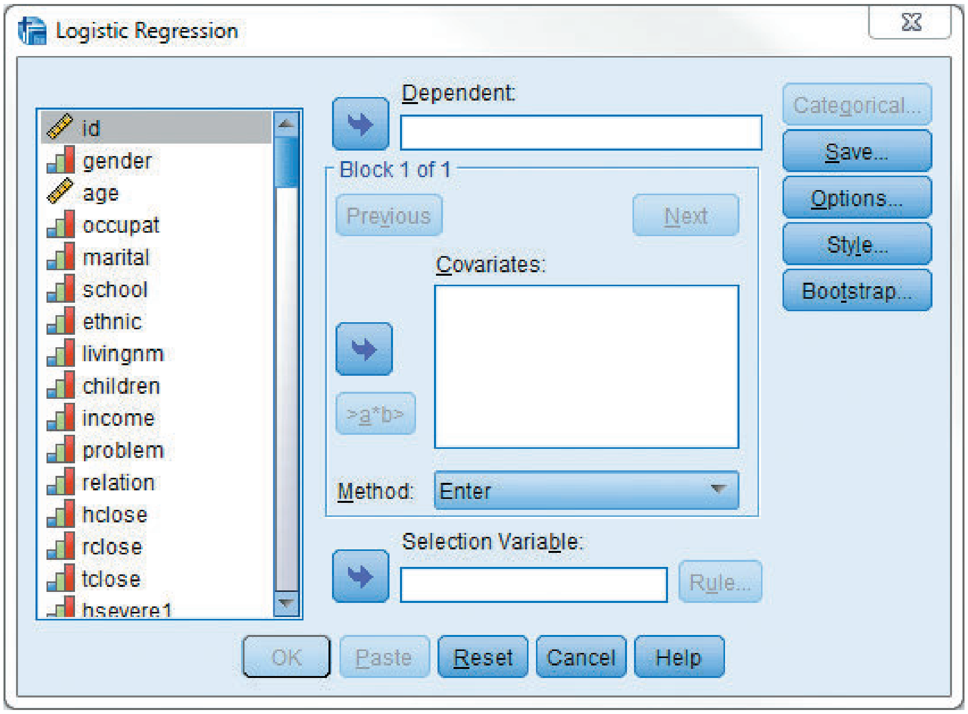
| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | ✱ **Analyze** → **Regression** → ✱ **Binary Logistic** | 25.1 |

After this sequence step is performed, a window opens (Screen 25.1, following page) in which you select the predicted variable, predictor variables, and other specifications that allow you to do a logistic regression analysis. On the left side of the window is the list of all of the variables in the file; in the top center of the window is the **Dependent** box. In this box, you will place the categorical variable you wish to predict based on your other variables. In our example, we will use the variable **cathelp**, which equals 1 if help was *not* rated as useful and 2 if it *was* rated useful. Because logistic regression uses 0s and 1s (instead of 1s and 2s) to code the dependent variable, SPSS will recode the variable from 1 and 2 to 0 and 1. It does this automatically; you don't have to worry about it.

In the center of the window is the **Covariates** box. In this box, you will enter the predictor variables by clicking on each variable that you want to include in the analysis, and then clicking on the ➔ button. It is possible to enter interaction terms into the regression equation, by selecting all of the variables that you want in the interaction term and then clicking on the >a*b> button. It is often difficult to interpret interaction terms, so you probably shouldn't use interactions unless you are familiar with them, and you have strong theoretical rationale for including them.

The **Method** specifies the way that SPSS will build the regression equation. **Enter**, shown in Screen 25.1, tells SPSS to build the equation by entering all of the variables at once, whether or not they significantly relate to the dependent variable. In **Forward: LR**, SPSS builds the equation by entering variables one at a time, using likelihood

Screen 25.1 Logistic Regression Dialog Window



Method: Enter

ratio estimates to determine which variable will add the most to the regression equation. In the **Backward: LR** method, SPSS builds the equation starting with all variables and then removes them one by one if they do not contribute enough to the regression equation. These are the most commonly used methods.

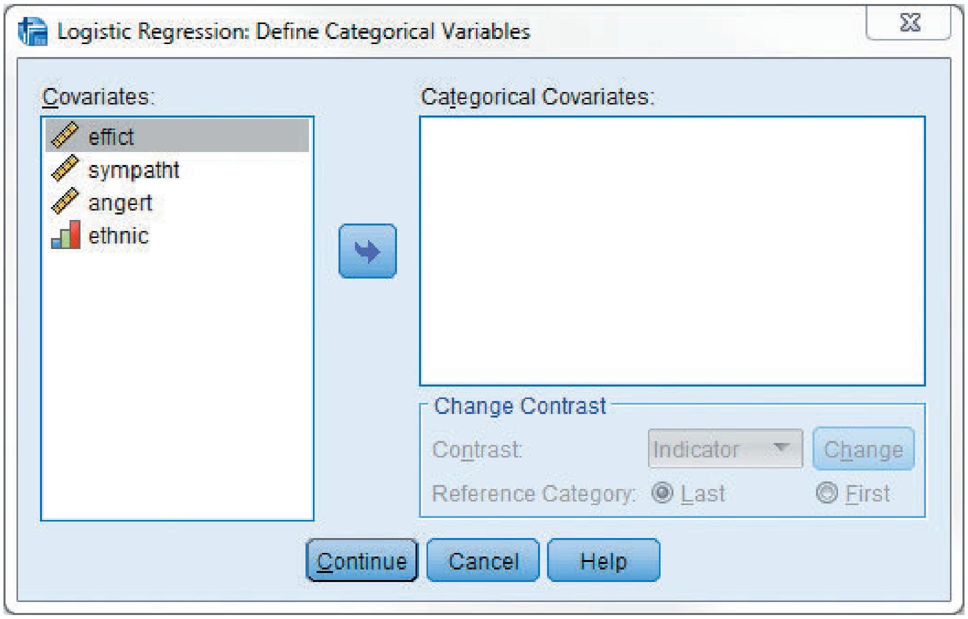
If you wish to use a different method for different variables, then you will need to use the **Previous** and **Next** buttons to set up different blocks of variables. For each block you set up, you can enter several variables and the method that you want SPSS to use to enter those variables (SPSS will analyze one block at a time and then move on to the next block). Then, click the **Next** button to set up the next block. Different blocks can have different methods. In our example, we will only use a single block and use the **Forward: LR** method.

The **Selection Variable** option operates like the **If** procedure described in Chapter 4. If, for instance, you paste **gender** into the Selection Variable window, the Value button would brighten up. A click on this button opens a tiny dialog window that allows you to select which level of the variable (in this case, 1 = women, 2 = men) you wish to select for the procedure. The analysis will then be conducted with the selected subjects or cases. This selection may be reversed later if you wish.

Once you have set up your Covariates (predictor variables), you need to select the **Categorical** button if a categorical predictor variable is included. In our examples, individuals may be White, Black, Hispanic, Asian, or other. Because there is no particular order to these ethnicities, the **Categorical** button is clicked and Screen 25.2 appears (following page).

In the **Covariates** box to the left are all of the predictor variables that were entered in the **Logistic Regression** window (Screen 25.1). For each of these variables that are categorical, select the variable and click on the ➡ button to move it into the **Categorical Covariates** box. It is possible to handle the categorical variable in several ways,

**Screen 25.2** Logistic Regression: Categorical Variables Dialog Window



Logistic Regression: Define Categorical Variables

Covariates:
- effict
- sympatht
- angert
- ethnic

Categorical Covariates:

Change Contrast

Contrast: Indicator

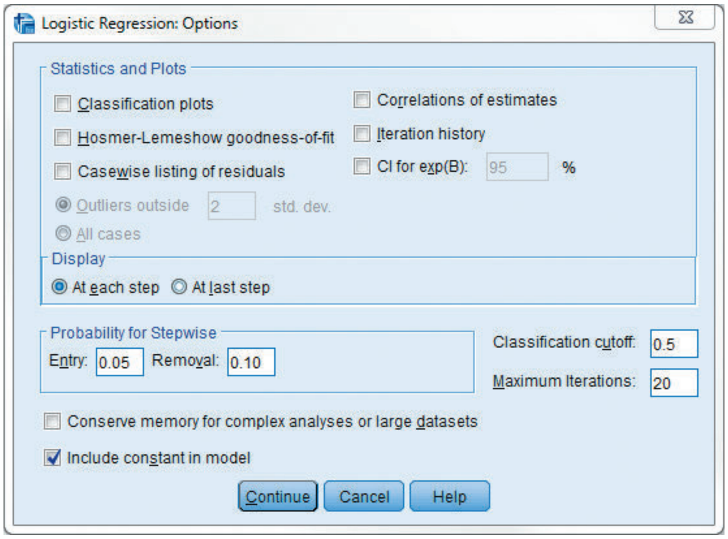
Reference Category: ● Last ○ First

by setting the **Contrast** in the **Change Contrast** box in the lower right of Screen 25.2. We will limit our discussion to the **Deviation** contrast (**Indicator** is the default), which converts this simple categorical variable into a series of variables (each of them a contrast between one of the first variables and the final variable) that may be entered into the logistic regression equation. Note that the **Change Contrast** box allows you to change the current contrast to a different contrast, required, of course in the shift from **Indicator** to **Deviation**.

The **Options** button (Screen 25.1) allows you to select a number of additional options that are often useful. Upon clicking this button, Screen 25.3 (below) appears that allows you to select which additional statistics and plots you wish, as well as the probability for entry and removal of variables when performing stepwise regression.

Clicking on the **Classification plots** box will produce a graph in which each case is plotted along its predicted probability from the regression equation produced, indicating its classification based on the predicted variable in the original data. This provides a graphical representation of how well the regression equation is working. **Correlations of estimates** will produce correlations between all variables that have

**Screen 25.3** Logistic Regression: Options Window



Logistic Regression: Options

Statistics and Plots

□ Classification plots
□ Hosmer-Lemeshow goodness-of-fit
□ Casewise listing of residuals
● Outliers outside 2 std. dev.
○ All cases
□ Correlations of estimates
□ Iteration history
□ CI for exp(B): 95 %

Display

● At each step ○ At last step

Probability for Stepwise

Entry: 0.05 Removal: 0.10

Classification cutoff: 0.5

Maximum Iterations: 20

□ Conserve memory for complex analyses or large datasets

☑ Include constant in model

been entered into the regression equation. Values for the **Probability for Stepwise Entry** and **Removal** are important when stepwise regression (the Forward and Backward methods) is used. The **Probability for Stepwise Entry** specifies the probability value used to add a variable to the regression equation. Similarly, the **Probability for Stepwise Removal** specifies the values used to drop a variable from the regression equation. As default, SPSS uses a probability for **Entry** of .05, and a probability for **Removal** of .10. These values may be changed if the researcher desires.

*To perform a logistic regression analysis with* **cathelp** *as the dependent variable, and* **sympatht, angert, effict**, *and* **ethnic** *as predictor variables, perform the following sequence of steps. The starting point is Screen 25.1. Please perform whichever of Steps 1–4 (pages 323–324) are necessary to arrive at this screen.*

| In Screen | Do This | Step 5 |
|---|---|---|
| 25.1 | ✳ **cathelp** → ✳ *top* ➡ → ✳ **sympatht** → ✳ *middle* ➡ | |
| | → ✳ **angert** → ✳ *middle* ➡ → ✳ **effict** → ✳ *middle* ➡ | |
| | → ✳ **ethnic** → ✳ *middle* ➡ | |
| | → ✳ ▼ *to the right of* **Method** → ✳ **Forward:LR** → ✳ **Categorical** | |
| 25.2 | ✳ **ethnic** → ✳ ➡ → ✳ ▼ *to the right of* **Contrast:** → ✳ **Deviation** | |
| | → ✳ **Continue** | |
| 25.1 | ✳ **Options** | |
| 25.3 | ✳ **Classification plots** → ✳ **Correlations of estimates** → ✳ **Continue** | |
| 25.1 | ✳ **OK** | Back1 |

Upon completion of Step 5, the output screen will appear (Screen 1, inside back cover). All results from the just-completed analysis are included in the Output Navigator. Make use of the scroll bar arrows (▲▼▶◀) to view the results. Even when viewing output, the standard menu of commands is still listed across the top of the window. Further analyses may be conducted without returning to the data screen.

## 25.2 Printing Results

Results of the analysis (or analyses) that have just been conducted require a window that displays the standard commands (**File Edit Data Transform Analyze. . .**) across the top. A typical print procedure is shown below beginning with the standard output screen (Screen 1, inside back cover).

*To print results, from the Output screen perform the following sequence of steps:*

| In Screen | Do This | Step 6 |
|---|---|---|
| Back1 | *Select desired output or edit (see pages 16–22)* → ✳ **File** → ✳ **Print** | |
| 2.9 | *Consider and select desired print options offered, then* → ✳ **OK** | |

*To exit you may begin from any screen that shows the* **File** *command at the top.*

| In Screen | Do This | Step 7 |
|---|---|---|
| **Menu** | ✳ **File** → ✳ **Exit** | **2.1** |

Note: After clicking **Exit**, there will frequently be small windows that appear asking if you wish to save or change anything. Simply click each appropriate response.

# 25.3 Output

## 25.3.1 Logistic Regression Analysis

What follows is partial output from sequence Step 5 the previous page.

Because forward- and backward-stepping analyses involve reanalyzing the regression several times (adding or deleting variables from the equation each time), they may produce several sets of output, as illustrated here. The final set of output shows the final regression equation, but output is produced for each step in the development of the equation to see how the regression equation was formed. Only the final set of output is shown here.

## 25.3.2 Categorical Variables Codings

| | | | **Parameter coding** | | | |
|---|---|---|---|---|---|---|
| | | **Frequency** | **(1)** | **(2)** | **(3)** | **(4)** |
| ethnicity | **WHITE** | 293 | 1.000 | .000 | .000 | .000 |
| | **BLACK** | 50 | .000 | 1.000 | .000 | .000 |
| | **HISPANIC** | 80 | .000 | .000 | 1.000 | .000 |
| | **ASIAN** | 70 | .000 | .000 | .000 | 1.000 |
| | **OTHER/DTS** | 44 | .000 | .000 | .000 | .000 |

If **Contrasts** are used in the analysis, a table will be produced at the beginning of the output that shows how the computer has converted from the various values of your variable (the rows in the table) to coding values of several different variables within the computer (the columns). In this case, **ethnic** had five levels in the original data and has been broken down into four new variables, labeled **ethnic(1)** through **ethnic(4)**. These variables are a series of contrasts between the various ethnicities.

## 25.3.3 Omnibus Tests of Model Coefficients

**Omnibus Tests of Model Coefficients**

| | | **Chi-square** | **df** | **Sig.** |
|---|---|---|---|---|
| **Step 1** | Step | 114.843 | 1 | .000 |
| | Block | 114.843 | 1 | .000 |
| | Model | 114.843 | 1 | .000 |
| **Step 2** | Step | 29.792 | 1 | .000 |
| | Block | 144.635 | 2 | .000 |
| | Model | 144.635 | 2 | .000 |

**Model Summary**

| **Step** | **–2 Log likelihood** | **Cox & Snell R Square** | **Nagelkerke R Square** |
|---|---|---|---|
| **1** | 629.506 | .193 | .257 |
| **2** | 599.713 | .236 | .315 |

| Term | Definition/Description |
|---|---|
| **STEPS 1 AND 2** | Using the Forward: LR method of entering variables into the model, it took two steps for SPSS to enter all variables that significantly improved the model. We will find **which** variables are entered later (in the "Variables in Equation" section), but for now we will have to be content knowing that there are two variables in the model. |
| **STEP, BLOCK, AND MODEL $\chi^2$** | These values test whether or not all of the variables entered in the equation (for model), all of the variables entered in the current block (for block), or the current increase in the model fit (for step) have a significant effect. $\chi^2$ values are provided for each step of the model. In this case, a high $\chi^2$ value for the model and step for Step 1 (as Step 1 is the first step, it is the entire model; note that we are only working with one block, so $\chi^2$ values for the block are the same as for the model throughout) indicates that the first variable added to the model significantly impacts the dependent variable. The step $\chi^2$ for Step 2 indicates that adding a second variable significantly improves the model, and the model $\chi^2$ for Step 2 indicates that the model including two variables is significant. Note that the $\chi^2$ for the model in Step 2 is equal to the sum of the model in Step 1 and the step $\chi^2$ in Step 2. |
| **–2 LOG LIKELIHOOD** | This and the other model summary measures are used to indicate how well the model fits the data. Smaller –2 log likelihood values mean that the model fits the data better; a perfect model has a –2 log likelihood value of zero. |
| **COX & SNELL AND NAGELKERKE R SQUARE** | Estimates of the $R^2$ effect size value, indicating what percentage of the dependent variable may be accounted for by all included predictor variables. |

## 25.3.4 Classification Table

| | | | Predicted | | |
|---|---|---|---|---|---|
| | | | CATHELP | | |
| Observed | | | NOT HELPFUL | HELPFUL | Percentage Correct |
| Step 1 | CATHELP | NOT HELPFUL | 176 | 89 | 66.4 |
| | | HELPFUL | 79 | 193 | 71.0 |
| | Overall Percentage | | | | 68.7 |
| Step 2 | CATHELP | NOT HELPFUL | 181 | 84 | 68.3 |
| | | HELPFUL | 75 | 197 | 72.4 |
| | Overall Percentage | | | | 70.4 |

The classification table compares the predicted values for the dependent variable, based on the regression model, with the actual observed values in the data. When computing predicted values, SPSS simply computes the probability for a particular case (based on the current regression equation) and classifies it into the two categories possible for the dependent variable based on that probability. If the probability is less than .50, then SPSS classifies it as the first value for the dependent variable (*Not Helpful* in this example), and if the probability is greater than .50, then SPSS classifies that case as the second value for the dependent variable (*Helpful*). This table compares these predicted values with the values observed in the data. In this case, the Model-2 variables can predict which value of cathelp is observed in the data 70% of the time.

## 25.3.5 Variables in the Equation

| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| Step 1[a] | EFFICT | 1.129 | .122 | 85.346 | 1 | .000 | 3.094 |
| | Constant | –5.302 | .585 | 82.019 | 1 | .000 | .005 |
| Step 2[b] | SYMPATHT | .467 | .089 | 27.459 | 1 | .000 | 1.596 |
| | EFFICT | 1.114 | .128 | 76.197 | 1 | .000 | 3.046 |
| | Constant | –7.471 | .775 | 93.006 | 1 | .000 | .001 |

a. Variable(s) entered on step 1: EFFICT.
b. Variable(s) entered on step 2: SYMPATHT.

At last, we learn which variables have been included in our equation. For each step in building the equation, SPSS displays a summary of the effects of the variables that are currently in the regression equation. The constant variable indicates the constant $B_0$ term in the equation.

| Term | Definition/Description |
|---|---|
| B | The weighting value of *B* used in the equation; the magnitude of *B*, along with the scale of the variable that *B* is used to weight, indicates the effect of the predictor variable on the predicted variable. In this case, for example, sympathy and efficacy both have a positive effect on helping. |
| S.E. | Standard error; a measure of the dispersion of *B*. |
| WALD | A measure of the significance of *B* for the given variable; higher values, in combination with the degrees of freedom, indicate significance. |
| SIG | The significance of the **WALD** test. |
| EXP(B) | $e^B$, used to help in interpreting the meaning of the regression coefficients (as you may remember from the introduction to this chapter, the regression equation may be interpreted in terms of *B* or $e^B$). |

## 25.3.6 Correlation Matrix

| | | Constant | EFFICT | SYMPATHT |
|---|---|---|---|---|
| Step 1 | Constant | 1.000 | −.986 | |
| | EFFICT | −.986 | 1.000 | |
| Step 2 | Constant | 1.000 | −.825 | −.619 |
| | SYMPATHT | −.619 | .085 | 1.000 |
| | EFFICT | −.825 | 1.000 | .085 |

This is the correlation matrix for all variables in the regression equation, and it is only printed if you request **Correlations of estimates**. It is useful because if some variables are highly correlated, then the regression may have multicollinearity and be unstable.

## 25.3.7 Model if Term Removed

**Model if Term Removed**

| Variable | | Model Log Likelihood | Change in −2 Log Likelihood | df | Sig. of the Change |
|---|---|---|---|---|---|
| **Step 1** | effict | −372.174 | 114.843 | 1 | .000 |
| **Step 2** | sympatht | −314.753 | 29.792 | 1 | .000 |
| | effict | −350.086 | 100.458 | 1 | .000 |

All variables in the model are tested here to see if they should be removed from the model. If the significance of the change in the −2-Log-Likelihood-for-the-model-if-a-variable-is-dropped is larger than the value entered in the **Probability for Stepwise Removal** in the Options window (Screen 25.3), then that variable will be dropped. In this example, since all of the Significance of the Change values are less than .10 (the default), no variables are removed.

## 25.3.8 Variables not in the Equation

| | | | Score | df | Sig. |
|---|---|---|---|---|---|
| Step 2 | Variables | ANGERT | .640 | 1 | .424 |
| | | ETHNIC | 4.892 | 4 | .299 |
| | | ETHNIC(1) | .464 | 1 | .496 |
| | | ETHNIC(2) | 2.304 | 1 | .129 |
| | | ETHNIC(3) | .357 | 1 | .550 |
| | | ETHNIC(4) | 2.169 | 1 | .141 |
| | Overall Statistics | | 5.404 | 5 | .369 |

All variables that are not entered into the equation that could possibly be entered are listed here. The **Sig** indicates for each variable whether it has a significant impact on the predicted variable, independently from the other predictor variables. Here, **angert** does not have an impact on **cathelp**. Note here that **ethnic** is divided into four variables, indicating the four different contrasts that SPSS is performing. Because no variables can be deleted or added, the logistic regression equation is now complete. SPSS will not try to add or delete more variables from the equation.

