Durmuş Özdemir

Applied Statistics for Economics and Business

Second Edition

EXTRAS ONLINE

Springer

Applied Statistics for Economics and Business

Durmuş Özdemir

Applied Statistics for Economics and Business

Second Edition

Springer

Durmuş Özdemir
Faculty of Business
Yaşar University
İzmir
İzmir, Turkey

Additional material to this book can be downloaded from http://extras.springer.com.

ISBN 978-3-319-26495-0 ISBN 978-3-319-26497-4 (eBook)
DOI 10.1007/978-3-319-26497-4

Library of Congress Control Number: 2016940371

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Contents

Appendix: Tables

About the Editor

Durmuş Özdemir is a professor in economics at Yaşar University, İzmir, Turkey.

He received his Bsc from Hacettepe University, his PG Dip. from the University of Glasgow, Scotland, and his PhD from the University of Leicester, England. He taught economics at the University of Leicester (1991–1995) and was a lecturer in Economics at Sussex University (1996–1998), UK. He has also taught as a visiting lecturer at Bilkent University, Ankara, and he had worked for the Université libre de Bruxelles (ULB), Belgium (2008), and Istanbul Bilgi University of Turkey (1999–2014). More than ten years of his academic life was spent in Europe.

He is interested in Growth, Trade, European Economic Integration, Macroeconomics, CGE and Structural Economic Modelling. He is a research associate at Economic Research Forum and an instructor and research fellow for the ECOMOD global economic modelling group where he is involved in building several country models, including the model used for the peace dividend for Greece and Turkey.

He has edited several conference proceedings and published number of papers, books, and chapters in edited volumes in his research area. He is a project Evaluator for the Scientific and Technological Research Council of Turkey. He is also an advisor to the European Parliament on the issues of trade and economic relations between Turkey and the EU.

Amongst other things, he has extensive experience in teaching statistics at Universities both in Turkey and in the UK.

Introduction

How to Use the Textbook

At the mere mention of the word 'statistics', the average undergraduate is often filled with an expectation of mystification and failure. In turn, this well-known fact often instils in lecturers, who have been presented the unhappy task of having to teach the subject to their less than eager pupils, a similar expectation. The origin of these assumptions is difficult to pinpoint. It may, to some extent, be due to the fact that, at first, some statistical concepts are not always easy to grasp. This may lead the student to become discouraged at an early stage. Moreover, a large percentage of available textbooks, in their efforts to be as informative as possible, can appear frighteningly unwieldy and bombard the already nervous student with an information overload.

The aim of this textbook is to introduce the subject in as non-threatening a manner as possible with an emphasis on concise, easily understandable explanations in plain English. It is aimed primarily at undergraduates of business and economics, in order to provide them with the basic statistical skills necessary for further study of their subject. However, students of other disciplines will also find it of relevance. Throughout the text, it is assumed that the student has no prior knowledge of statistics.

Within the text, we have included worked examples, and at the end of each chapter, there are further exercises which the reader is strongly urged to attempt. In the same way that one would not learn to drive a car without some amount of practice, the student will find that statistical concepts will not stick very long unless they have actually worked something out for themselves. As with many things in life, to succeed in statistics, practice is of paramount importance.

On completing the text, students should expect to be able to understand and be familiar with graphs and the plotting of data, measures of location and dispersion, probability theory, statistical inference and confidence intervals, correlation, and regression analysis.

Within a university setting, these concepts will generally be taught within an applied framework that exploits the spreadsheet package Excel. To this end, where

appropriate, complementary exercises are included towards the conclusion of the relevant chapter.

CD-ROM An extensive number of data files for examples, cases, and problems in the text are included on the CD-ROM, in Excel format. The text references these data files with a special CD-ROM icon with a data file name attached to it. The CD-ROM will save your time for retyping all the relevant data. An extensive number of data files for examples, cases, and problems can be downloaded in Excel-format from http://extras.springer.com. The text references these data files with a special CD-ROM icon with a data file name attached to it.