## 25.3.9 Observed Groups and Predicted Probabilities

**_Observed Groups and Predicted Probabilities table_:**

```
      20 +                                                                                                    +
         I                                                                                                    I
         I                                                                     H                              I
 F       I                                     H                               H                              I
 R    15 +                                     H                      H        H                              +
 E       I                                     H                H H   H        H                              I
 Q       I                         H       H   H    H     H     H H   H   H    H                              I
 U       I                         N       H   H    H     H H   H H   H   H    H                              I
 E    10 +        H                N       H   N    H     H H   H H   H   H  H H     H                        +
 N       I        H                N       N   N    H     H H   H H H H   H  H H  H  H          H             I
 C       I        N         H  H   NH N    N H N H  H H   H H H N H H H H H  H H  H HH    H     HH H          I
 Y       I      N N H       N  N   NN N    N H N H  HHHH  N H H N H H N HHH  N H  H HH  H H   H HH H          I
       5 +      N NNN       NH N HHNHNN  NHN H N NHHHHHH  NHNHH N N NHN HHH  NHH  H HH  HHH   HHHH H    H     +
         I      N NNN  N    NN N HHNNNN HNHN N N NNHNHHNHHNHNHN N N NHN HHHHHNHN  N HH  HHHH HHHHH H  HHHH    I
         I  NN NNNNNNNHNHHHHNNHNNNNNNNNNNNNN N NNNNHNNNNNNNNNNNHNHNHNNNNHNNNHNNN HNHNN HHHNHHHNHHHHH NHHHHH   I
         I NNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNHHNNNN NHHNNNNNNHNNH NNNHHHH  I
Predicted ---------+---------+---------+---------+---------+---------+---------+---------+---------+----------
  Prob:   0        .1        .2        .3        .4        .5        .6        .7        .8        .9         1
  Group:  NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH

          Predicted Probability is of Membership for HELPFUL
          The Cut Value is .50
          Symbols: N - NOT HELPFUL
                          H - HELPFUL
          Each Symbol Represents 1.25 Cases.
```

This graph is produced if you request **Classification plots** on Screen 25.3. First note that this graph will simply use the first letter of the value label for coding. With labels of *helpful* and *not helpful*, the graph is composed of Ns and Hs. These codes are plotted along the *x*-axis based on its predicted probability from the regression equation. If the logistic regression equation were perfect, all of the Ns would be to the left of all of the Hs.

# Chapter 26
# Hierarchical Log-Linear Models

BOTH THIS chapter and the chapter that follows focus on log-linear models: the present chapter on *hierarchical* models, and the next (Chapter 27) on *non* hierarchical models. The first part of this introduction will discuss log-linear models in general and is applicable (and indeed, necessary) for both this chapter and the next.

## 26.1 Log-Linear Models

Log-linear models allow the analysis of chi-square-type data using regression-like models, and in many ways they appear to be similar to ANOVAs. As in chi-square analysis, log-linear models deal with frequency tables in which several categorical variables categorize the data. If only two categorical variables are in the data for analysis, then chi-square analyses are indeed the simplest analysis available. For example, in Chapter 8 on chi-square analysis, the relationship between gender and ethnicity was examined in the **grades.sav** file. If, however, you wish to analyze several different categorical variables together, then it quickly becomes difficult or impossible to interpret chi-square tables visually. For example, a chi-square table and analysis of **gender** by **ethnic** (ethnicity) by **year** (in school) by **review** (whether or not the student attended the review session) is virtually impossible to interpret. It is for this purpose that log-linear models were developed.

In this chapter, the example will be drawn from the same **helping3.sav** file ($N = 537$) used in the previous chapter. In particular, hierarchical log-linear models will be tested for **gender** × **ethnic** × **income** level × **cathelp** (whether or not the person receiving help thought the help was useful or not). Note that in this example, **income** is a categorical variable, indicating whether subjects earned less than $15,000/year, between $15,000 and $25,000, between $25,000 and $50,000, or greater than $50,000/year.

Log-linear models are essentially multiple linear regression models in which the classification variables (and their interaction terms) are the independent (predictor) variables, and the dependent variable is the natural logarithm of the frequency of cases in a cell of the frequency table. Using the natural log (*ln*) of the frequencies produces a linear model. A log-linear model for the effects of gender, ethnicity, income level, and their interactions on a particular cell in the crosstabulation table might be represented by the following equation:

$$In(FREQUENCY) = \mu + \lambda^{G} + \lambda^{E} + \lambda^{I} + \lambda^{GxE} + \lambda^{GxI} + \lambda^{ExI} + \lambda^{GxExI}$$

There will be different values for each of these variables for each of the cells in the model. In this equation, *Frequency* represents the frequency present within a particular cell of the data. $\mu$ represents the overall grand mean of the effect; it is equivalent to the constant in multiple regression analysis. Each of the λs (lambdas) represents the effect of one or more independent variables. $\lambda^{G}$ represents the main effect (here is where log-linear models start sounding similar to ANOVAs) of **gender**, $\lambda^{E}$ represents the main effect (also known as first-order effect) of **ethnic**, and $\lambda^{I}$ represents the main effect of **income** level on the frequency. λ values with multiple superscripts are interaction terms; for example, $\lambda^{G \times E}$ represents the two-way (second-order) interactive effect of **gender** and **ethnic** on frequency, and $\lambda^{G \times E \times I}$ represents the three-way (third-order) interactive effect of **gender, ethnic**, and **income** on the frequency. The model presented here is a *saturated* model because it contains *all* possible main effects and interaction terms. Because it is a saturated model, it can perfectly reproduce the data; however, it is not parsimonious and usually not the most desirable model.

The purpose of SPSS's **Model selection** option of the **Loglinear** procedures is to assist you in choosing an unsaturated log-linear model that will fit your data, as well as to calculate parameters of the log-linear model (the $\mu$s and $\lambda$s).

## 26.2 The Model Selection Log-Linear Procedure

Although any of the terms may be deleted from the saturated model in order to produce a simpler, more parsimonious model, many researchers explore *hierarchical* log-linear models. In hierarchical models, in order for an effect of a certain order to be present, all effects of a lower order must be present. In other words, in a hierarchical model, in order for a two-way interaction (second-order effect) of **gender** and **ethnic** to be present, there must also be main effects (first-order effects) of both **gender** and **ethnic**. Likewise, in order for a third-order interactive effect of **gender, ethnic**, and **income** level to be present, second-order interactive effects of **gender** by **ethnic** level, **gender** by **income** level, and **ethnic** by **income** level must be present.

There are three primary techniques that SPSS provides in assisting with model selection. All three techniques are useful and will usually yield similar or identical results. Ultimately, however, the choice of which model or models you will use has to rely on both the statistical results provided by SPSS *and* your understanding of the data and what the data mean. The three techniques are summarized here, with a more detailed example provided in the Output section:

- **Examine parameter estimates:** One technique used in developing a model is to calculate parameter estimates for the saturated model. SPSS provides, along with these parameter estimates, *standardized* parameter estimates. If these standardized parameter estimates are small, then they probably do not contribute very much to the model and might be considered for removal.
- **Partitioning the chi-square statistic:** SPSS can, in addition to providing parameter estimates for the model, calculate a chi-square value that indicates how well the model fits the data. This chi-square value may also be subdivided and may be useful in selecting a model. SPSS can test whether all one-way and higher effects are nonsignificant, whether all two-way and higher effects are nonsignificant, whether all three-way and higher effects are nonsignificant, and so on. The program can also test that all one-way effects are zero, all two-way effects are zero, and so on. These tests examine a combination of all first-order effects, second-order effects, and so forth. However, just because the second-order effects *overall* may not be significant doesn't mean that *none* of the individual second-order effects is significant. Similarly, just because the second-order effects are significant overall doesn't mean that *all* of the second-order effects are significant. Because of this, SPSS can also examine partial chi-square values for individual main effects and interactive effects.
- **Backward elimination:** Another way to select a model is to use backward elimination; this is very similar to backward elimination in multiple regression analysis. In backward elimination, SPSS starts with a saturated model and removes effects that are not contributing to the model significantly. This model-building technique is subject to the constraints of hierarchical log-linear modeling; third-order effects are not examined as candidates for exclusion if fourth-order effects are present, since if they were removed the assumptions of hierarchical models

would be violated. The model is considered to fit best when all remaining effects contribute significantly to the model's fit.

We acknowledge that hierarchical log-linear models are complex and direct you to three other sources for a more complete picture than has been presented here: The SPSS manuals do a fairly thorough job of description and have the advantage of explaining the procedure within the context of SPSS documentation. Two other textbooks are quite good: Agresti (1996) and Wickens (1989); please see the reference section for a more complete description.

## 26.3 Step by Step

### 26.3.1 Model Selection (Hierarchical) Log-Linear Models

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | 🖱 ⊞ → 🖱 All Programs ▷ → 🖱 IBM SPSS Statistics → 🖱 IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | 🖱 🚀 → 🖱 ■ → 🖱 Σα | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| | Step 2 |
|---|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* | |

*Screens 1 and 2 (on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **helping3.sav** *file for further analyses:*

| In Screen | Do This | Step 3 |
|---|---|---|
| Front1 | 🖱 File → Open → 🖱 Data [*or* 🖱 📂] | |
| Front2 | type `helping3.sav` → 🖱 Open [*or* 🖱 🖱 helping3.sav] | Data |

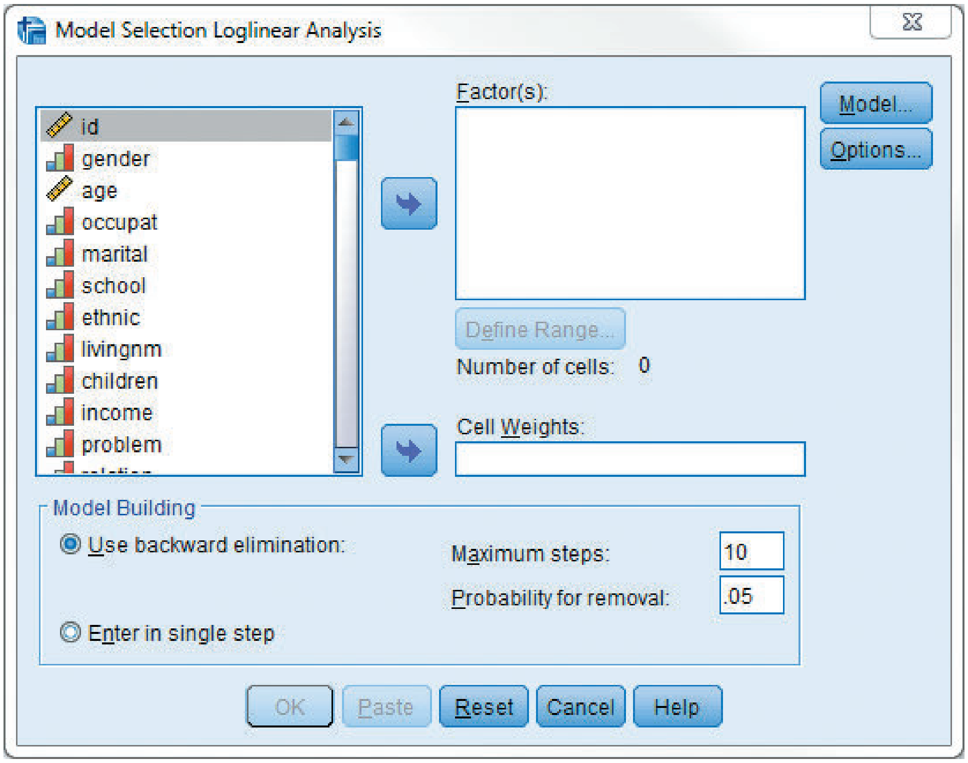
Whether first entering SPSS or returning from earlier operations the standard menu of commands across the top is required. As long as it is visible you may perform any analyses. It is not necessary for the data window to be visible.

*After completion of Step 3 a screen with the desired menu bar appears. When you click a command (from the menu bar), a series of options will appear (usually) below the selected command. With each new set of options, click the desired item. The sequence to access hierarchical log-linear models begins at any screen with the menu of commands visible:*

| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | 🖱 Analyze → Loglinear → 🖱 Model Selection | 26.1 |

After this sequence step, a window opens (Screen 26.1, following page) in which you specify the factors to include in the analysis and select the model building procedure. In the box at the left of Screen 26.1 is the list of variables in your data file. Any variable that will be included in the analysis should be moved to the **Factor(s)** box to the right of the variable list, by clicking on the upper ➡. Because all variables used in the analysis will be categorical, it is important to click on **Define Range** for each variable. SPSS will prompt you (Screen 26.2, below) to enter the **Minimum** and **Maximum** values of the variable. For example, in most cases gender variables will have a range of 1 through 2.

Screen 26.1 Model Selection (Hierarchical) Loglinear Analysis Dialog Window



Number of cells: 0

● Use backward elimination:

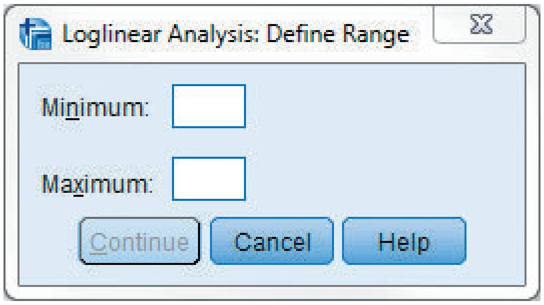
Maximum steps: 10

Probability for removal: .05

○ Enter in single step

**Cell Weights** are used if your model contains structural zeros. Structural zeros are beyond the scope of this book; please see the SPSS manuals for details.

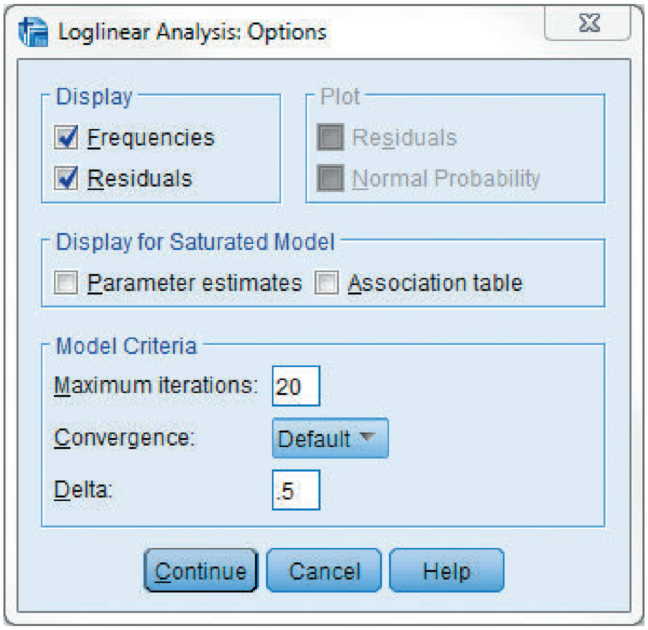
Screen 26.2 Loglinear Analysis Define Range Dialog Window



The **Use backward elimination** selection in the **Model Building** box at the bottom of Screen 26.1 instructs SPSS to start with a saturated model and remove terms that do not significantly contribute to the model, through a process of backward elimination. Use the **Enter in single step** option if you do not wish to eliminate terms from the model, but test the model containing all of the terms. Assuming you do want to use backward elimination, the **Probability for removal** provides a *p* value that SPSS will use as a criterion for dropping effects from the model. SPSS drops effects that are not significant at the level of the designated *p* value. The default is *p* = .05, so you do not need to enter a value here unless you wish to specify a different *p* value. Each time that SPSS removes an effect from the model, this is called a *step*. **Maximum steps** allows you to set the maximum number of steps that SPSS will perform, to something other than the default of 10. You may need to increase the maximum number of steps if you are testing a very large model.

Once the factors for analysis are defined, and the model building procedure is identified, the **Options** and **Model** may be specified. When the **Options** button is clicked, Screen 26.3 appears. This window allows you to select which output you desire, as well as specifying some technical aspects of the way model selection will proceed (these technical aspects are not covered in this book). **Frequencies** produces a listing of the number of subjects in each cell of the model. **Residuals** displays the difference between the actual number of subjects within each cell, and the number of subjects predicted to be in that cell based on the model.

Screen 26.3 Model Selection (Hierarchical) Loglinear Analysis Options Dialog Window



Loglinear Analysis: Options

Display
☑ Frequencies
☑ Residuals

Plot
□ Residuals
□ Normal Probability

Display for Saturated Model
□ Parameter estimates □ Association table

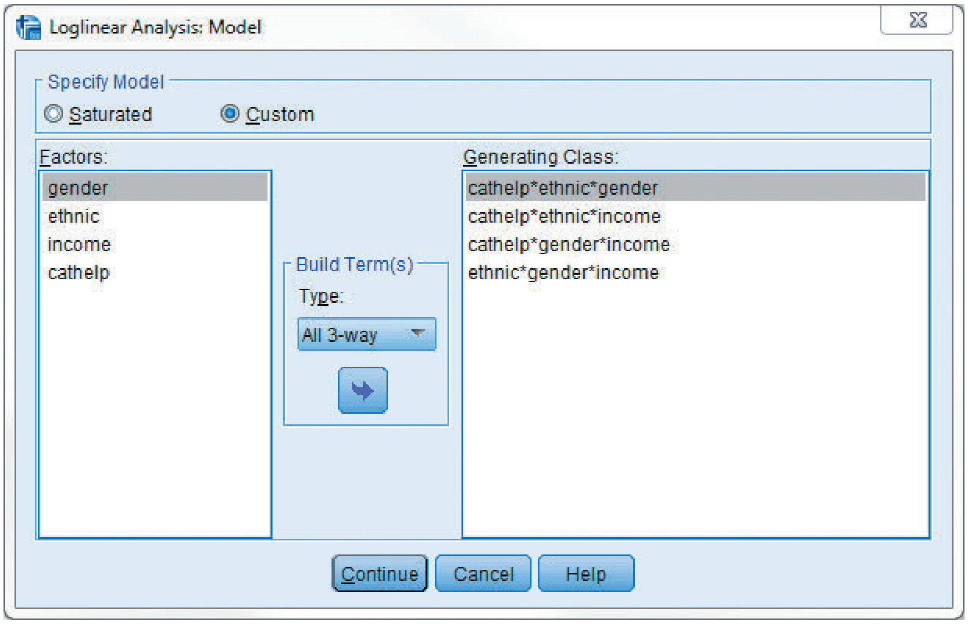
Model Criteria
Maximum iterations: 20
Convergence: Default
Delta: .5

Other frequently used options include **Parameter estimates** and an **Association table**. For a saturated model, these options will produce parameter estimates (which may be useful in model selection; see the introduction to this chapter) and an association table that partitions the chi-square values for each of the main and interactive effects (again useful in model selection). These options are strongly recommended when model selection with a saturated model is desired.

After all desired options are selected (Screen 26.3), one more button on Screen 26.1 may be needed: **Model**. If you wish to test a saturated model (in which all effects and interaction effects are included), then this step is not necessary. The **Model** button should only be used when you wish to use a nonsaturated model, using only some of the first-order effects and interactive effects.

The **Model** button calls up Screen 26.4 (shown on the following page with a custom model already specified). After **Custom** is selected, **Factors** may be selected by clicking on them, in the list at the left of the window. Once they have been selected, the type of effects (both main effects and interactions) may be chosen from the list underneath the ➡. **Factors** may be selected as **Main effects, Interactions**, or **All 2-way, All 3-way, All 4-way**, or **All 5-way**. In Screen 26.4, we have selected **All 3-way** interactions. Because this is a hierarchical model, all three-way interactions also assume that all two-way interactions and main effects are also present. **Custom** models are generally not used unless the researcher has strong theoretical reasons for choosing to use a nonsaturated model, and/or the saturated model would contain higher-order interactions that would be too difficult to interpret.

**Screen 26.4** Model Selection (Hierarchical) Loglinear Analysis Model Specification Dialog Window



Loglinear Analysis: Model

Specify Model

○ Saturated ● Custom

Factors:

gender
ethnic
income
cathelp

Build Term(s)

Type: All 3-way

Generating Class:

cathelp*ethnic*gender
cathelp*ethnic*income
cathelp*gender*income
ethnic*gender*income

*The following sequence instructs SPSS to perform a hierarchical log-linear analysis of* **gender** *(two levels),* **ethnic** *(five levels),* **income** *(four levels), and* **cathelp** *(two levels, whether or not the individuals helped felt that they had benefited). The saturated model is tested, and* **Parameter estimates** *will be produced, along with partitioning the chi-square into individual effects. Finally,* **backward elimination** *will be used to build a model containing only effects that significantly contribute to the model. The step begins with the initial dialog box, Screen 26.1.*

| In Screen | Do This | Step 5 |
|---|---|---|
| 26.1 | ✳ **gender** → ✳ *top* ➡ → ✳ **Define Range** | |
| 26.2 | **type** 1 → press **TAB** → **type** 2 → ✳ **Continue** | |
| 26.1 | ✳ **ethnic** → ✳ *top* ➡ → ✳ **Define Range** | |
| 26.2 | **type** 1 → press **TAB** → **type** 5 → ✳ **Continue** | |
| 26.1 | ✳ **income** → ✳ *top* ➡ → ✳ **Define Range** | |
| 26.2 | **type** 1 → press **TAB** → **type** 4 → ✳ **Continue** | |
| 26.1 | ✳ **cathelp** → ✳ *top* ➡ → ✳ **Define Range** | |
| 26.2 | **type** 1 → press **TAB** → **type** 2 → ✳ **Continue** | |
| 26.1 | ✳ **Options** | |
| 26.3 | ✳ **Parameter estimates** → ✳ **Association table** → ✳ **Continue** | |
| 26.1 | ✳ **OK** | **Back1** |

Upon completion of Step 5 the output screen will appear (Screen 1, inside back cover). All results from the just-completed analysis are included in the Output Navigator. Make use of the scroll bar arrows (▲▼▶◀) to view the results. Even when viewing output, the standard menu of commands is still listed across the top of the window. Further analyses may be conducted without returning to the data screen.

## 26.4 Printing Results

Results of the analysis (or analyses) that have just been conducted require a window that displays the standard commands (**File Edit Data Transform Analyze...**) across the top. A typical print procedure is shown below beginning with the standard output screen (Screen 1, inside back cover).

*To print results, from the Output screen perform the following sequence of steps:*

| In Screen | Do This | Step 6 |
|---|---|---|
| Back1 | *Select desired output or edit (see pages 16–22)* → 🖱 **File** → 🖱 **Print** | |
| 2.9 | *Consider and select desired print options offered, then* → 🖱 **OK** | |

*To exit you may begin from any screen that shows the* **File** *command at the top.*

| In Screen | Do This | Step 7 |
|---|---|---|
| Menu | 🖱 **File** → 🖱 **Exit** | 2.1 |

Note: After clicking **Exit**, there will frequently be small windows that appear asking if you wish to save or change anything. Simply click each appropriate response.

## 26.5 Output

### 26.5.1 Model Selection (Hierarchical) Log-Linear Models

Below is an explanation of the key output from the Hierarchical Loglinear Model selection procedure (sequence Step 5, previous page). Only representative output is provided, because the actual printout from this procedure is approximately 20 pages long.

**Tests that K-Way and Higher-Order Effects (are zero)**

```
K         DF      L.R. Chisq      Prob      Pearson Chisq      Prob      Iteration
1         79        428.287       .000        618.095          .000          0
2         70        112.633       .001        102.374          .007          2
3         43         62.361       .028         59.044          .052          3
4         12         13.408       .340         12.330          .420          4
```

These tests examine whether all effects at a certain order and above are zero. For example, the last line in this table tests whether all fourth-order effects are equal to zero. As indicated by the small chi-square values and the large $p$ values, there is no fourth-order effect. The third line ($K = 3$) indicates that the third- and fourth-order effects may be significantly different from zero (note $p$ values of .028 and .052), and the second line ($K = 2$) indicates that the second-, third-, and fourth-order effects *are* significantly different from zero (note $p = .007$). The first line ($K = 1$) suggests that, in the first- through fourth-order effects, there are some effects that are not equal to zero.

| Term | Definition/Description |
|---|---|
| **K** | The order of effects for each row of the table (4 = fourth-order and higher effects, 3 = third-order and higher effects, etc.). |
| **DF** | The degrees of freedom for *K*th and higher-order effects. |
| **L.R. CHISQ** | The likelihood-ratio chi-square value testing that *K*th and higher-order effects are zero. |
| **PEARSON CHISQ** | The Pearson chi-square value testing that *K*th and higher-order effects are zero. |
| **PROB** | The probability that *K* order and higher effects are equal to zero; small *p* values suggest that one or more of the effects of order *K* and higher are not equal to zero. |
| **ITERATION** | The number of iterations that SPSS took to estimate the chi-square values. |

**Tests that K-Way Effects (are zero)**

| K | DF | L.R. Chisq | Prob | Pearson Chisq | Prob | Iteration |
|---|---|---|---|---|---|---|
| 1 | 9 | 315.653 | .000 | 515.721 | .000 | 0 |
| 2 | 27 | 50.272 | .004 | 43.331 | .024 | 0 |
| 3 | 31 | 48.953 | .021 | 46.714 | .035 | 0 |
| 4 | 12 | 13.408 | .340 | 12.330 | .420 | 0 |

| Term | Definition/Description |
|---|---|
| **K** | The order of effects for each row of the table (1 = first-order effects, 2 = second-order effects, etc.). |
| **DF** | The degrees of freedom for *K*th-order effects. |
| **L.R. CHISQ** | The likelihood-ratio chi-square value testing that *K*th-order effects are zero. |
| **PEARSON CHISQ** | The Pearson chi-square value testing that *K*th-order effects are zero. |
| **PROB** | The probability that *K*-order effects are equal to zero; small *p* values suggest that one or more of the effects of order *K* are not equal to zero. |
| **ITERATION** | For this table, this column will always be zero. |

**Partial Associations**

| Effect | df | Partial Chi-Square | Sig. | Number of Iterations |
|---|---|---|---|---|
| gender*ethnic*income | 12 | 17.673 | .126 | 4 |
| gender*ethnic*cathelp | 4 | 7.939 | .094 | 3 |
| gender*income*cathelp | 3 | 12.470 | .006 | 4 |
| ethnic*income*cathelp | 12 | 14.680 | .259 | 4 |
| gender*ethnic | 4 | 3.214 | .523 | 3 |
| gender*income | 3 | 1.605 | .658 | 3 |
| ethnic*income | 12 | 32.394 | .001 | 3 |
| gender*cathelp | 1 | 4.169 | .041 | 3 |
| ethnic*cathelp | 4 | 5.399 | .249 | 3 |
| income*cathelp | 3 | 4.918 | .178 | 3 |
| gender | 1 | 10.248 | .001 | 2 |
| ethnic | 4 | 236.046 | .000 | 2 |
| income | 3 | 66.886 | .000 | 2 |
| cathelp | 1 | 2.473 | .116 | 2 |

This table (previous page) examines partial chi-square values for each effect in the saturated model. Each partial chi-square examines the unique contribution of that effect to the model; those with low *p* values contribute to the model significantly. In this case, there are main effects of **gender**, **ethnic**, and **income**, two-way interactions between **gender** by **cathelp** and **ethnic** by **income**, as well as a three-way interaction of **gender** by **income** by **cathelp**. Because these partial chi-squares are not necessarily independent, their partial associations (displayed here) as a portion of the total chi-square for the saturated model may not be equivalent to the chi-square in nonsaturated models.

| Term | Definition/Description |
|---|---|
| **DF** | These degrees of freedom refer to the particular effects listed in each line. |
| **PARTIAL CHI-SQUARE** | The partial chi-square for each effect. |
| **SIG** | The probability that the effect is equal to zero; small probabilities indicate that the given effect has a large contribution to the model. |
| **N OF ITERATIONS** | The number of iterations that the computer took to calculate the partial association for each effect. This procedure takes quite a while; the number of iterations is there to remind you how long it took to compute each partial association. |

**Estimates for Parameters**

```
.
.
ethnic*income
Parameter    Coeff.   Std. Err.    Z-Value      sig    Lower 95 CI  Upper 95 CI
 1            -.109       .191        -.574     .566       -.483         .264
 2            -.170       .222        -.767     .443       -.605         .264
 3            -.113       .159        -.707     .480       -.425         .200
 4            -.682       .370       -1.845     .065      -1.407         .042
 5             .726       .269        2.701     .007        .199        1.252
 6            -.023       .253        -.093     .926       -.519         .472
 7            -.097       .245        -.395     .693       -.576         .383
 8             .226       .248         .910     .363       -.260         .712
 9             .110       .201         .546     .585       -.284         .503
10             .733       .236        3.103     .002        .270        1.196
11            -.483       .371       -1.302     .193      -1.211         .244
12            -.428       .264       -1.623     .105       -.944         .089
.
.
income
Parameter    Coeff.   Std. Err.    Z-Value      sig    Lower 95 CI  Upper 95 CI
1             -.183       .142       -1.287     .198       -.462         .096
2             -.491       .158       -3.107     .002       -.800        -.181
3              .244       .117        2.088     .037        .015         .472
```

Parameter estimates are provided for each main effect and interaction. These are the λs from the log-linear equation presented in the introduction of this chapter. In this example, only one interaction (**ethnic** × **income**) and one main effect (**income**) are presented. Because the parameters are constrained to sum to zero, only some of the parameters appear here. In this example, the **ethnic** × **income** parameters may be interpreted as shown in the following table, where parameters in **bold** print are calculated by SPSS and those in *italics* are calculated by summing each row and column in the table to zero.

| | Ethnicity | | | | | |
|---|---|---|---|---|---|---|
| **Income** | **White** | **Black** | **Hispanic** | **Asian** | **Other** | **SUM** |
| <15,000 | **–.109** | **–.682** | **–.097** | **.733** | *.155* | 0 |
| <25,000 | **–.170** | **.726** | **.226** | **–.483** | *–.299* | 0 |
| <50,000 | **–.113** | **–.023** | **.110** | **–.428** | *.453* | 0 |
| >50,000 | *.391* | *–.021* | *–.239* | *.178* | *–.309* | 0 |
| **SUM** | 0 | 0 | 0 | 0 | 0 | 0 |

These parameters suggest that Whites with higher incomes are far more frequent in this sample than those with lower incomes and that Blacks, Hispanics, and Asians are more common with lower incomes. Additional detail may be observed: Notice that the Asians, although well represented in the lowest income level, are also quite prominent in the highest income level. The parameters for **income** may be interpreted in the same way as the **ethnic** × **income** interaction:

**Income**

| <15,000 | <25,000 | <50,000 | >50,000 | SUM |
|---|---|---|---|---|
| **–.183** | **–.491** | **.244** | *.430* | 0 |

These parameters suggest that there are few individuals with less than $25,000/year income and more with greater than $50,000/year income. In this case, the fact that the $z$ values for most of these parameters are large (greater than 1.96 or less than −1.96) imply that the parameters are significant (at $p < .05$) and support this interpretation of the data.

| Term | Definition/Description |
|---|---|
| **PARAMETER** | The number of the parameter. Parameters go from low codes to high codes of the independent variables. See the above **gender** × **ethnic** table for an example of the way interactions are coded. |
| **COEFFICIENT** | The $\lambda$ value in the log-linear model equation. |
| **STANDARD ERROR** | A measure of the dispersion of the coefficient. |
| **Z-VALUE** | A standardized measure of the parameter coefficient; large $z$ values (those whose absolute value is greater than 1.96) are significant ($\alpha = .05$). |
| **LOWER AND UPPER 95% CONFIDENCE INTERVAL** | There is a 95% chance that the actual (rather than estimated) coefficient is between the lower and upper confidence interval. |

## 26.5.2 Backward Elimination

In the output below, the chi-square values, probability values, and interactions involved in backward elimination were already explained and are not different here from the previous printouts in which the chi-square values were partitioned. Backward elimination begins with, in this case, the saturated model. SPSS calculates a likelihood ratio chi-square for the model including the fourth-order and all first-, second-, and third-order effects; this chi-square is 0, because it is a saturated model. SPSS calculates that if the four-way effect was deleted, the chi-square change would not be significant. This suggests that the four-way interactive effect does not contribute significantly to the model, and it is eliminated.

```
Backward Elimination (p = .050) for DESIGN 1 with generating class


gender*ethnic*income*cathelp

.
.
.

If Deleted Simple Effect is  L.R. Chisq Change     DF        Prob        Iter
   gender*ethnic*income*-          13.408          12        .340          4
           cathelp
Step 1

The best model has generating class
   gender*ethnic*income
   gender*ethnic*cathelp
   gender*income*cathelp
   ethnic*income*cathelp

Likelihood ratio chi square = 13.408            DF = 12   P = .340

.
.
.
```

*(Continued)*

```
Step 6
The best model has generating class
   gender*income*cathelp
   ethnic*income
                                  (output continues)
Likelihood ratio chi square = 59.380            DF = 48   P = .126
If Deleted Simple Effect is  L.R. Chisq Change     DF       Prob       Iter
 gender*income*cathelp            11.809            3       .008         3
 ethnic*income                    31.567           12       .002         2
Step 7
The best model has generating class
   gender*income*cathelp
   ethnic*income
 Likelihood ratio chi square = 59.380           DF = 48   P = .126
The final model has generating class
   gender*income*cathelp
   ethnic*income
```

In Step 1, the model with all three-way interactions is tested, and all effects are tested to see whether or not they significantly contribute to the model. This process continues (through seven steps in this example); lower-order effects are evaluated and considered for elimination only if they may be removed from the model without violating the hierarchical assumptions. In this case, the final model is composed of **gender** × **income** × **cathelp** and **ethnic** × **income** interactions. The model has a chi-square of 59.38 with 48 degrees of freedom; because this is not significant, the model does appear to fit the data well.

### 26.5.3 Observed, Expected Frequencies and Residuals

For the output below, observed and expected frequencies are presented for each cell in the design. Note that in order to avoid cells with zero frequencies, SPSS adds .5 to each cell frequency when saturated models are examined. Here, expected frequencies for the final model are chosen through backward elimination. Large residuals indicate possible problems with the model. Because the chi-square values are not large and the *p* values are not small, the model does seem to fit the data well.

```
 Factor            Code          OBS count    EXP count   Residual    Std Resid
gender         FEMALE
 ethnic         CAUCASIAN
  income            <15,000
   cathelp          NOT HELPFUL      4.000        5.753     -1.753       -.731
   cathelp          HELPFUL         14.000       13.904       .096        .026
  income            <25,000
   cathelp          NOT HELPFUL      7.000        5.392      1.608        .692
   cathelp          HELPFUL          9.000        7.843      1.157        .413
  income            <50,000
   cathelp          NOT HELPFUL     11.000       13.679     -2.679       -.724
   cathelp          HELPFUL         16.000       17.925     -1.925       -.455
  income            >50,000
   cathelp          NOT HELPFUL     20.000       26.679     -6.679      -1.293
   cathelp          HELPFUL         34.000       31.126      2.874        .515
          .
          .
          .
```

*(Continued)*

```
 ethnic       BLACK
  income      <15,000
   cathelp    NOT HELPFUL          1.000        .493        .507        .722
   cathelp    HELPFUL              2.000       1.192        .808        .740
         .
         .
         .
Goodness-of-fit test statistics:
Likelihood ratio chi square = 59.380        DF = 48      P = .126
         Pearson chi square = 54.598        DF = 48      P = .238
```

| Term | Definition/Description |
|---|---|
| **OBSERVED COUNT** | The observed cell count derives from the data and indicates the number of cases in each cell. |
| **EXPECTED COUNT** | The expected cell counts indicate the expected frequencies for the cells, based on the model being tested. |
| **RESIDUAL AND STANDARDIZED RESIDUAL: OBSERVED-EXPECTED** | Cells with high values here are those that are not well predicted by the model. |
| **GOODNESS-OF-FIT TEST STATISTICS** | The likelihood ratio chi-square and the Pearson chi-square statistics examine the fit of the model. Large chi-square values and small *p* values indicate that the model does not fit the data well; in this case, since *p* > .05, the data do fit the model adequately. |

# Chapter 27 Nonhierarchical Log-Linear Models

LOG-LINEAR modeling, using the SPSS General Loglinear procedure, is very flexible. Because of its great flexibility, it is difficult to provide a simple step-by-step approach to these operations. Using the General Loglinear procedure often involves executing the procedure multiple times, testing various models in an attempt to discover the best fit. Because of these subtleties in using log-linear models, this chapter will attempt to (a) provide a general introduction to log-linear modeling for those who need a review of the basic ideas involved and (b) provide step-by-step computer instructions for running the most common types of models. For more complex models, beyond the scope of this book, we refer you to the *SPSS* manuals for details.

If you are not familiar with the basic concepts of log-linear models, and if you have not read Chapter 26, then be sure to read the first portion of that chapter before continuing with this chapter. The general introduction there applies to creating both hierarchical log-linear models (Chapter 26) and general, nonhierarchical log-linear models (this chapter).

## 27.1 Models

### 27.1.1 Nonhierarchical versus Hierarchical Log-Linear Models

In hierarchical log-linear models (using procedures described in the previous chapter), if an effect of a given order is present, then all effects of lesser order must also be present. This constraint is unique to *hierarchical* log-linear models. In the General Loglinear procedure of SPSS, this is not a necessary constraint: You may include any main or interactive effect, or omit an effect if you don't want the effect included. Furthermore, although the **Loglinear** → **Model Selection** procedure can calculate only parameter estimates for saturated models, the general log-linear procedure can calculate parameter estimates whether or not the model is saturated.

As in the previous two chapters, we use the **helping3.sav** file ($N = 537$) in this chapter. A model is tested including the main effects of **gender, ethnic** (ethnicity), and **cathelp** (whether or not the help given was useful or not), along with the two-way interactive effects of **gender** by **ethnic** and **gender** by **cathelp**. One thing to remember: Testing a single model, as is done in the example in this chapter, is usually only one step among many in the quest for the best possible model.

### 27.1.2 Using Covariates with Log-Linear Models

It is possible to use one or more variables as covariates in the model. Because these covariates are not categorical variables, it is possible to use covariates to test for particular types of trends within categorical variables. For example, in the **helping3.sav** file, **income** seems to be related to the *number of people* in each income category: There are 73 people with less than $15,000/year income, 51 people between $15,000 and $25,000, 106 people between $25,000 and $50,000, and 159 people with greater than $50,000/year. In the Step by Step section of this chapter, we discuss a model that examines **income** and **income**$^3$ (**income** cubed) to see if a model that includes these two factors predicts well the number of subjects from each income group (that is, produces a model that fits the data well).

### 27.1.3 Logit Models

SPSS has the ability to work with logit models, through the **Loglinear** → **Logit** procedure. Conceptually, this modeling procedure is very similar to general log-linear modeling, with several exceptions. First of all, logit models allow dichotomous variables to be treated as dependent variables and one or more categorical variables to be treated as independent variables. Second, instead of predicting the frequency within a particular cell, a dichotomous dependent variable is designated and membership into one of two distinct categories (the logit—see Chapter 25 for a description) is predicted for each cell.

The execution of logit models in SPSS is very similar to the execution of general log-linear models. One difference is important: A dependent variable, always dichotomous, is specified. In the Step by Step section of this chapter, we discuss a model in which **gender** and a **gender** by **ethnic** interaction (the two categorical independent variables) predict **cathelp** (the dichotomous dependent variable).

## 27.2 A Few Words about Model Selection

The procedures described in this chapter assume that you know what model you want to test. In practice, this is not an easy task. Model selection usually depends on a tight interplay of theory and testing of multiple models. If the different models tested have different degrees of freedom, then they may be compared using chi-square differences (using the tables in the back of your statistics textbooks you thought you'd never need again) in order to determine whether or not one model is significantly better than another model. The goal of model selection is to find a model with the best fit possible *and* as parsimonious as possible. Obviously, this process is too complex and involves too much artistry to fully describe in a step-by-step format. However, the model selection process is likely to *use* (if not consist of) the procedures described in the Step by Step section.

## 27.3 Types of Models Beyond the Scope of This Chapter

In addition to the complexities of model selection, this chapter does not describe several types of models used with the general log-linear procedure because their complexity makes it difficult to describe in fewer words than are used in the SPSS manuals. Procedures and techniques discussed in the SPSS manuals, but not in this chapter, include:

- equiprobability models,
- linear-by-linear association models,
- row- and column-effects models,
- cell weights specification for models,
- tables with structural zeros,
- linear combinations of cells, and
- using contrasts.

For additional information about log-linear models that extends beyond the SPSS manuals, we refer you to Agresti (1996) and Wickens (1989).

# 27.4 Step by Step

## 27.4.1 General (Nonhierarchical) Log-Linear Models

*To access the initial SPSS screen from the Windows display, perform the following sequence of steps:*

| In Screen | Do This | Step 1 |
|---|---|---|
| 2.1 | ✳ → ✳ All Programs ▷ → ✳ IBM SPSS Statistics → ✳ IBM SPSS Statistics | Front1 |

*Mac users: To access the initial SPSS screen, successively click the following icons:*

| In Screen | Do This | Step 1a |
|---|---|---|
| 2.1 | ✳ → ✳ → ✳ | Front1 |

*After clicking the SPSS program icon, Screen 1 appears on the monitor.*

| Step 2 |
|---|
| *Create and name a data file or edit (if necessary) an already existing file (see Chapter 3).* |

*Screens 1 and 2 (on the inside front cover) allow you to access the data file used in conducting the analysis of interest. The following sequence accesses the* **helping3.sav** *file for further analyses:*

| In Screen | Do This | Step 3 |
|---|---|---|
| Front1 | ✳ File → Open → ✳ Data [*or* ✳ ] | |
| Front2 | type `helping3.sav` → ✳ Open [*or* ✳ ✳ helping3.sav] | Data |

Whether first entering SPSS or returning from earlier operations the standard menu of commands across the top is required. As long as it is visible you may perform any analyses. It is not necessary for the data window to be visible.

*After completion of Step 3 a screen with the desired menu bar appears. When you click a command (from the menu bar), a series of options will appear (usually) below the selected command. With each new set of options, click the desired item. The sequence to access general linear models begins at any screen with the menu of commands visible:*

| In Screen | Do This | Step 4 |
|---|---|---|
| Menu | ✳ Analyze → Loglinear → ✳ General | 27.1 |

After this sequence step, a window opens (Screen 27.1, next page) in which you specify the factors and covariates to include in the analysis. In the box at the left of Screen 27.1 is the list of variables in your data file. Any categorical variable that will be included in the analysis should be moved to the **Factor(s)** box to the right of the variable list, by clicking on the upper ➡. In the example in sequence Step 5 (page 349), we will examine **gender, ethnic**, and **cathelp**. Because these are all categorical variables, they will all be moved into the **Factor(s)** box.

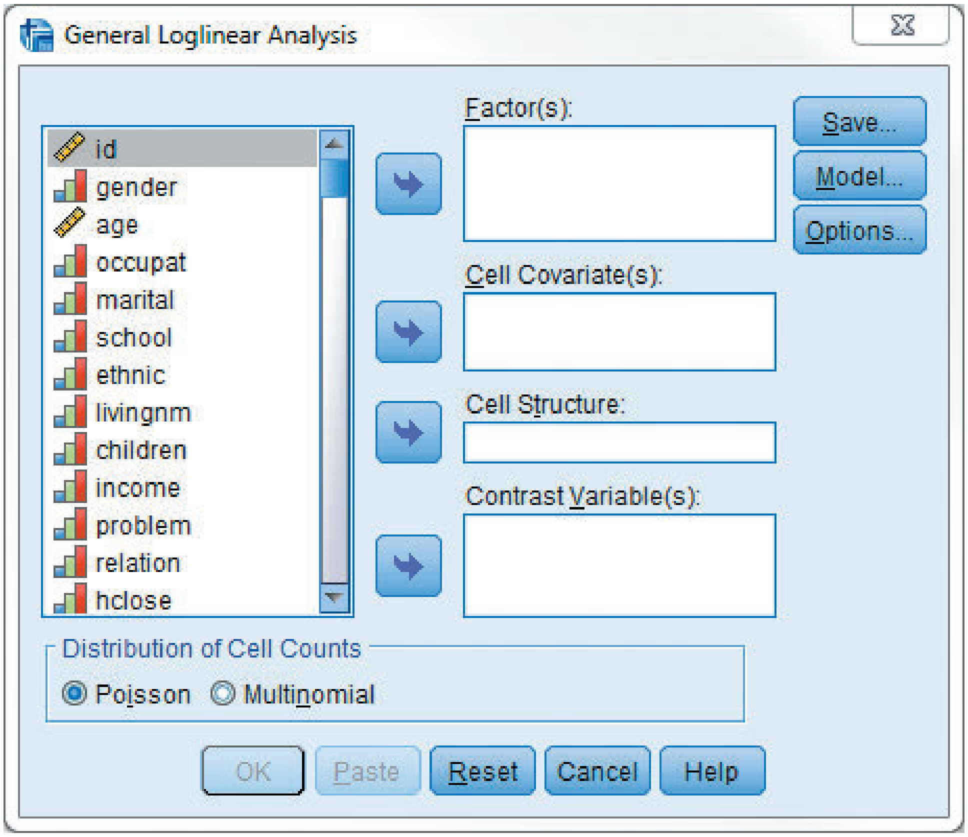
If you wish to examine the effects of any non-categorical variable on the categorical factor(s), then the variable should be moved into the **Cell Covariate(s)** box by selecting the variable and clicking on the second ➡ from the top. For example, if you wished to test for the influence of linear and cubed effects of **income**, the procedure would be:

- Click **Transform**, click **Compute**, then compute two new variables, **income1** (equal to **income**) and **income3** (equal to **income**$^3$).
- Access Screen 27.1, and enter **income** in the **Factor(s)** box. Then, enter **income1** and **income3** in the **Cell Covariate(s)** box.
- Be sure that when you select a model (Screen 27.3, the next page) you include the covariates in the model.

The **Cell Structure**, **Contrast Variable(s)**, **Distribution of Cell Counts**, and **Save** options on Screen 27.1 are not frequently used and extend beyond the scope of this book. The **Model** and **Options** buttons will, however, be used quite often. We begin our discussion with the **Options** button, used to specify desired output and criteria used in the analysis.

When the **Options** button is selected, Screen 27.2 (below) appears. The **Plot** options produce charts; these are useful if you are familiar with residuals analysis, or if you wish to examine the normal probabilities involved in the analysis. They will not be described in this chapter; for a discussion of residuals in general, see Chapter 28.

**Screen 27.1** General Loglinear Analysis Dialog Window



General Loglinear Analysis

id
gender
age
occupat
marital
school
ethnic
livingnm
children
income
problem
relation
hclose

Factor(s):

Cell Covariate(s):

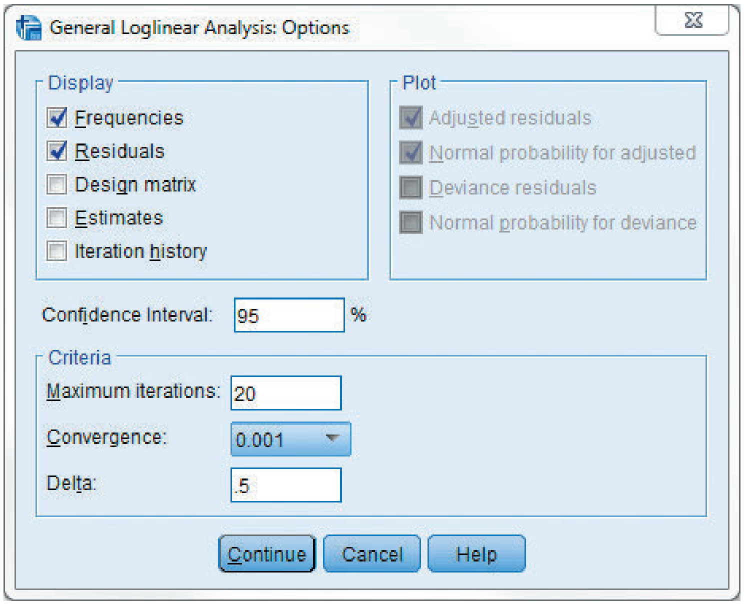
Cell Structure:

Contrast Variable(s):

Distribution of Cell Counts

● Poisson ○ Multinomial

**Screen 27.2** General Loglinear Analysis Options Dialog Window



General Loglinear Analysis: Options

Display

☑ Frequencies
☑ Residuals
☐ Design matrix
☐ Estimates
☐ Iteration history

Plot

☑ Adjusted residuals
☑ Normal probability for adjusted
☐ Deviance residuals
☐ Normal probability for deviance

Confidence Interval: 95 %

Criteria

Maximum iterations: 20

Convergence: 0.001

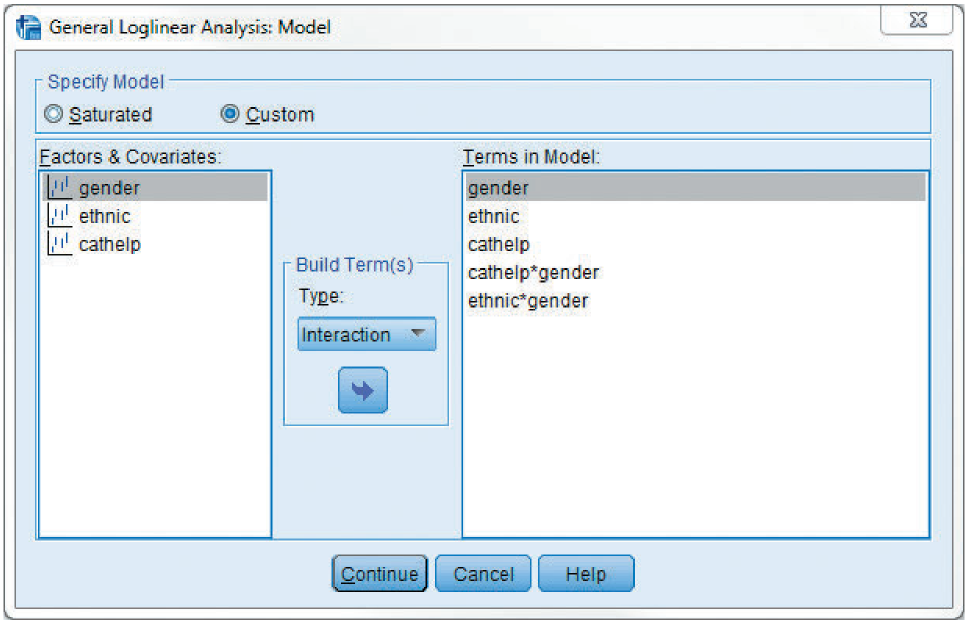
Delta: .5

The **Display** options are frequently used. **Frequencies** outputs tables of observed and estimated frequencies for all cells in the model. **Residuals** produces actual and standardized residuals from these cells, and **Estimates** produces parameter estimates for the model. The **Design matrix** option is rarely desired; it produces information about the internal design matrix used by SPSS in producing the log-linear model.

Occasionally, the **Criteria** options may be accessed, but usually these settings do not need to be changed. **Convergence** specifies the accuracy that must be attained in order to accept convergence of the model's equations. **Maximum iterations** specifies how long SPSS should keep iterating the model to achieve a fit.

By default, SPSS produces a saturated log-linear model, including all main effects and all interactions. This is often not desired; the saturated model perfectly predicts the data, but is no more parsimonious than the data. If you wish to specify a custom, non-saturated model, then select the **Model** button (on Screen 27.1) to activate Screen 27.3. When you click the **Custom** button, you may specify which terms (main effects and interactions) you want included in the model. For each main effect you wish to include, click on the variable name in the **Factors & Covariates** box to the left of the window, select **Main effects** from the drop-down menu in the **Build Term(s)** box, and click on the ➜ button. For each interaction you wish included, click on the variable names of the variables involved in the interaction (in the **Factors & Covariates** box), select **Interaction** from the drop-down menu in the **Build Term(s)** box, and click the ➜ button. To enter several interactions at once, you may select all variables you want in the interactions, and select **All 2-way, All 3-way, All 4-way**, or **All 5-way** interactions from the drop-down menu in the **Build Term(s)** box. Screen 27.3 shows the following custom model already specified: Main effects for **cathelp, ethnic**, and **gender**, plus **cathelp** × **gender** and **ethnic** × **gender** interactions.

**Screen 27.3** General Loglinear Model Specification Dialog Window



General Loglinear Analysis: Model

Specify Model

○ Saturated ● Custom

Factors & Covariates:

gender
ethnic
cathelp

Build Term(s)

Type: Interaction

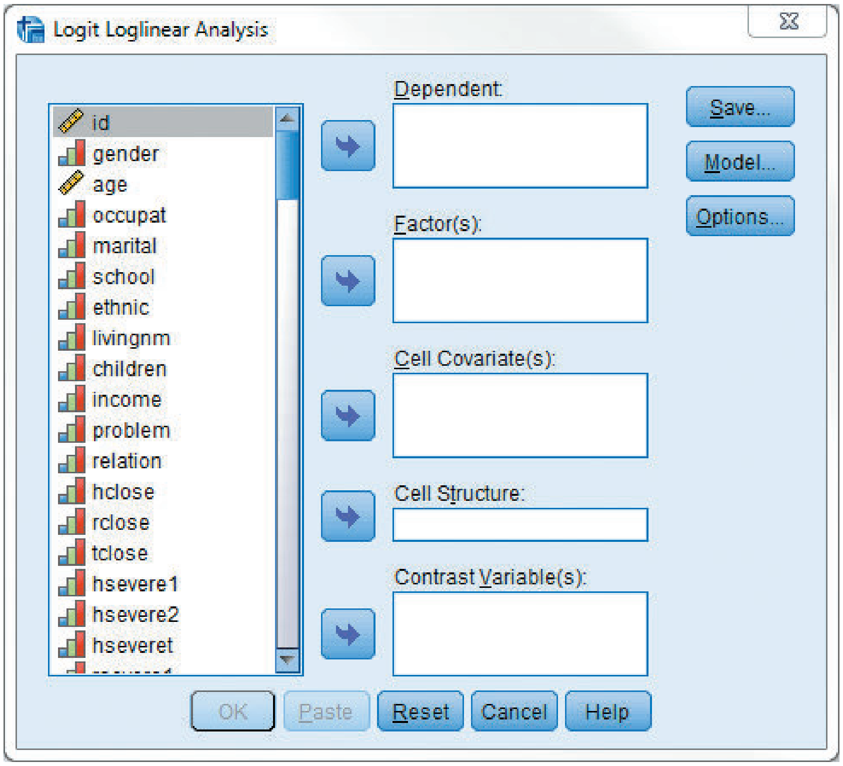
Terms in Model:

gender
ethnic
cathelp
cathelp*gender
ethnic*gender

SPSS has the ability to test logit models, in which one or more dichotomous dependent variables are included in the analysis. This procedure is virtually identical to the General Loglinear procedure and is selected through the **Analyze** → **Loglinear** → **Logit** sequence instead of the **Analyze** → **Loglinear** → **General** sequence.

This sequence calls up Screen 27.4. This procedure is nearly identical to the General Loglinear procedure, and Screen 27.4 is nearly identical to Screen 27.1. The primary difference is that this window has an additional box, **Dependent**, in which one or more categorical dependent variables may be placed.

**Screen 27.4** Logit Loglinear Analysis Window



*The following procedure instructs SPSS to test a log-linear model including the main effects of* **gender, ethnic**, *and* **cathelp**, *as well as interactive effects of* **gender** *by* **ethnic** *and* **gender** *by* **cathelp**.

| In Screen | Do This | Step 5 |
|---|---|---|
| 27.1 | 🖱 **gender** → 🖱 *top* ➡ → 🖱 **ethnic** → 🖱 *top* ➡ | |
| | → 🖱 **cathelp** → 🖱 *top* ➡ → 🖱 **Options** | |
| 27.2 | 🖱 **Estimates** → 🖱 **Continue** | |
| 27.1 | 🖱 **Model** | |
| 27.3 | 🖱 **Custom** → 🖱 ▼ *in* **Build Term(s)** *box* → 🖱 **Main effects** → 🖱 **gender** | |
| | → 🖱 ➡ → 🖱 **ethnic** → 🖱 ➡ → 🖱 **cathelp** → 🖱 ➡ → 🖱 ▼ *in* | |
| | **Build Term(s)** *box* → 🖱 **Interaction** → 🖱 **gender** → *hold* **Ctrl** *down then* | |
| | 🖱 **cathelp** 🖱 ➡ → 🖱 **gender** → *hold* **Ctrl** *down* 🖱 **ethnic** → 🖱 ➡ → | |
| | 🖱 **Continue** | |
| 27.1 | 🖱 **OK** | Back1 |

Upon completion of Step 5 the output screen will appear (Screen 1, inside back cover). All results from the just-completed analysis are included in the Output Navigator. Make use of the scroll bar arrows (▲▼▶◀) to view the results. Even when viewing output, the standard menu of commands is still listed across the top of the window. Further analyses may be conducted without returning to the data screen.

## 27.5 Printing Results

Results of the analysis (or analyses) that have just been conducted require a window that displays the standard commands (**File Edit Data Transform Analyze . . .**) across the top. A typical print procedure is shown below beginning with the standard output screen (Screen 1, inside back cover).

*To print results, from the Output screen perform the following sequence of steps:*

| In Screen | Do This | Step 6 |
|---|---|---|
| Back1 | *Select desired output or edit (see pages 16–22)* → ✳ **File** → ✳ **Print** | |
| 2.9 | *Consider and select desired print options offered, then* → ✳ **OK** | |

*To exit you may begin from any screen that shows the* **File** *command at the top.*

| In Screen | Do This | Step 7 |
|---|---|---|
| Menu | ✳ **File** → ✳ **Exit** | 2.1 |

Note: After clicking **Exit**, there will frequently be small windows that appear asking if you wish to save or change anything. Simply click each appropriate response.

## 27.6 Output

### 27.6.1 General (Nonhierarchical) Log-Linear Models

Below is an explanation of the key output from the General (nonhierarchical) Loglinear Model procedure (sequence Step 5, previous page). Only representative output is provided, because the actual printout from this procedure is many pages long.

Note: Because the interpretation of log-linear models with covariates (or designs that are logit models) is virtually identical to log-linear models without covariates (or that are not logit models), we include only one sample output. Logit models also produce an analysis of dispersion and measures of association. These measures indicate the dependent variable's dispersion, and how much of the total dispersion of the dependent variable comes from the model. These measures are difficult to interpret without considerable experience and so are not discussed here.

### 27.6.2 Goodness-of-Fit Tests

| | Value | df | Sig. |
|---|---|---|---|
| Likelihood Ratio | 8.281 | 8 | .406 |
| Pearson Chi-Square | 8.298 | 8 | .405 |

The likelihood-ratio chi-square and the Pearson chi-square examine the fit of the model. Large chi-square values and small $p$ values indicate that the model does *not* fit the data well. Note that this is the opposite of thinking in interpretation of most types of analyses. Usually one looks for a large test statistic and a small $p$ value to indicate a significant effect. In this case, a large chi-square and a small $p$ value would indicate that your data differ significantly (or do not fit well) from the model you have created. The degrees of freedom refers to the number of non-zero cells, minus the number of parameters in the model.

**Predicted and Actual Cell Counts and Residuals**

| gender | ethnic | cathelp | Observed | | Expected | | Residual | Standardized Residual | Adjusted Residual | Deviance |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Count | % | Count | % | | | | |
| FEMALE | WHITE | NOT HELPFUL | 70 | 13.0% | 74.123 | 13.8% | –4.123 | –.479 | –.920 | –.483 |
| | | HELPFUL | 95 | 17.7% | 90.877 | 16.9% | 4.123 | .433 | .920 | .429 |
| | BLACK | NOT HELPFUL | 13 | 2.4% | 15.274 | 2.8% | –2.274 | –.582 | –.829 | –.597 |
| | | HELPFUL | 21 | 3.9% | 18.726 | 3.5% | 2.274 | .525 | .829 | .515 |
| | HISPANIC | NOT HELPFUL | 22 | 4.1% | 22.462 | 4.2% | –.462 | –.097 | –.143 | –.098 |
| | | HELPFUL | 28 | 5.2% | 27.538 | 5.1% | .462 | .088 | .143 | .088 |
| | ASIAN | NOT HELPFUL | 28 | 5.2% | 20.215 | 3.8% | 7.785 | 1.731 | 2.513 | 1.635 |
| | | HELPFUL | 17 | 3.2% | 24.785 | 4.6% | –7.785 | –1.564 | –2.513 | –1.659 |
| | OTHER/ DTS | NOT HELPFUL | 13 | 2.4% | 13.926 | 2.6% | –.926 | –.248 | –.352 | –.251 |
| | | HELPFUL | 18 | 3.4% | 17.074 | 3.2% | .926 | .224 | .352 | .222 |
| MALE | WHITE | NOT HELPFUL | 75 | 14.0% | 71.849 | 13.4% | 3.151 | .372 | .892 | .369 |
| | | HELPFUL | 53 | 9.9% | 56.151 | 10.5% | –3.151 | –.420 | –.892 | –.425 |
| | BLACK | NOT HELPFUL | 8 | 1.5% | 8.981 | 1.7% | –.981 | –.327 | –.514 | –.334 |
| | | HELPFUL | 8 | 1.5% | 7.019 | 1.3% | .981 | .370 | .514 | .362 |
| | HISPANIC | NOT HELPFUL | 15 | 2.8% | 16.840 | 3.1% | –1.840 | –.448 | –.731 | –.457 |
| | | HELPFUL | 15 | 2.8% | 13.160 | 2.5% | 1.840 | .507 | .731 | .496 |
| | ASIAN | NOT HELPFUL | 15 | 2.8% | 14.033 | 2.6% | .967 | .258 | .415 | .255 |
| | | HELPFUL | 10 | 1.9% | 10.967 | 2.0% | –.967 | –.292 | –.415 | –.296 |
| | OTHER/DTS | NOT HELPFUL | 6 | 1.1% | 7.297 | 1.4% | –1.297 | –.480 | –.748 | –.496 |
| | | HELPFUL | 7 | 1.3% | 5.703 | 1.1% | 1.297 | .543 | .748 | .524 |

| Term | Definition/Description |
|---|---|
| **OBSERVED COUNT AND %** | The observed cell count and percentage derives from the data and indicates the number of cases in each cell. |
| **EXPECTED COUNT AND %** | The expected cell counts and percentages indicate the expected frequencies for the cells, based on the model being tested. |
| **RESIDUAL, STANDARDIZED, AND ADJUSTED RESIDUAL** | The residuals are the observed counts minus the expected counts. Standardized residuals are adjusted for the standard error; adjusted residuals are standardized residuals adjusted for *their* standard error. High residuals indicate that the model is not adequate. |
| **DEVIANCE** | The sign (positive or negative) of the deviance is always the same as the residual; the magnitude of the deviance (or, how far away it is from zero) is an indication of how much of the chi-square statistic comes from each cell in the model. |

**Parameter Estimates**

| Parameter | Estimate | Std. Error | Z | Sig. | 95% Confidence Interval | |
|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound |
| Constant | 1.741 | .288 | 6.044 | .000 | 1.176 | 2.305 |
| [cathelp = 1.00] | .247 | .138 | 1.781 | .075 | –.025 | .518 |
| [cathelp = 2.00] | $0^a$ | . | . | . | . | . |
| [ethnic = 1] | 2.287 | .291 | 7.857 | .000 | 1.717 | 2.858 |
| [ethnic = 2] | .208 | .373 | .556 | .578 | –.524 | .939 |
| [ethnic = 3] | .836 | .332 | 2.518 | .012 | .185 | 1.487 |
| [ethnic = 4] | .654 | .342 | 1.912 | .056 | –.016 | 1.324 |
| [ethnic = 5] | $0^a$ | . | . | . | . | . |
| [gender = 1] | 1.097 | .343 | 3.196 | .001 | .424 | 1.769 |
| [gender = 2] | | . | . | . | . | . |
| [gender = 1] * [cathelp = 1.00] | –.450 | .178 | –2.533 | .011 | –.799 | –.102 |
| [gender = 1] * [cathelp = 2.00] | $0^a$ | . | . | . | . | . |

*(Continued)*

| Parameter | Estimate | Std. Error | Z | Sig. | 95% Confidence Interval | |
|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound |
| [gender = 2] * [cathelp = 1.00] | 0[a] | . | . | . | . | . |
| [gender = 2] * [cathelp = 2.00] | 0[a] | . | . | . | . | . |
| [gender = 1] * [ethnic = 1] | −.615 | .351 | −1.754 | .080 | −1.303 | .072 |
| [gender = 1] * [ethnic = 2] | −.115 | .448 | −.257 | .797 | −.994 | .764 |
| [gender = 1] * [ethnic = 3] | −.358 | .403 | −.889 | .374 | −1.148 | .432 |
| [gender = 1] * [ethnic = 4] | −.281 | .414 | −.679 | .497 | −1.093 | .530 |
| [gender = 1] * [ethnic = 5] | 0[a] | . | . | . | . | . |
| [gender = 2] * [ethnic = 1] | 0[a] | . | . | . | . | . |
| [gender = 2] * [ethnic = 2] | 0[a] | . | . | . | . | . |
| [gender = 2] * [ethnic = 3] | 0[a] | . | . | . | . | . |
| [gender = 2] * [ethnic = 4] | 0[a] | . | . | . | . | . |
| [gender = 2] * [ethnic = 5] | 0[a] | . | . | . | . | . |

[a] This parameter is set to zero because it is redundant.

Parameter estimates are provided for each effect. Because some of the parameters are not necessary to predict the data, many parameters are zero (that is, they really aren't part of the model). For example, if you know the parameter for **cathelp** = 1, you don't need to know the parameter for **cathelp** = 2, because people who are not in the **cathelp** = 1 condition will be in the **cathelp** = 2 condition.

| Term | Definition/Description |
|---|---|
| **ESTIMATE** | The $\lambda$ value in the log-linear model equation. |
| **STANDARD ERROR** | A measure of the dispersion of the coefficient. |
| **Z-VALUE** | A standardized measure of the parameter coefficient: Large $z$ values (those whose absolute value is greater than 1.96) are significant ($\alpha = .05$). |
| **LOWER AND UPPER 95% CONF INTERVAL** | There is a 95% chance that the coefficient is between the lower and upper confidence interval. |

## Chapter 28

# Residuals: Analyzing Left-Over Variance

THIS IS a somewhat unusual chapter. It does not contain step-by-step instructions for performing a particular analysis in SPSS (although it does include one- or two-line sequences to reproduce graphs illustrated in this chapter); indeed, it doesn't even deal with a single SPSS procedure. Instead, it deals with residuals, a statistical concept that is present in many SPSS procedures. The crosstabs, ANOVA and MANOVA models, regression, and log-linear model procedures can all analyze residuals. This chapter takes a didactic format, rather than step-by-step format, because the analysis of residuals takes quite a bit of statistical finesse and experience. Because of this it is useful to examine the way residuals work separately, so that the common aspects of residuals in all applicable SPSS procedures can be examined at once. If residuals were discussed in each procedure in which they are present, it would add several pages to each chapter, rather than several pages for the entire book. Also, for those with little experience in analyzing residuals (or little desire to gain experience), placing the discussion of residuals throughout the book would likely confuse rather than clarify.

With that apology, we begin. This chapter commences with a brief description of residuals, and their primary usefulness. The bulk of the chapter presents two case studies, in which a linear regression and a general log-linear model are examined. Each case study illustrates a different use of residuals. The chapter concludes with a summary of the various methods of analyzing residuals within SPSS, along with the ways of accessing the appropriate output.

## 28.1 Residuals

Most of the sophisticated statistical procedures commonly done are designed to test a theoretical model. For example:

- In 2 × 2 chi-square analyses, expected frequencies are computed; these expected frequencies are based on a model with the effects of the two categorical variables independent: **Expected frequencies = effects of the categorical variables + residuals**.
- In log-linear modeling, the relationship between categorical variables is examined, and a model is developed containing main effects and interactions between the variable: **Expected frequencies main effects + interactions + residuals**.
- In regression analyses, one or more predictor variables predict a criterion or dependent variable. The regression procedure examines whether or not each predictor variable significantly relates to the dependent variable. Thus, a model of predictor variables is developed: **Dependent variable = effects of predictor variables + residual**.
- In an ANOVA, some main effects and interactions are examined. These main effects and interactions may be considered a model of independent variable(s) that influence the dependent variable(s). It is with this understanding that SPSS

groups most of the ANOVA procedures under the heading "General Linear Model": **Dependent variable(s) = main effects + interactions + covariates + residuals**.

In all of these techniques, models are developed to predict the data that have been collected. The goal is to develop a model that predicts the data perfectly, and parsimoniously. That is, the best model will predict the data collected, and also be very simple. Because of this desire for parsimony, there is generally some discrepancy between the actual, observed data, and the expected data predicted by the model. This discrepancy is the residual, or "error." The analysis of residuals can be useful, for instance, for determining how good the model is that you are testing. It can also be useful in examining unusual cases in your data that do not fit the model well, or in finding patterns in the data that you hadn't noticed or predicted.

Our first case study will test a simple linear regression analysis to determine the adequacy of the model developed by SPSS.

## 28.2 Linear Regression: A Case Study

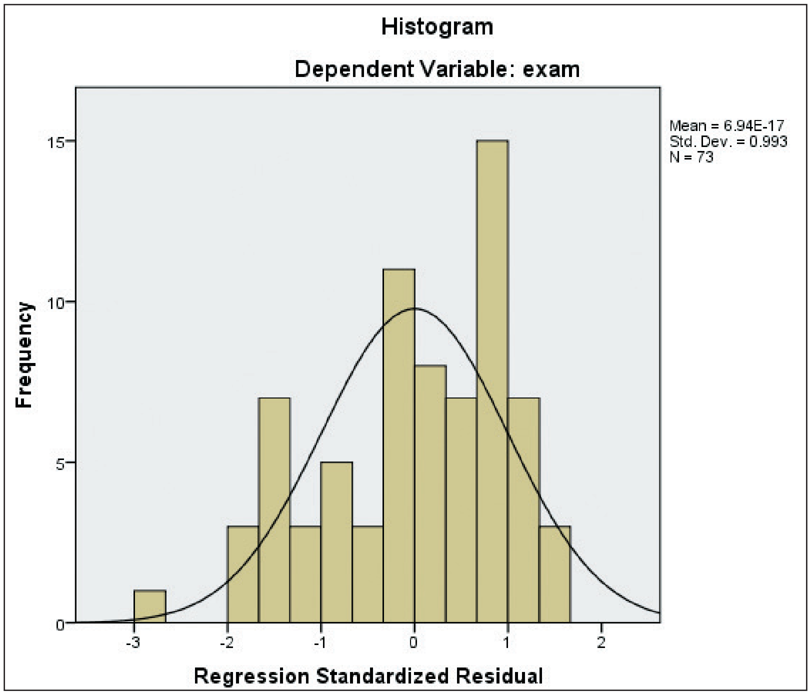
A researcher hypothesizes that students with low anxiety do poorly on an exam and those with high anxiety do better. This is the same example used in Chapter 15. The researcher expects a linear relationship. The description of the model developed is described fully in Chapter 15, pages 191–195, and the SPSS procedure for developing this model is described in Steps 4 and 5 in that chapter. In summary, the model is that:

$$\textbf{exam}_{(\text{true})} = \text{some } \textbf{constant} + \text{a coefficient} \times \textbf{anxiety} + \text{residual}$$

or

$$\textbf{exam}_{(\text{true})} = 64.247 + 2.818\textbf{(anxiety)} + \textbf{residual}$$

In order to examine how well this model of anxiety predicts exam scores, we need to look at the residual values for all of the cases to see if there is any pattern to the residuals. In this case, we can request a histogram, normal probability plot, as well as customized graphs relating a number of different forms (e.g., standardized and studentized) of the predicted values, residuals, and dependent measure. Graphs are requested from the plots window, selected from the Linear Regression window in Screen 15.3.

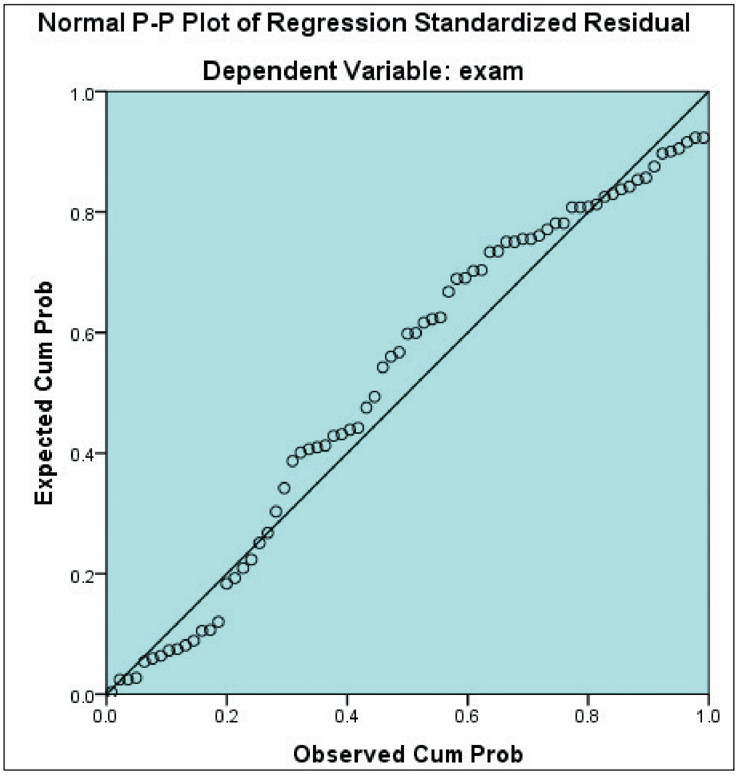


The histogram (at right) displays the standardized residuals (the residuals divided by the estimate of the standard error) across the horizontal axis, and the number of subjects within each range of standardized residuals along the vertical axis. The standard deviations will always be 1 (or practically 1, due to rounding error) and the mean will always be 0. This is because the residuals are standardized (therefore, the standard deviation is 1) and the model is based on the data (thus, the mean is 0).

**anxiety.sav:** Sequence: analyze → regression → linear → exam (DV) → anxiety (IV) → plots → histogram → continue → OK

A line is drawn on the histogram, which shows where the normal curve predicts the residual to fall. What is important in the analysis of residuals is any pattern in the relationship between the histogram bars and the normal probability curve. In this case, the bars on the left side of the graph tend to be lower than the normal curve, and the bars on the right side of the graph tend to be higher than the normal curve. This indicates that there may be a curvilinear relationship between the anxiety and exam scores.

The normal probability plot (below) takes a bit more explanation, but in this case makes the curvilinear relationship between the anxiety and exam scores even clearer. The plot places the observed cumulative probability along the horizontal axis and the expected cumulative probability along the vertical axis. Cumulative probabilities are calculated by ranking all of the subjects, and then comparing the value of that variable with the value expected from a normal distribution. If the residuals are normally distributed (that's what we want with a good model), then there should be a linear relationship between the expected and observed cumulative probabilities. Any other pattern suggests a problem or weakness in the model.



In our example, we can see that values near the middle (probabilities around .5) tend to have the higher expected values than observed values, and the values near the ends (probabilities near 0 or 1) tend to have lower expected values than observed values. This indicates the possibility of a nonlinear relationship between **anxiety** and **exam**. An analysis of the curvilinear relationship between **anxiety** and **exam** (in Chapter 15) indeed reveals that the inclusion of the **anxiety**$^2$ variable improves the fit of the model.

Other nonlinear patterns on the normal probability plot will indicate different problems in the regression equation. Nearly any pattern is possible and it is beyond the scope of this book to describe them all, but the histogram and the normal probability plot are good places to look for these patterns.

**anxiety.sav:** Sequence: analyze ⟶ regression ⟶ linear ⟶ exam (DV) ⟶ anxiety (IV) ⟶ plots ⟶ normal probability plot ⟶ continue ⟶ OK

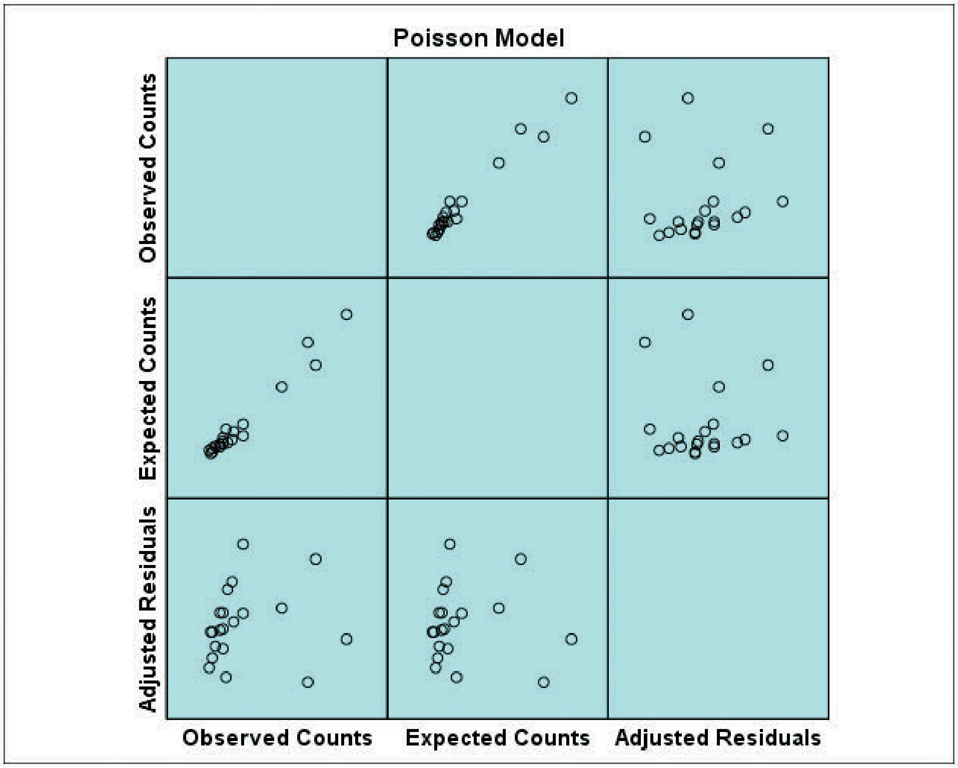
## 28.3 General Log-Linear Models: A Case Study

In addition to looking for curvilinear relationships present in data but not in the model being examined, the examination of residuals can be useful in searching for additional variables that should be included in a model. This is particularly true in log-linear modeling, in which the development of the model is a fairly complex procedure.

In this case study we will examine a model based on the **helping3.sav** file. This model will include main effects of **gender**, **ethnicity**, and **cathelp** (referring to whether or not help was helpful, a dichotomous variable), and an interactive effect of **gender** by **cathelp**. Residuals-related output may be requested by selecting **Display Residuals** and all four of the **Plot** options in the **Options** window (Screen 27.2). Note that plots are not available if a saturated model is selected (i.e., a model containing all main effects and interactions), because a saturated model predicts the data perfectly, and thus all residuals are equal to zero.

| Gender | Ethnicity | Cathelp | Residual | Adj. Residuals | Dev. Residuals |
|---|---|---|---|---|---|
| MALE | White | Helpful | –9.66 | –1.88 | –1.11 |
| | | Not Helpful | –2.67 | –.49 | –.27 |
| | Black | Helpful | –.59 | –.20 | –.16 |
| | | Not Helpful | 4.33 | 1.37 | 1.02 |
| | Hispanic | Helpful | .25 | .07 | .05 |
| | | Not Helpful | 1.33 | .34 | .26 |
| | Asian | Helpful | 8.97 | 2.58 | 1.92 |
| | | Not Helpful | –6.33 | –1.72 | –1.38 |
| | Other/DTS | Helpful | 1.04 | .37 | .30 |
| | | Not Helpful | 3.33 | 1.11 | .84 |
| FEMALE | White | Helpful | 10.07 | 2.10 | 1.22 |
| | | Not Helpful | 2.26 | .52 | .31 |
| | Black | Helpful | –3.08 | –1.10 | –.97 |
| | | Not Helpful | –.66 | –.26 | –.23 |
| | Hispanic | Helpful | –2.73 | –.80 | –.67 |
| | | Not Helpful | 1.15 | .37 | .30 |
| | Asian | Helpful | –.51 | –.16 | –.13 |
| | | Not Helpful | –2.12 | –.72 | –.63 |
| | Other/DTS | Helpful | –3.75 | –1.42 | –1.29 |
| | | Not Helpful | –.62 | –.26 | –.23 |

The residuals output includes a list of residuals, adjusted residuals (like the standardized residuals), and the deviance residuals (based on the probability for being in a particular group). Although there are no cells in this table with very high adjusted residuals (the highest is 2.58), there are quite a number of cells with somewhat high adjusted residuals. It appears that there may be an interaction between **gender** and **ethnic**, because the pattern of residuals is quite different across ethnicities for males and females.

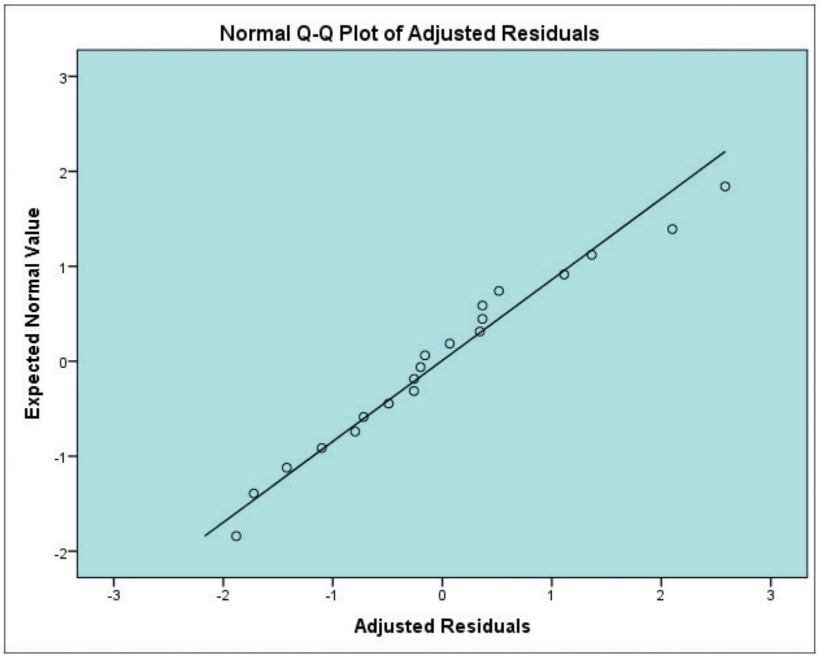


The adjusted residuals plot (above) is, at first glance, quite confusing. Upon closer examination, however, it is manageable. It is essentially a correlation table with scatterplots instead of correlation values within each cell. As in a correlation table, cells below the diagonal are reflected in cells above the diagonal. The diagonal cells identify each row and column. So, the left center cell is a scatterplot of observed and expected cell counts, the bottom left cell is a scatterplot of observed cell counts with adjusted residual, and the middle cell at the bottom shows expected counts with adjusted residuals. The other scatterplots are mirror images of these bottom left three scatterplots.

**Helping3.sav:** Sequence: follow the entire sequence on page 349 (Chapter 27) but do not include the gender by ethnic interaction, Then, click Options ⟶ Deviance residuals ⟶ Normal Probability for deviance ⟶ Continue ⟶ OK

Two particularly important plots are the observed- versus expected-cell-count plot, and the expected-cell-count versus adjusted-residual plot. A good model will produce a linear relationship between the observed and the expected cell counts—that is, the observed data will match the data predicted by the model. A good model wouldn’t show a relationship between the expected cell count and the adjusted residual, either. If there is a relationship between these variables, then another variable may need to be added to the model, or else some assumption of the model may be violated (like homogeneity of variance).

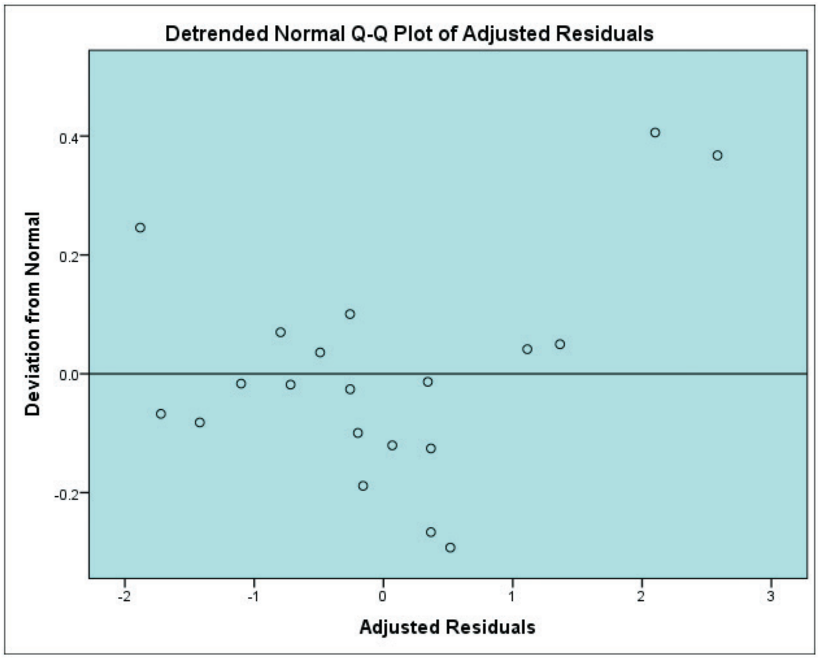
In this example, there is a generally linear relationship between the observed and the expected values (though not perfectly linear). There does, however, appear to be a pattern in the scatterplot between expected cell counts and the adjusted residuals.



The normal plot of adjusted residuals above displays adjusted residuals and expected normal values of those residuals. A good model should produce a linear relationship between expected normalized residuals and the actual residuals. Note that this graph is similar to the bottom center cell of the figure on the previous page, except that the expected cell counts have been standardized. In this case, because there is little nonlinear trend present, no problems are present. If there was a nonlinear trend present, that pattern in the residuals could suggest problems or inadequacies in the model.

**Helping3.sav:** Sequence on page 357 also produces this chart

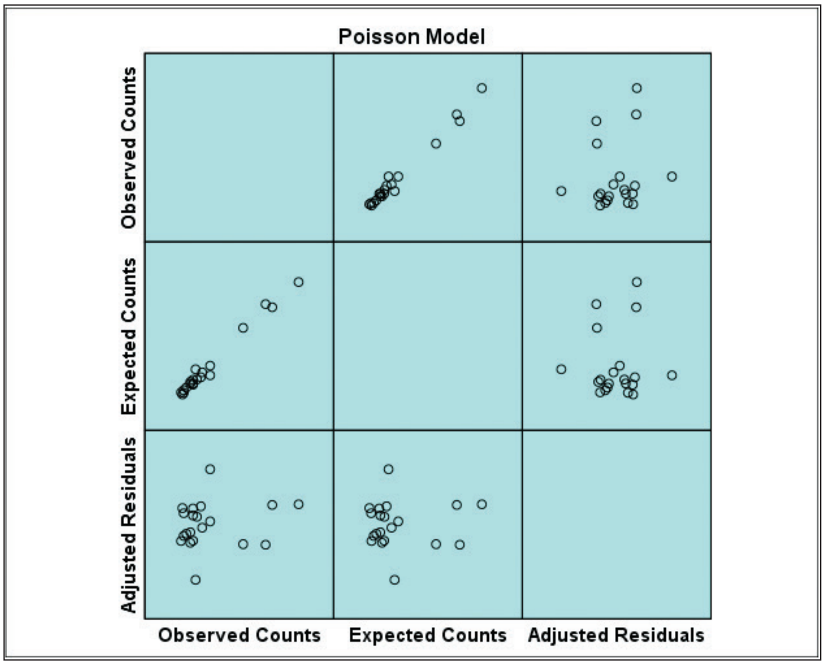
It does appear that the points to the far right of the graph have higher residual than expected, indicating that there may be something unusual going on with those cells. However, the cells are probably not far enough away from the trend line to be a problem.



The detrended normal plot of adjusted residuals, at the bottom of the previous page, is a variation on the normal plot of adjusted residuals, at the top of the previous page. In this case, however, the diagonal trend line in the chart is moved to a horizontal line, and the vertical axis graphs how far away the cell is from the normal trend line. This serves to magnify any deviation away from the normal value and is particularly useful in examining small, subtle patterns that may be present in a normal plot of residuals in which no cells are very far from the normal trend line (as in our example).

**Helping3.sav:** Sequence on page 357 also produces this chart

Note that the two "outlying" cells clearly stand out, on the top right of the chart. However, an examination of the scale on the vertical axis points out that even these cells are actually quite close to the expected normal values.



Based on an examination of the pattern of adjusted residuals, another log-linear model was tested, adding an interaction of **gender × ethnicity**. This model produced smaller adjusted residuals overall, although there was still one cell with a relatively high adjust residual (2.51). As can be seen in the new adjusted residuals plot (above), the plot of observed versus expected cell counts is even more linear than before, and the plot of expected cell count versus adjusted residuals (as well as observed counts versus adjusted residual) does not show as strong a pattern as before. Adding the **gender × ethnicity** interaction did, indeed, improve the model.

**Helping3.sav:** Follow sequence on page 357, then Model ⟶ gender ⟶ ethnic ⟶ interaction ⟶ paste ⟶ continue ⟶ OK

It should be noted that the residual list and plots of this model are a little bit too symmetrical and perfect. This is because the authors created the cathelp variable to provide a good fit for the model tested. So if you examine these plots, don't read too much meaningful significance into the patterns present.

## 28.4 Accessing Residuals in SPSS

We have only begun to explain the analysis of residuals. Many SPSS procedures allow you to save residuals in your data file, using new variable names; this allows you to perform any statistical analysis on those residuals that you desire.

We conclude our brief foray into residuals with a table summarizing the various SPSS options to analyze residuals. For each entry in the table, we present the chapter that describes the SPSS procedure, the window and options within that window needed to access residuals, and a brief description of the output produced.

| Chapter | Window | Options | Descriptions |
|---|---|---|---|
| **8**<br>**Crosstabs** | Cells<br>8.2 | **Unstandardized**<br>**Standardized**<br>**Adjusted standardized** | Produces residuals for each cell in the crosstabs table. Note that standardized residuals are transformed so that there is a mean of 0 and a standard deviation of 1, and the adjusted standardized residual is transformed to indicate the number of standard deviations above or below the mean. |
| **15, 16**<br>**Regression** | Save<br>16.3 | **Unstandardized**<br>**Standardized**<br>**Studentized**<br>**Deleted**<br>**Studentized deleted** | Saves various kinds of residuals in the data file, with new variable names. SPSS will produce the new variable names, and the output will define them. Studentized residuals are standardized on a case-by-case basis, depending on the distance of the case from the mean. Deleted residuals are what the residual for a case would be if the case were excluded from the analysis. |
| | Plots<br>16.5 | **Histogram**<br>**Normal probability plot**<br>And customized plots of the dependent variable with various kinds of residuals | Produces histograms and normal probability plots, as well as letting you produce scatterplots for any two of the following: the dependent variable, the standardized predicted values and residuals, the deleted residuals, the adjusted predicted residuals, studentized residuals, and deleted residuals. |
| **23, 24**<br>**General Linear Model**<br>**(MANOVA)** | Options<br>23.5 | **Residual plot**<br>**Residual SSCP matrix** | Residual plot produces for each dependent variable scatterplots for the observed, predicted, and standardized residuals. The residual sum of squares and the cross-product matrix allows you to examine if there are any correlations between the dependent variables' residuals. |
| **25**<br>**Logistic Regression** | Options<br>25.3 | **Casewise listing of residuals** | Lists residuals for each case, for either all cases or outliers more than a specified number of standard deviations from the mean (the default is 2). |

| Chapter | Window | Options | Descriptions |
|---|---|---|---|
| **25 Logistic Regression (Continued)** | Save | **Residuals:**<br>**Unstandardized**<br>**Logit**<br>**Studentized**<br>**Standardized**<br>**Deviance** | Saves various kinds of residuals in the data file, with new variable names. SPSS will choose the new variable names. Studentized residuals are standardized on a case-by-case basis, depending on the distance of the case from the mean. Deviance residuals are based on the probability of a case being in a group. Logit residuals are residuals predicted on logit scales. |
| **26**<br>**Hierarchical Log-linear Models** | Options<br>26.3 | **Under Plot**<br>**Residuals**<br>**Normal Probability** | Displays residuals (including standardized residuals), produces scatterplots of expected and observed values with residuals, and normal probability plots (see Glossary for a description of normal probability plots). |
| **27**<br>**General Log-linear Models** | Options<br>27.2 | **Under Plot**<br>**Adjusted residuals**<br>**Normal probability for adjusted**<br>**Deviance residuals**<br>**Normal probability: for deviance** | Displays residuals for each case, and produces plots of adjusted residuals, and normal and detrended normal probability plots for adjusted residuals and deviance residuals. Note: Plots are only available when a non-saturated model is tested. |
| | Save | **Residuals**<br>**Standardized residuals**<br>**Adjusted residuals**<br>**Deviance residuals** | Saves various kinds of residuals in the data file with new variable names. SPSS will select the new variable names. |

Note: If you want to analyze residuals for an ANOVA, you must use the GLM command.

# Data Files

Available for download from **www.spss-step-by-step.net** are 11 different data files that have been used to demonstrate procedures in this book. There is also a 12th file not utilized to demonstrate procedures in the book but included on the website (and in the Instructor's Manual) for additional exercises. There are also data files for all of the exercises, with the exception of the exercises for Chapter 3 (as the whole point of those exercises is to practice entering data!). You can tell from the name of the data file which exercise it goes with: for example, **ex14-1.sav** is the dataset for Chapter 14, Exercise 1. The **grades.sav** file is the most thoroughly documented and demonstrates procedures in 16 of the 28 chapters. This file is described in detail in Chapters 1 and 3. For analyses described in the other 12 chapters, it was necessary to employ different types of data to illustrate. On the website, *all* files utilized in this book are included. What follows are brief narrative introductions to each file, and when appropriate, critical variables are listed and described. Before presenting this, we comment briefly on how to read a data file from an external source (as opposed to reading it from the hard drive of your computer, as is illustrated in all chapters of this book).

In every one of Chapters 6 through 27, there is a sequence Step 3 that gives instructions on how to access a particular file. The instructions assume that you are reading a file saved in your hard drive. Depending on your computer set-up, you may have downloaded the files to your hard drive or to an external source (memory stick, CD, external drive). In case you find downloading files intimidating, here's a step by step to assist you:

## Do This

- type **www.spss-step-by-step.net** *in the address box of the internet browser*
- press **ENTER**
- 🖱 **Data Sets** under the SPSS 23 version of this book
- *right-*🖱 **Complete Data Sets [.zip]** *(or whatever file or files you wish to download)*
- *From the drop-down menu that appears* 🖱 **Save Target as***. . .*
- *If you don't want the files saved to the default location selected by your computer, then select the drive or file in which you wish to save the data sets, then* 🖱 **Save**

**grades.sav** The complete data file is reproduced on pages 55–57. This is a fictional file ($N = 105$) created by the authors to demonstrate a number of statistical procedures. This file is used to demonstrate procedures in Chapters 3–14, 17, and 23–24. In addition to the data, all variable names and descriptions are also included on page 54. Be aware that in addition to the key variable names listed there, additional variables are included in the file:

- **total** Sum of the five quizzes and the final
- **percent** The percent of possible points in the class
- **grade** The grade received in the class (A, B, C, D, or F)
- **passfail** Whether or not the student passed the course (P or F)

**graderow.sav** This file includes 10 additional subjects with the same variables as those used in the grades.sav file. It is used to demonstrate merging files in Chapter 4.

**gradecol.sav** This file includes the same 105 subjects as the **grades.sav** file but includes an additional variable **(IQ)** and is used in Chapter 4 to demonstrate merging files.

**anxiety.sav** A fictional data file ($N = 73$) that lists values to show the relationship between pre-exam anxiety and exam performance. It is used to demonstrate simple linear and curvilinear regression (Chapter 15). It contains two variables:

- **exam** The score on a 100-point exam
- **anxiety** A measure of pre-exam anxiety measured on a *low* (1) to *high* (10) scale

**helping1.sav** A file of real data ($N = 81$) created to demonstrate the relationship between several variables and the amount of time spent helping a friend in need. It is used to demonstrate multiple regression analysis (Chapter 16). Although there are other variables in the file, the ones used to demonstrate regression procedures include:

- **zhelp** *Z* scores of the amount of time spent helping a friend on a −3 to +3 scale
- **sympathy** Helper's sympathy in response to friend's need: *little* (1) to *much* (7) scale
- **anger** Anger felt by helper in response to friend's need; same 7-point scale
- **efficacy** Self-efficacy of helper in relation to friend's need; same scale
- **severity** Helper's rating of the severity of the friend's problem; same scale
- **empatend** Helper's empathic tendency measured by a personality test; same scale

**helping2.sav** A file of real data ($N = 517$) dealing with issues similar to those in the **helping1.sav** file. Although the file is large (both in number of subjects and number of variables), only the 15 measures of self-efficacy and the 14 empathic tendency questions are used to demonstrate procedures: reliability analysis (Chapter 18) and factor analysis (Chapter 20). (The full dataset is included in **helping2a.sav**; the file has been reduced to work with the student version of SPSS.) Variable names utilized in analyses include:

- **effic1 to effic15** The 15 self-efficacy questions used to demonstrate factor analysis
- **empathy1 to empath14** The 14 empathic tendency questions used to demonstrate reliability analysis

**helping3.sav** A file of real data ($N = 537$) dealing with issues similar to the previous two files. This is the same file as **helping2.sav** except it has been expanded by 20 subjects and all missing values in the former file have been replaced by predicted values. Although the *N* is 537, the file represents over 1,000 subjects because both the helpers and help recipients responded to questionnaires. In the book these data are used to demonstrate logistic regression (Chapter 25) and log-linear models (Chapters 26 and 27). We describe it here in greater detail because it is a rich data set that is able to illustrate every procedure in the book. Many exercises from this data set are included at the end of chapters and in the Instructor's Manual. Key variables include:

- **thelplnz** Time spent helping (*z* score scale, −3 to +3)
- **tqualitz** Quality of the help given (*z* score scale, −3 to +3)

- **tothelp** A help measure that weights time and quality equally ($z$ score scale, –3 to +3)
- **cathelp** A coded variable; 1 = the help was not helpful ($z$ score for **tothelp** $< 0$); 2 = the help was helpful ($z$ score for **tothelp** $> 0$)
- **empahelp** Amount of time spent in empathic helping [scale: *little* (1) to *much* (10)]
- **insthelp** Amount of time spent in instrumental (doing things) helping (same scale)
- **infhelp** Time spent in informational (e.g., giving advice) helping (same scale)
- **gender** 1 = female, 2 = male
- **age** Ranges from 17 to 89
- **school** 7-point scale; from 1 (lowest level education, < 12 yr) to 7 (the highest, > 19 yr)
- **problem** Type of problem: 1 = goal disruptive, 2 = relational, 3 = illness, 4 = catastrophic

[All variables that follow are scored on 7-point scales ranging from *low* (1) to *high* (7).]

- **angert** Amount of anger felt by the helper toward the needy friend
- **effict** Helper's feeling of self-efficacy (competence) in relation to the friend's problem
- **empathyt** Helper's empathic tendency as rated by a personality test
- **hclose** Helper's rating of how close the relationship was
- **hcontrot** Helper's rating of how controllable the cause of the problem was
- **hcopet** Helper's rating of how well the friend was coping with his or her problem
- **hseveret** Helper's rating of the severity of the problem
- **obligat** The feeling of obligation the helper felt toward the friend in need
- **sympathi** The extent to which the helper felt sympathy toward the friend
- **worry** Amount the helper worried about the friend in need

**graduate.sav** A fictitious data file ($N = 50$) that attempts to predict success in graduate school based on 17 classifying variables. This file is utilized to demonstrate discriminant analysis (Chapter 22). All variables and their metrics are described in detail in that chapter.

**grades-mds.sav** A fictitious data file ($N = 20$) is used in the multidimensional scaling chapter (Chapter 19). For variables **springer** through **shearer**, the rows and columns of the data matrix represent a 20 × 20 matrix of disliking ratings, in which the rows (cases) are the raters and the variables are the ratees. For variables **quiz1** through **quiz5**, these are quiz scores for each student. Note that these names and quiz scores are derived from the **grades.sav** file.

**grades-mds2.sav** A fictitious data file used to demonstrate individual differences multidimensional scaling (Chapter 19). This file includes four students' ratings of the similarity between four television shows. The contents and format of this data file are described on page 245.

**dvd.sav** A fictitious data file ($N = 21$) that compares 21 different brands of DVD players on 21 classifying features. The fictitious brand names and all variables are described in detail in the cluster analysis chapter (Chapter 21).

**divorce.sav** This is a file of 229 divorced individuals recruited from communities in central Alberta. The objective of researchers was to identify cognitive or interpersonal factors that assisted in recovery from divorce. No procedures in this book utilize this file, but there are a number of exercises at the end of chapters and in the Instructor's Manual that do. Key variables include:

- **lsatisy** A measure of life satisfaction based on weighted averages of satisfaction in 12 different areas of life functioning. This is scored on a 1 (low satisfaction) to 7 (high satisfaction) scale.
- **trauma** A measure of the trauma experienced during the divorce recovery phase based on the mean of 16 different potentially traumatic events, scored on a 1 (low trauma) to 7 (high trauma) scale.
- **sex** Gender [*women* (1), *men* (2)]
- **age** Range from 23 to 76
- **sep** Years separated accurate to one decimal
- **mar** Years married prior to separation, accurate to one decimal
- **status** Present marital status [*married* (1), *separated* (2), *divorced* (3), *cohabiting* (4)]
- **ethnic** Ethnicity [*White* (1), *Hispanic* (2), *Black* (3), *Asian* (4), *other or DTS* (5)]
- **school** [*1–11 yr* (1), *12 yr* (2), *13 yr* (3), *14 yr* (4), *15 yr* (5), *16 yr* (6), *17 yr* (7), *18 yr* (8), *19+* (9)]
- **childneg** Number of children negotiated in divorce proceedings
- **childcst** Number of children presently in custody
- **income** [*DTS* (0), *<10,000* (1), *10–20* (2), *20–30* (3), *30–40* (4), *40–50* (5), *50+* (6)]
- **cogcope** Amount of cognitive coping during recovery [*little* (1) to *much* (7)]
- **behcope** Amount of behavioral coping during recovery [*little* (1) to *much* (7)]
- **avoicop** Amount of avoidant coping during recovery [*little* (1) to *much* (7)]
- **iq** Intelligence or ability at abstract thinking [*low* (1) to *high* (12)]
- **close** Amount of physical (nonsexual) closeness experienced [*little* (1) to *much* (7)]
- **locus** Locus of control [*external locus* (1) to *internal locus* (10)]
- **asq** Attributional style questionnaire [*pessimistic style* (≈–7) to *optimistic style* (≈+9)]
- **socsupp** Amount of social support experienced [*little* (1) to *much* (7)]
- **spiritua** Level of personal spirituality [*low* (1) to *high* (7)]
- **d-variables** There are 16 variables (all beginning with the letter "d") that specify 16 different areas that may have been destructive in their efforts at recovery. Read labels in the data file to find what each one is. Great for bar chart, *t* test, or reliability demonstrations.
- **a-variables** There are 13 variables (all beginning with the letter "a") that specify 13 different areas that may have assisted in their efforts toward recovery. Read labels in the data file to find what each one is. Great for bar chart, *t* test, or reliability demonstrations.

# Glossary

**adjusted *r* square** In multiple regression analysis, $R^2$ is an accurate value for the sample drawn but is considered an optimistic estimate for the population value. The Adjusted $R^2$ is considered a better population estimate and is useful when comparing the $R^2$ values between models with different numbers of independent variables.

**agglomerative hierarchical clustering** A procedure used in cluster analysis by which cases are clustered one at a time until all cases are clustered into one large cluster. See Chapter 25 for a more complete description.

**alpha** Also, coefficient alpha or $\alpha$ is a measure of internal consistency based on the formula $\alpha = rk/[1 + (k - 1)r]$, where $k$ is the number of variables in the analysis and $r$ is the mean of the inter-item correlations. The alpha value is inflated by a larger number of variables, so there is no set interpretation as to what is an acceptable alpha value. A rule of thumb that applies to most situations is: $\alpha > .9$—excellent, $\alpha > .8$—good, $\alpha > .7$—acceptable, $\alpha > .6$—questionable, $\alpha > .5$—poor, $\alpha < .5$—unacceptable.

**alpha if item deleted** In reliability analysis, the resulting alpha if the variable to the left is deleted.

**analysis of variance (ANOVA)** A statistical test that identifies whether there are any significant differences between three or more sample means. See Chapters 12–14 for a more complete description.

**asymptotic values** Determination of parameter estimates based on asymptotic values (the value a function is never expected to exceed). This process is used in nonlinear regression and other procedures where an actual value is not possible to calculate.

**B** In regression output, the *B* values are the regression coefficients and the constant for the regression equation. The *B* may be thought of as a weighted constant that describes the magnitude of influence a particular independent variable has on the dependent variable. A positive value for *B* indicates a corresponding increase in the value of the dependent variable, whereas a negative value for *B* decreases the value of the dependent variable.

**bar graph** A graphical representation of the frequency of categorical data. A similar display for continuous data is called a histogram.

**Bartlett test of sphericity** This is a measure of the multivariate normality of a set of distributions. A significance value < .05 suggests that the data do not differ significantly from multivariate normal. See page 264 for more detail.

**beta ($\beta$)** In regression procedures, the standardized regression coefficients. This is the *B* value for standardized scores (*z* scores) of the variables. These values will vary strictly between ±1.0 and may be compared directly with beta values in other analyses.

**beta in** In multiple regression analysis, the beta values for the excluded variables if these variables were actually in the regression equation.

**between-groups sum of squares** The sum of squared deviations between the grand mean and each group mean weighted (multiplied) by the number of subjects in each group.

**binomial test** A nonparametric test that measures whether a distribution of values is binomially distributed (each outcome equally likely). For instance, if you tossed a coin 100 times, you would expect a binomial distribution (approximately 50 heads and 50 tails).

**Bonferroni test** A post hoc test that adjusts for experimentwise error by dividing the alpha value by the total number of tests performed.

**bootstrap** In many of the statistical-procedure screens a button with the word "Bootstrap…" occurs. The bootstrap procedure is not central to any of the functions we describe in the book and is thus not addressed there. In statistics, bootstrapping is a computer-based method for assigning measures of accuracy to sample estimates. The SPSS bootstrap procedure, by default, takes 1,000 random samples from your data set to generate accurate parameter estimates. Description of standard error (page 115) adds detail.

**Box's *m*** A measure of multivariate normality based on the similarities of determinants of the covariance matrices for two or more groups.

**canonical correlation** In discriminant analysis, the canonical correlation is a correlation between the discriminant scores for each subject and the levels of the dependent variable for each subject. See Chapter 26 to see how a canonical correlation is calculated.

**canonical discriminant functions** The linear discriminant equation(s) calculated to maximally discriminate between levels of the dependent (or criterion) variable. This is described in detail in Chapter 26.

**chi-square analysis** A nonparametric test that makes comparisons (usually of crosstabulated data) between two or more samples on the observed frequency of values with the expected frequency of values. Also used as a test of the goodness-of-fit of log-linear and structural models. For the latter, the question being asked is: Does the actual data differ significantly from results predicted from the model that has been created? The formula for the Pearson chi-square is:

$$\chi^2 = \Sigma[(f_0 - f_e)^2/f_e]$$

**cluster analysis** A procedure by which subjects, cases, or variables are clustered into groups based on similar characteristics of each.

**Cochran's *c* and Bartlett-box *f*** Measure whether the variances of two or more groups differ significantly from each other (heteroschedasticity). A high probability value (for example, $p > .05$) indicates that the variances of the groups do not differ significantly.

**column percent** A term used with crosstabulated data. It is the result of dividing the frequency of values in a particular cell by the frequency of values in the entire column. Column percents sum to 100% in each column.

**column total** A term used with crosstabulated data. It is the total number of subjects in each column.

**communality** Used in factor analysis, a measure designed to show the proportion of variance that factors contribute to explaining a particular variable. In the SPSS default procedure, communalities are initially assigned a value of 1.00.

**confidence interval** The range of values within which a particular statistic is likely to fall. For instance, a 95% confidence interval for the mean indicates that there is a 95% chance that the true population mean falls within the range of values listed.

**converge** To converge means that after some number of iterations, the value of a particular statistic does not change more than a pre-specified amount and parameter estimates are said to have "converged" to a final estimate.

**corrected item-total correlation** In reliability analysis, correlation of the designated variable with the sum of all other variables in the analysis.

**correlation** A measure of the strength and direction of association between two variables. See Chapter 10 for a more complete description.

**correlation between forms** In split-half reliability analysis, an estimate of the reliability of the measure if each half had an equal number of items.

**correlation coefficient** A value that measures the strength of association between two variables. This value varies between ±1.0 and is usually designated by the lowercase letter *r*.

**count** In crosstabulated data, the top number in each of the cells indicating the actual number of subjects or cases in each category.

**covariate** A variable that has substantial correlation with the dependent variable and is included in an experiment as an adjustment of the results for differences existing among subjects prior to the experiment.

**Cramer's *V*** A measure of the strength of association between two categorical variables. Cramer's *V* produces a value between 0 and 1 and (except for the absence of a negative relation) may be interpreted in a manner similar to a correlation. Often used within the context of chi-square analyses. The equation follows. (Note: *k* is the smaller of the number of rows and columns.)

$$V = \sqrt{\chi^2/[N(k - 1)]}$$

**crosstabulation** Usually a table of frequencies of two or more categorical variables taken together. However, crosstabulation may also be used for different ranges of values for continuous data. See Chapter 8.

**cumulative frequency** The total number of subjects or cases having a given score or any score lower than the given score.

**cumulative percent** The total percent of subjects or cases having a given score or any score lower than the given score.

***d* or Cohen's *d*** The measure of effect size for a *t* test. It measures the number of standard deviations (or fraction of a standard deviation) by which the two mean values in a *t* test differ.

**degrees of freedom (DF)** The number of values that are free to vary, given one or more statistical restrictions on the entire set of values. Also, a statistical compensation for the failure of a range of values to be normally distributed.

**dendogram** A branching-type graph used to demonstrate the clustering procedure in cluster analysis. See Chapter 21 for an example.

**determinant of the variance-covariance matrices:** The determinant provides an indication of how strong a relationship there is among the variables in a correlation matrix. The smaller the number, the more closely the variables are related to each other. This is used primarily by the computer to compute the Box's *M* test. The determinant of the pooled variance-covariance matrix refers to all the variance-covariance matrices present in the analysis.

**deviation** The distance and direction (positive or negative) of any raw score from the mean.

**(difference) mean** In a *t* test, the difference between the two means.

**discriminant analysis** A procedure that creates a regression formula to maximally discriminate between levels of a categorical dependent variable. See Chapter 22.

**discriminant scores** Scores for each subject, based on substitution of values for the corresponding variables into the discriminant formula.

**DTS** Decline to state.

**eigenvalue** In factor analysis, the proportion of variance explained by each factor. In discriminant analysis, the between-groups sums of squares divided by within-groups sums of squares. A large eigenvalue is associated with a strong discriminant function.

**equal-length Spearman-Brown** Used in split-half reliability analysis when there is an unequal number of items in each portion of the analysis. It produces a correlation value that is inflated to reflect what the correlation would be if each part had an equal number of items.

**eta** A measure of correlation between two variables when one of the variables is discrete.

**eta squared** The proportion of the variance in the dependent variable accounted for by an independent variable. For instance, an eta squared of .044 would indicate that 4.4% of the variance in the dependent variable is due to the influence of the independent variable.

**exp(B):** In logistic regression analysis, $e^B$ is used to help in interpreting the meaning of the regression coefficients. (Remember that the regression equation may be interpreted in terms of *B* or $e^B$.)

**expected value** In the crosstabulation table of a chi-square analysis, the number that would appear if the two variables were perfectly independent of each other. In regression analysis, it is the same as a predicted value, that is, the value obtained by substituting data from a particular subject into the regression equation.

**factor** In factor analysis, a factor (also called a component) is a combination of variables whose shared correlations explain a certain amount of the total variance. After rotation, factors are designed to demonstrate underlying similarities between groups of variables.

**factor analysis** A statistical procedure designed to take a larger number of constructs (measures of some sort) and reduce them to a smaller number of factors that describe these measures with greater parsimony. See Chapter 20.

**factor transformation matrix** If the original unrotated factor matrix is multiplied by the factor transformation matrix, the result will be the rotated factor matrix.

***F*-change** In multiple regression analysis, the *F*-change value is associated with the additional variance explained by a new variable.

***F*-ratio** In an analysis of variance, an *F* ratio is the between-groups mean square divided by the within-groups mean square. This value is designed to compare the between-groups variation to the within-groups variation. If the between-groups variation is substantially larger than the within-groups variation, then significant differences between groups will be demonstrated. In multiple regression analysis, the *F* ratio is the mean square (regression) divided by the mean square (residual). It is designed to demonstrate the strength of association between variables.

**frequencies** A listing of the number of times certain events take place.

**Fridman 2-way ANOVA** A nonparametric procedure that tests whether three or more groups differ significantly from each other, based on average rank of groups rather than comparison of means from normally distributed data.

**goodness-of-fit test statistics** The likelihood-ratio chi-square and the Pearson chi-square statistics examine the fit of log-linear models. Large chi-square values and small *p* values indicate that the model does not fit the data well. Be aware that this is the opposite of thinking in interpretation of most types of analyses. Usually one looks for a large test statistic and a small *p* value to indicate a significant effect. In this case, a large chi-square and a small *p* value would indicate that your data differs significantly from (or does not fit well) the model you have created.

**group centroids** In discriminant analysis, the average discriminant score for subjects in the two (or more) groups. More specifically, the discriminant score for each group is determined when the variable means (rather than individual values for each subject) are entered into the discriminant equation. If you are discriminating between exactly two outcomes, the two scores will be equal in absolute value but have opposite signs. The dividing line between group membership in that case will be zero (0).

**Guttman split-half** In split-half reliability, a measure of the reliability of the overall test, based on a lower-bounds procedure.

**hypothesis SS** The between-groups sum of squares; the sum of squared deviations between the grand mean and each group mean, weighted (multiplied) by the number of subjects in each group.

**icicle plot** A graphical display of the step-by-step clustering procedure in cluster analysis. See Chapter 21 for an example.

**interaction** The idiosyncratic effect of two or more independent variables on a dependent variable over and above the independent variables' separate (main) effects.

**intercept** In regression analysis, the point where the regression line crosses the Y-axis. The intercept is the predicted value of the vertical-axis variable when the horizontal-axis variable value is zero.

**inter-item correlations** In reliability analysis, this is descriptive information about the correlation of each variable with the sum of all the others.

**item means** In reliability analysis (using coefficient alpha), this is descriptive information about all subjects' means for all the variables. On page 239, an example clarifies this.

**item variances** A construct similar to that used in item means (the previous entry). The first number in the SPSS output is the mean of all the variances, the second is the lowest of all the variances, and so forth. Page 239 clarifies this with an example.

**iteration** The process of solving an equation based on preselected values, and then replacing the original values with the computer-generated values and solving the equation again. This process continues until some criterion (in terms of amount of change from one iteration to the next) is achieved.

*K* In hierarchical log-linear models, the order of effects for each row of the table (1 = first-order effects, 2 = second-order effects, and so on).

**Kaiser-Mayer-Olkin** A measure of whether the distribution of values is adequate for conducting factor analysis. A measure > .9 is generally thought of as excellent, > .8 as good, > .7 as acceptable, > .6 as marginal, > .5 as poor, and < .5 as unacceptable.

**Kolmogorov-Smirnov one-sample test** A nonparametric test that determines whether the distribution of the members of a single group differ significantly from a normal (or uniform, Poisson, or exponential) distribution.

***k*-sample median test:** A nonparametric test that determines whether two or more groups differ on the number of instances (within each group) greater than the grand median value or less than the grand median value.

**kurtosis** A measure of deviation from normality that measures the peakedness or flatness of a distribution of values. See Chapter 7 for a more complete description.

**Leveine's test** A test that examines the assumption that the variance of each dependent variable is the same as the variance of all other dependent variables.

**log determinant** In discriminant analysis, the natural log of the determinant of each of the two (or more) covariance matrices. This is used to test the equality of group covariance matrices using Box's *M*.

**linear dependency** Linear dependency essentially describes when two or more variables are highly correlated with each other. There are two kinds numeric linear dependency when one variable is a composite of another variable (such as final score and total points) and conceptual linear dependency is when two concepts are highly similar (such as tension and anxiety) and are highly correlated.

**LSD** "Least Significant Difference" post hoc test; performs a series of *t* tests on all possible combinations of the independent variable on each dependent variable. A liberal test.

**main effects** The influence of a single independent variable on a dependent variable. See Chapters 13 and 14 for examples.

**MANCOVA** A MANOVA that includes one or more covariates in the analysis.

**Mann-Whitney and Wilcoxon rank-sum test** A nonparametric alternative to the *t* test that measures whether two groups differ from each other based on ranked scores.

**MANOVA** Multivariate analysis of variance. A complex procedure similar to ANOVA except that it allows for more than one dependent variable in the analysis.

**MANOVA repeated measures** A multivariate analysis of variance in which the same set of subjects experiences several measurements on the variables of interest over time. Computationally, it is the same as a within-subjects MANOVA.

**MANOVA within-subjects** A multivariate analysis of variance in which the same set of subjects experience all levels of the dependent variable.

**Mantel-Haenszel test for linear association** Within a chi-square analysis, this procedure tests whether the two variables correlate with each other. This measure is often meaningless unless there is some logical or numeric relation to the order of the levels of the variables.

**Mauchly's sphericity test** A test of multivariate normality. SPSS computes a $\chi^2$ approximation for this test, along with its significance level. If the significance level associated with Mauchly's sphericity test is small (i.e., $p < .05$), then the data may not be spherically distributed.

**maximum** Largest observed value for a distribution.

**mean** A measure of central tendency; the sum of a set of scores divided by the total number of scores in the set.

**mean square** Sum of squares divided by the degrees of freedom. In ANOVA, the most frequently observed mean squares are the within-groups sum of squares divided by the corresponding degrees of freedom and the between-groups sum of squares divided by the associated degrees of freedom. In regression analysis, it is the regression sum of squares and the residual sum of squares divided by the corresponding degrees of freedom. For both ANOVA and regression, these numbers are used to determine the *F* ratio.

**median** A measure of central tendency; the middle point in a distribution of values.

**median test** See *K*-Sample Median Test.

**minimum** Lowest observed value for a distribution.

**minimum expected frequency** A chi-square analysis identifies the value of the cell with the minimum expected frequency.

**−2 log likelihood** This is used to indicate how well a log-linear model fits the data. Smaller −2 log likelihood values mean that the model fits the data better; a perfect model has a −2 log likelihood value of zero. Significant $\chi^2$ values indicate that the model differs significantly from the theoretically "perfect" model.

**mode** A measure of central tendency; it is the most frequently occurring value.

**model $\chi^2$** In logistic regression analysis, this value tests whether or not all the variables entered in the equation have a significant effect on the dependent variable. A high $\chi^2$ value indicates that the variables in the equation significantly impact the dependent variable. This test is functionally equivalent to the overall *F* test in multiple regression.

**multiple regression analysis** A statistical technique designed to predict values of a dependent (or criterion) variable from knowledge of the values of two or more independent (or predictor) variables. See Chapter 16 for a more complete description.

**multivariate test for homogeneity of dispersion matrices** Box's *M* test examines whether the variance-covariance matrices are the same in all cells. To evaluate this test, SPSS calculates an *F* or $\chi^2$ approximation for the *M*. These values, along with their associated *p* values, appear in the SPSS output. Significant *p* values indicate differences between the variance-covariance matrices for the two groups.

**multivariate tests of significance** In MANOVA, there are several methods of testing for differences between the dependent variables due to the independent variables. Pillai's method is considered the best test by many, in terms of statistical power and robustness.

**95% confidence interval** See confidence interval.

**nonlinear regression** A procedure that estimates parameter values for intrinsically nonlinear equations.

**nonparametric tests** A series of tests that make no assumptions about the distribution of values (usually meaning the distribution is not normally distributed) and performs statistical analyses based upon rank order of values, comparisons of paired values, or other techniques that do not require normally distributed data.

**normal distribution** A distribution of values that, when graphed, produces a smooth, symmetrical, bell-shaped distribution that has skewness and kurtosis values equal to zero.

**oblique rotations** A procedure of factor analysis in which rotations are allowed to deviate from orthogonal (or from perpendicular) in an effort to achieve a better simple structure.

**observed value or count** In a chi-square analysis, the frequency results that are actually obtained when conducting an analysis.

**one-sample chi-square test** A nonparametric test that measures whether observed scores differ significantly from expected scores for levels of a single variable.

**one-tailed test** A test in which significance of the result is based on deviation from the null hypothesis in only one direction.

**overlay plot** A type of scatterplot that graphs two or more variables along the horizontal axis against a single variable on the vertical axis.

**parameter** A numerical quantity that summarizes some characteristic of a population.

**parametric test** A statistical test that requires that the characteristics of the data being studied be normally distributed in the population.

**partial** A term frequently used in multiple regression analysis. A partial effect is the unique contribution of a new variable after variation from other variables has already been accounted for.

**partial chi-square** The chi-square value associated with the unique additional contribution of a new variable on the dependent variable.

**P(D/G)** In discriminant analysis, given the discriminant value for that case (D), what is the probability of belonging to that group (G)?

**Pearson product-moment correlation** A measure of correlation ideally suited for determining the relationship between two continuous variables.

**percentile** A single number that indicates the percent of cases in a distribution falling below that single value. See Chapter 6 for an example.

**P(G/D)** In discriminant analysis, given that this case belongs to a given group (G), how likely is the observed discriminant score (D)?

**phi coefficient** A measure of the strength of association between two categorical variables, usually in a chi-square analysis. Phi is computed from the formula:

$$\phi = \sqrt{\chi^2/N}$$

**pooled within-group correlations** Pooled within group differs from values for the entire (total) group in that the pooled values are the average (mean) of the group correlations. If the *N*s are equal, then this would be the same as the value for the entire group.

**pooled within-groups covariance matrix** In discriminant analysis, a matrix composed of the means of each corresponding value within the two (or more) matrices for each level of the dependent variable.

**population** A set of individuals or cases who share some characteristic of interest. Statistical inference is based on drawing samples from populations to gain a fuller understanding of characteristics of that population.

**power** Statistical power refers to the ability of a statistical test to produce a significant result. Power varies as a function of the type of test (parametric tests are usually more powerful than nonparametric tests) and the size of the sample (greater statistical power is usually observed with large samples than with small samples).

**principal-components analysis** The default method of factor extraction used by SPSS.

**prior probability for each group** The .5000 value usually observed indicates that groups are weighted equally.

**probability** Also called significance. A measure of the rarity of a particular statistical outcome given that there is actually no effect. A significance of $p < .05$ is the most widely accepted value by which researchers accept a certain result as statistically significant. It means that there is less than a pearson chance that the given outcome could have occurred by chance.

**quartiles** Percentile ranks that divide a distribution into the 25th, 50th, and 75th percentiles.

***R*** The multiple correlation between a dependent variable and two or more independent (or predictor) variables. It varies between 0 and 1.0 and is interpreted in a manner similar to a bivariate correlation.

**$R^2$** Also called the multiple coefficient of determination. The proportion of variance in the dependent (or criterion) variable that is explained by the combined influence of two or more independent (or predictor) variables.

**$R^2$ change** This represents the unique contribution of a new variable added to the regression equation. It is calculated by simply subtracting the $R^2$ value for the given line from the $R^2$ value of the previous line.

**range** A measure of variability; the difference between the largest and smallest scores in a distribution.

**rank** Rank or size of a covariance matrix.

**regression** In multiple regression analysis, this term is often used to indicate the amount of explained variation and is contrasted with residual, which is unexplained variation.

**regression analysis** A statistical technique designed to predict values of a dependent (or criterion) variable from knowledge of the values of one or more independent (or predictor) variable(s). See Chapters 15 and 16 for greater detail.

**regression coefficients** The *B* values. These are the coefficients of the variables within the regression equation plus the constant.

**regression line** Also called the line of best fit. A straight line drawn through a scatterplot that represents the best possible fit for making predictions from one variable to the other.

**regression plot** A scatterplot that includes the intercepts for the regression line in the vertical axes.

**residual** Statistics relating to the unexplained portion of the variance. See Chapter 28.

**residuals and standardized residuals** In log-linear models, the residuals are the observed counts minus the expected counts. High residuals indicate that the model is not adequate. SPSS calculates the adjusted residuals using an estimate of the standard error. The distribution of adjusted residuals is a standard normal distribution, and numbers greater than 1.96 or less than −1.96 are not likely to have occurred by chance ($\alpha = .05$).

**rotation** A procedure used in factor analysis in which axes are rotated in order to yield a better simple structure and a more interpretable pattern of values.

**row percent** A term used with crosstabulated data. It is the result of dividing the frequency of values in a particular cell by the frequency of values in the entire row. Row percents sum to 100% in each row.

**row total** The total number of subjects in a particular row.

**runs test** A nonparametric test that determines whether the elements of a single dichotomous group differ from a random distribution.

**sample** A set of individuals or cases taken from some population for the purpose of making inferences about characteristics of the population.

**sampling error** The anticipated difference between a random sample and the population from which it is drawn based on chance alone.

**scale mean if item deleted** In reliability analysis, for each subject all the variables (excluding the variable to the left) are summed. The values shown are the means for all variables across all subjects.

**scale variance if item deleted** In reliability analysis, the variance of summed variables when the variable to the left is deleted.

**scatterplot** A plot showing the relationship between two variables by marking all possible pairs of values on a bi-coordinate plane. See Chapter 10 for greater detail.

**Scheffé procedure** The Scheffé test allows the researcher to make pairwise comparisons of means after a significant *F* value has been observed in an ANOVA.

**scree plot** A plot of the eigenvalues in a factor analysis that is often used to determine how many factors to retain for rotation.

**significance** Frequently called probability. A measure of the rarity of a particular statistical outcome given that there is actually no effect. A significance of $p < .05$ is the most widely accepted value by which researchers accept a certain result as statistically significant. It means that there is less than a 5% chance that the given outcome could have occurred by chance.

**sign test** A nonparametric test that determines whether two distributions differ based on a comparison of paired scores.

**singular variance-covariance matrices** Cells with only one observation or with singular variance-covariance matrices indicate that there may not be enough data to accurately compute MANOVA statistics or that there may be other problems present in the data, such as linear dependencies (where one variable is dependent on one or more of the other variables). Results from any analysis with only one observation or with singular variance-covariance matrices for some cells should be interpreted with caution.

**size** The number associated with an article of clothing in which the magnitude of the number typically correlates positively with the size of the associated body part.

**skewness** In a distribution of values, this is a measure of deviation from symmetry. Negative skewness describes a distribution with a greater number of values above the mean; positive skewness describes a distribution with a greater number of values below the mean. See Chapter 7 for a more complete description.

**slope** The angle of a line in a bi-coordinate plane based on the amount of change in the Y variable per unit change in the X variable. This is a term most frequently used in regression analysis and can be thought of as a weighted constant indicating the influence of the independent variable(s) on a designated dependent variable.

**split-half reliability** A measure of reliability in which an instrument is divided into two equivalent sections (or different forms of the same test or the same test given at different times) and then intercorrelations between these two halves are calculated as a measure of internal consistency.

**squared euclidean distance** The most common method (and the SPSS default) used in cluster analysis to determine how cases or clusters differ from each other. It is the sum of squared differences between values on corresponding variables.

**squared multiple correlation** In reliability analysis, these values are determined by creating a multiple regression equation to generate the predicted correlation based on the correlations for all other variables.

**sscon = 1.000E−08** This is the default value at which iteration ceases in nonlinear regression. 1.000E−08 is the computer's version of $1.000 \times 10^{-8}$, scientific notation for .00000001 (one hundred-millionth). This criterion utilizes the residual sum of squares as the value to determine when iteration ceases.

**standard deviation** The standard measure of variability around the mean of a distribution. The standard deviation is the square root of the variance (the sum of squared deviations from the mean divided by $N - 1$).

**standard error** This term is most frequently applied to the mean of a distribution but may apply to other measures as well. It is the standard deviation of the statistic-of-interest given a large number of samples drawn from the same population. It is typically used as a measure of the stability or of the sampling error of the distribution and is based on the standard deviation of a single random sample.

**standardized item alpha** In reliability analysis, this is the alpha produced if the included items are changed to *z* scores before computing the alpha.

**statistics for summed variables** In reliability analysis, there are always a number of variables being considered. This line lists descriptive information about the sum of all variables for the entire sample of subjects.

**stepwise variable selection:** This procedure enters variables into the discriminant equation, one at a time, based on a designated criterion for inclusion ($F \geq 1.00$ is default) but will drop variables from the equation if the inclusion requirement drops below the designated level when other variables have been entered.

**string variable** A type of data that typically consists of letters or letters and numbers but cannot be used for analysis except some types of categorization.

**sum of squares** A standard measure of variability. It is the sum of the square of each value subtracted from the mean.

***t* test** A procedure used for comparing exactly two sample means to see if there is sufficient evidence to infer that the means of the corresponding population distributions also differ.

***t* test—independent samples** A *t* test that compares the means of two distributions of some variable in which there is no overlap of membership of the two groups being measured.

***t* test—one sample** A *t* test in which the mean of a distribution of values is compared to a single fixed value.

***t* test—paired samples** A *t* test in which the same subjects experience both levels of the variable of interest.

***t* tests in regression analysis** A test to determine the likelihood that a particular correlation is statistically significant. In the regression output, it is *B* divided by the standard error of *B*.

**tolerance level** The tolerance level is a measure of linear dependency between one variable and the others. In discriminant analysis, if a tolerance is less than .001, this indicates a high level of linear dependency, and SPSS will not enter that variable into the equation.

**total sum of squares** The sum of squared deviations of every raw score from the overall mean of the distribution.

**Tukey's HSD** (Honestly Significant Difference). A value that allows the researcher to make pairwise comparisons of means after a significant *F* value has been observed in an ANOVA.

**two-tailed test** A test in which significance of the result is based on deviation from the null hypothesis in either direction (larger or smaller).

**unequal-length spearman-brown** In split-half reliability, the reliability calculated when the two "halves" are not equal in size.

**univariate *f*-tests** An *F* ratio showing the influence of exactly one independent variable on a dependent variable.

**unstandardized canonical discriminant function coefficients** This is the list of coefficients (and the constant) of the discriminant equation.

**valid percent** Percent of each value excluding missing values.

**value** The number associated with each level of a variable (e.g., male = 1, female = 2).

**value label** Names or number codes for levels of different variables.

**variability** The way in which scores are scattered around the center of a distribution. Also known as variance, dispersion, or spread.

**variable labels** These are labels entered when formatting the raw data file. They allow up to 40 characters for a more complete description of the variable than is possible in the 8-character name.

**variables in the equation** In regression analysis, after each step in building the equation, SPSS displays a summary of the effects of the variables that are currently in the regression equation.

**variance** A measure of variability about the mean, the square of the standard deviation, used largely for computational purposes. The variance is the sum of squared deviations divided by $N - 1$.

**Wald** In log-linear models, a measure of the significance of *B* for the given variable. Higher values, in combination with the degrees of freedom, indicate significance.

**Wilcoxon matched-pairs signed-ranks test** A nonparametric test that is similar to the sign test except the positive and negative signs are weighted by the mean rank of positive versus negative comparisons.

**Wilks' lambda** The ratio of the within-groups sum of squares to the total sum of squares. This is the proportion of the total variance in the discriminant scores not explained by differences among groups. A lambda of 1.00 occurs when observed group means are equal (all the variance is explained by factors other than difference between these means), whereas a small lambda occurs when within-groups variability is small compared to the total variability. A small lambda indicates that group means appear to differ. The associated significance values indicate whether the difference is significant.

**within-groups sum of squares** The sum of squared deviations between the mean for each group and the observed values of each subject within that group.

***z* score** Also called *standard score*. A distribution of values that standardizes raw data to a mean of zero (0) and a standard deviation of one (1.0). A *z* score is able to indicate the direction and degree that any raw score deviates from the mean of a distribution. ***Z*** scores are also used to indicate the significant deviation from the mean of a distribution. A *z* score with a magnitude greater than ±1.96 indicates a significant difference at $p < .05$ level.

# References

**Three SPSS manuals (the three books authored by Marija Norusis) and one SPSS syntax guide cover (in great detail) all procedures that are included in the present book: Note: The "19" reflects that SPSS does not seem to have more recent manuals. These books are still available, but the shift seems to be to have most information online.**

Norušis, Marija. (2011). *IBM SPSS Statistics 19 Statistical Procedures Companion*. Upper Saddle River, NJ: Prentice Hall.

Norušis, Marija. (2011). *IBM SPSS Statistics 19 Guide to Data Analysis*. Upper Saddle River, NJ: Prentice Hall.

Norušis, Marija. (2011). *IBM SPSS Statistics 19 Advanced Statistical Procedures Companion*. Upper Saddle River, NJ: Prentice Hall.

Collier, Jacqueline. (2009). *Using SPSS Syntax: A Beginner's Guide*. Thousand Oaks, CA: Sage Publications.

(*Note*: It seems that SPSS no longer publishes a syntax guide, so we insert Collier's work.)

**Good introductory statistics texts that cover material through Chapter 13 (one-way ANOVA) and Chapter 18 (reliability):**

Fox, James; Levin, Jack; & Harkins, Stephen. (1994). *Elementary Statistics in Behavioral Research*. New York: Harper Collins College Publishers.

Hopkins, Kenneth; Glass, Gene; & Hopkins, B. R. (1995). *Basic Statistics for the Behavioral Sciences*. Boston: Allyn and Bacon.

Moore, David; McCabe, George; & Craig, Bruce A. (2010). *Introduction to the Practice of Statistics, Third Edition*. New York: W.H. Freeman and Company.

Welkowitz, Joan; Ewen, Robert; & Cohen, Jacob. (2006). *Introductory Statistics for the Behavioral Sciences, Sixth Edition*. New York: John Wiley & Sons.

Witte, Robert S. (2009). *Statistics, Eighth Edition*. New York: John Wiley & Sons.

**Comprehensive coverage of Analysis of Variance:**

Keppel, Geoffrey; & Wickens, Thomas. (2004). *Design and Analysis: A Researcher's Handbook, Fourth Edition*. Englewood Cliffs, NJ: Prentice Hall.

Lindman, Harold R. (1992). *Analysis of Variance in Experimental Design*. New York: Springer-Verlag.

Turner, J. Rick; & Thayer, Julian F. (2001). *Introduction to Analysis of Variance: Design, Analysis & Interpretation*. Thousand Oaks, CA: Sage Publications.

**Comprehensive coverage of MANOVA and MANCOVA:**

Lindman, Harold R. (1992). *Analysis of Variance in Experimental Design*. New York: Springer-Verlag.

Turner, J. Rick; & Thayer, Julian F. (2001). *Introduction to Analysis of Variance: Design, Analysis & Interpretation*. Thousand Oaks, CA: Sage Publications.

**Comprehensive coverage of simple and multiple regression analysis:**

Chatterjee, Samprit; & Hadi, Ali S. (2006). *Regression Analysis by Example, Third Edition*. New York: John Wiley & Sons.

Gonick, Larry; & Smith, Woollcott. (1993). *The Cartoon Guide to Statistics*. New York: Harper Perennial.

Pedhazur, Elazar J. (1997). *Multiple Regression in Behavioral Research*. New York: Holt, Rinehart and Winston.

West, Stephen G.; & Aiken, Leona S. (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. London: Routledge Academic.

Sen, Ashish; & Srivastava, Muni. (1997). *Regression Analysis: Theory, Methods, and Applications*. New York: Springer-Verlag.

Weisberg, Sanford. (2005). *Applied Linear Regression, Third Edition*. New York: John Wiley & Sons.

**Comprehensive coverage of factor analysis:**

Comrey, Andrew L.; & Lee, Howard B. (1992). *A First Course in Factor Analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Brown, Timothy A. (2006). *Confirmatory Factor Analysis for Applied Research (Methodology In The Social Sciences)*. New York: Guilford Press.

**Comprehensive coverage of cluster analysis:**

Everitt, Brian S.; Landau, Sabine; & Leese, Morven. (2011). *Cluster Analysis, Fourth Edition*. London: Hodder/Arnold.

**Comprehensive coverage of discriminant analysis:**

McLachlan, Geoffrey J. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. New York: John Wiley & Sons.

**Comprehensive coverage of nonlinear regression:**

Seber, G .A. F.; & Wild, C. J. (2003). *Nonlinear Regression*. New York: John Wiley & Sons.

**Comprehensive coverage of logistic regression analysis and loglinear models:**

Agresti, Alan. (2007). *An Introduction to Categorical Data Analysis*. New York: John Wiley & Sons.

McLachlan, Geoffrey J. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. New York: John Wiley & Sons.

Wickens, Thomas D. (1989). *Multiway Contingency Tables Analysis for the Social Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

**Comprehensive coverage of nonparametric tests:**

Bagdonavius, Vilijandas; Kruopis, Julius; & Mikulin, Mikhail. (2010). *Nonparametric Tests for Complete Data*. New York: John Wiley & Sons.

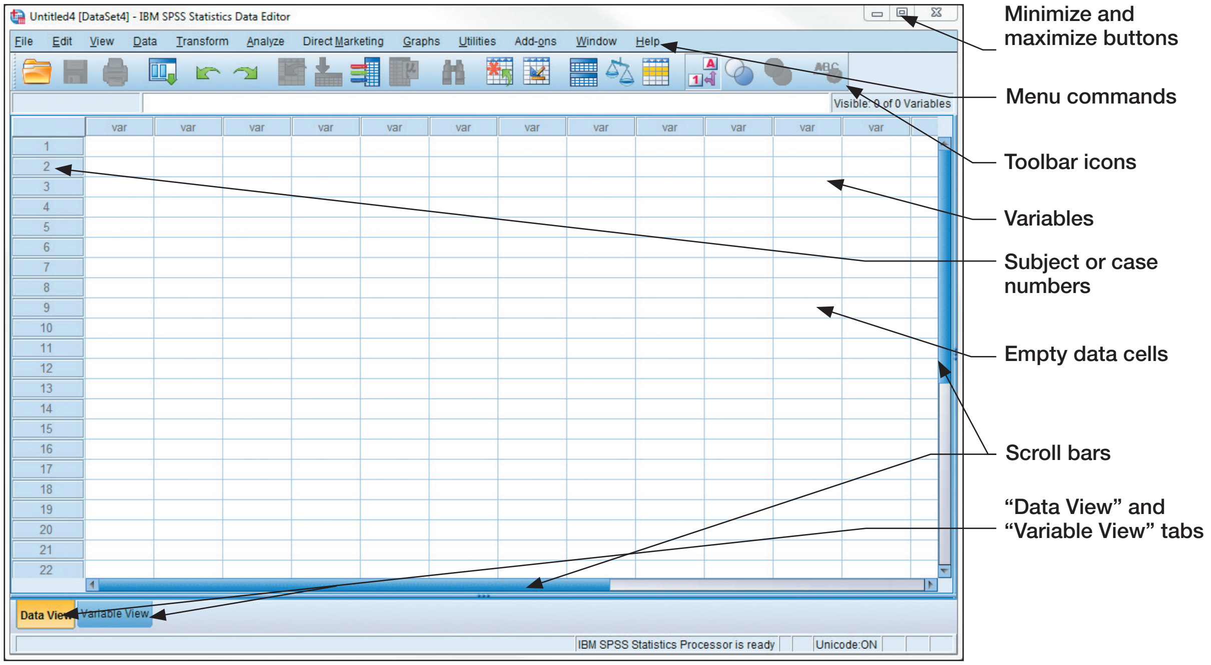
**Comprehensive coverage of multidimensional scaling:**

Davison, M. L. (1992). *Multidimensional Scaling*. New York: Krieger Publishing Company.

Young, F. W.; & Hamer, R. M. (1987). *Multidimensional Scaling: History, Theory, and Applications*. Hillsdale, NJ: Lawrence Erlbaum Associates.

# Index

## Front 1
Initial data screen


Untitled4 [DataSet4] - IBM SPSS Statistics Data Editor

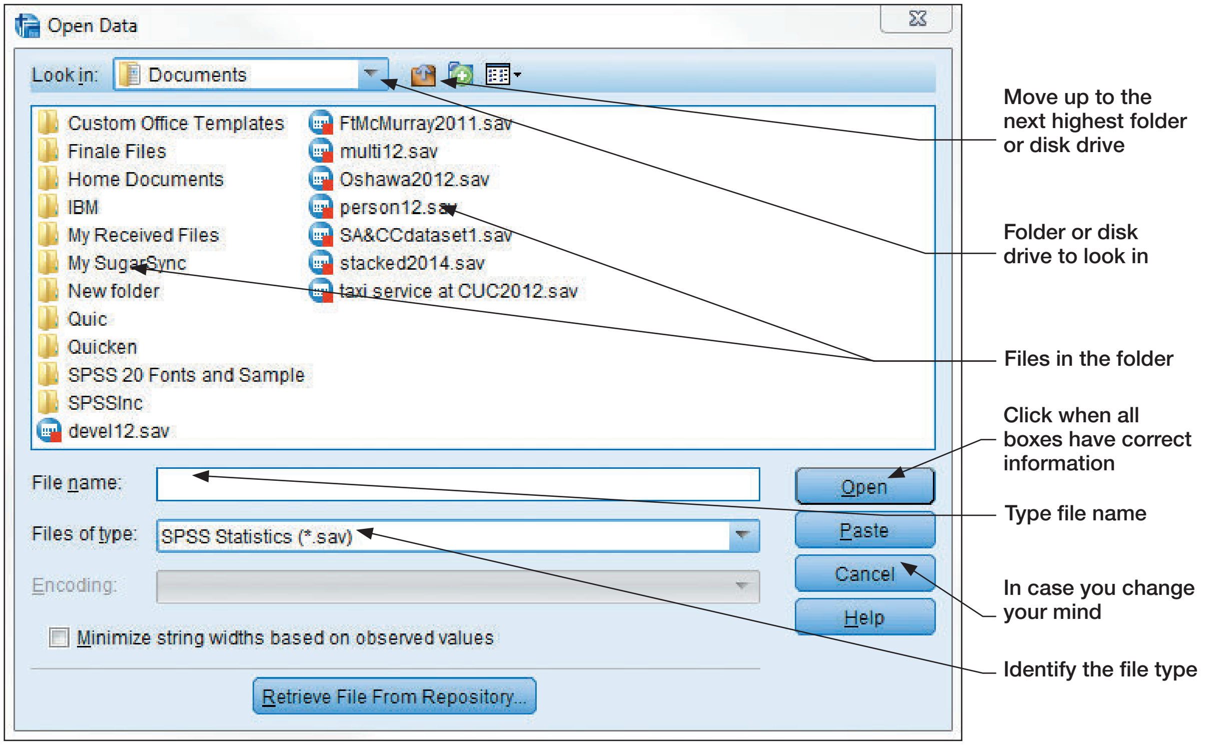
Visible: 0 of 0 Variables

IBM SPSS Statistics Processor is ready

Unicode:ON

| Icon | Function | Icon | Function | Icon | Function |
|---|---|---|---|---|---|
| | Click this to open a file | | Find data | | (upper left corner) the "+" sign indicates that this is the active file |
| | Save current file | | Insert subject or case into the data file | A 1 | Shifts between numbers and labels for variables with several levels |
| | Print file | | Insert new variable into the data file | | Go to a particular variable or case number |
| | Recall a recently-used command | | Split file into subgroups | | Use subsets of variables/use all variables |
| | Undo the last operation | | Weight cases | | Access information about the current variable |
| | Redo something you just undid | | Select cases | ABC | Spell check |

## Front 2
Open Data Screen


Open Data

Look in: Documents

Custom Office Templates
Finale Files
Home Documents
IBM
My Received Files
My SugarSync
New folder
Quic
Quicken
SPSS 20 Fonts and Sample
SPSSInc
devel12.sav
FtMcMurray2011.sav
multi12.sav
Oshawa2012.sav
person12.sav
SA&CCdataset1.sav
stacked2014.sav
taxi service at CUC2012.sav

File name:

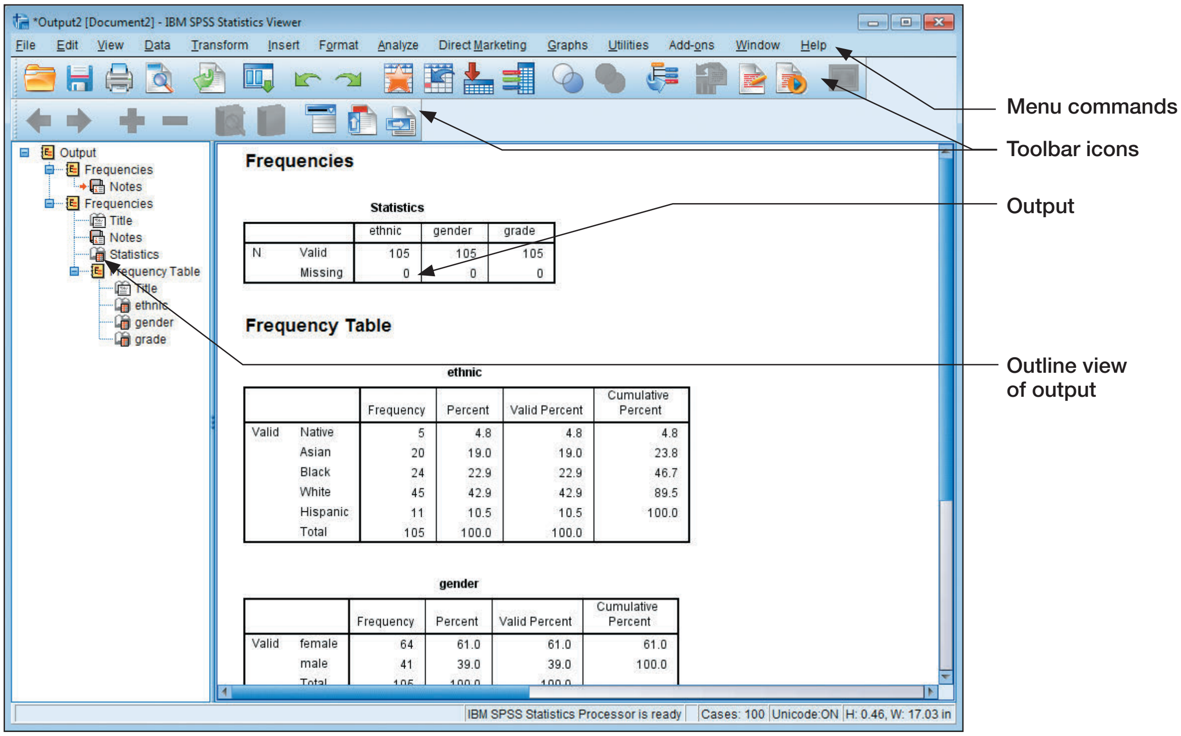
Files of type: SPSS Statistics (*.sav)

Encoding:

☐ Minimize string widths based on observed values

Open
Paste
Cancel
Help
Retrieve File From Repository...
Move up to the next highest folder or disk drive
Folder or disk drive to look in
Files in the folder
Click when all boxes have correct information
Type file name
In case you change your mind
Identify the file type

## Back 1
The SPSS Output navigator window



**Frequencies**

**Statistics**

| | | ethnic | gender | grade |
|---|---|---|---|---|
| N | Valid | 105 | 105 | 105 |
| | Missing | 0 | 0 | 0 |

**Frequency Table**

**ethnic**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Native | 5 | 4.8 | 4.8 | 4.8 |
| | Asian | 20 | 19.0 | 19.0 | 23.8 |
| | Black | 24 | 22.9 | 22.9 | 46.7 |
| | White | 45 | 42.9 | 42.9 | 89.5 |
| | Hispanic | 11 | 10.5 | 10.5 | 100.0 |
| | Total | 105 | 100.0 | 100.0 | |

**gender**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | female | 64 | 61.0 | 61.0 | 61.0 |
| | male | 41 | 39.0 | 39.0 | 100.0 |
| | Total | 105 | 100.0 | 100.0 | |

IBM SPSS Statistics Processor is ready | Cases: 100 | Unicode:ON | H: 0.46, W: 17.03 in

| Icon | Function | Icon | Function |
|---|---|---|---|
| 📂 | Click to open a file | ● | Show all variables |
| 💾 | Click to save a file | | Select most recent output |
| 🖨 | Print output | ← | Promote: Make the currently selected output into a heading |
| 🔍 | Print preview (see what the printout will look like) | → | Demote: Make the currently selected output into a subheading |
| | Export output to text or web (HTML) file | + | Display everything under the currently selected heading |
| | Recall a recently used command | − | Hide everything under the currently selected heading |
| ↶ ↷ | Undo or redo the last operation | | Display the currently selected object (if it is hidden) |
| ★ | Go to SPSS Data Editor (frequently used!) | | Hide the currently selected object (visible in outline view only) |
| | Go to a particular variable or case number | | Insert heading (above titles, tables, and charts) |
| | Get information about variables | | Insert title above output |
| ○ | User-defined subset of the data | | Insert text at strategic locations within the output |