What Is Statistics and Why Is it Important?

It is not an overstatement to say that statistics can be of relevance in all walks of life. The use of statistics forms an integral part of the fields of both business and economics. As humans, we are all innately programmed with the ability to perform statistical exercises. In a large proportion of our daily dealings, we are all manipulating incomplete information concerning our environment in order to make decisions. For example, when shopping we compare prices in the marketplace with those of our recent experience and are able to determine, within the limitations of our knowledge, whether we are being overcharged or not. In carrying out this simple exercise, we are utilising a simple, albeit flawed, form of statistics. By using tried and tested statistical techniques, in order to influence our decision-making, we can approach such questions in a more scientific manner.

The Role of Statistics

Making Sense of Numerical Information

In our role as a business person or economist, a great deal of the information which we may be called upon to deal with, be it stock market prices or unemployment figures, will be presented to us in numerical form. That is, it will be **QUANTITATIVE** data. Quite often it is the case that the sheer volume of numerical information with which we are presented can make it difficult to fully comprehend in its raw form. When this is the case, statistical techniques may be utilised for the process of summarising the most important features of such a body of information so that it is more easily understandable, whilst at the same time ensuring that little of value to us is lost.

The question of which statistical method is appropriate to use for any given set of data is often a sticking point for many students and one which it is hoped the ensuing text will adequately answer. In brief, one must bear in mind that this

depends largely upon the nature of the numerical information and on what it is intended to demonstrate. In some cases, detailed statistical analyses are necessary. At other times, simple graphs and figures are sufficient.

It must also be borne in mind that the way in which data are collected is often as important as the statistical methods that are later applied to it. In statistics, as with many things in life, only good questions get good answers.

Sampling

As a statistical term, the **POPULATION** refers to all the measurements or observations which are of interest to us. For example, if we decided that we wished to learn about the IQs of all economics students in Cyprus, then our population of interest would be all economics students in Cyprus. As the example highlights, it is not always the case that we are fortunate enough to have access to the entire population in which we are interested. At times, working with an entire population can be prohibitively time-consuming, expensive, or impractical. In order to overcome these difficulties, we often focus our attention on a smaller subset of the population in question, that is a **SAMPLE**. However, our ultimate objective is not to make statements about our sample but rather to draw conclusions about the population from which the sample has been drawn. Statistics offers us the opportunity to do this, and the following chapters will discuss sampling techniques in more detail.

Dealing With Uncertainty

In a world of uncertainty, statistics often enables us to make our best educated guess. As with any 'predictive' science, it is not able to inform us with any certainty what will happen in the future but is often useful when we wish to know what might happen.

For example by using statistical analysis of world oil markets, we could not categorically state that:

'The price of petrol WILL be higher in Turkey next month'

We could, however, venture that:

'The price of petrol is LIKELY to be higher in Turkey next month'

Taking our analysis one step further, we might also be able to comment upon our degree of certainty that the petrol prices are likely to increase.

The concept of uncertainty and prediction is dealt with further in the chapters concerning probability.

Analysing Relationships

In economics, we are often interested in the dynamics surrounding the relationship between two different variables. For example:

'Does minimum wage legislation affect the level of unemployment?'
'Does the rate of growth of the money supply influence the inflation rate?' Using simple economic theory, we can often say something useful about the relationship between the two or more variables in question. However, by using statistical analysis, we can go on to say by how much the changes in one variable will affect the other.

Forecasting

The desire to be able to foretell the future is an intrinsic human characteristic. In the business world, important investment decisions rely upon the ability to be able to predict possible demand and market conditions. When attempting to formulate coherent economic policies, a government requires outcome forecasts, of issues such as unemployment, GDP and inflation, under the influence of various policy options.

Our most reliable indicator of future developments in such cases comes from the statistical analysis of past trends.

Decision-Making in an Uncertain Environment

Sometimes it is not feasible to predict with any degree of certainty what will occur in the future. For example, an investor, when deciding how to balance their portfolio among stocks, bonds, and money market investments, must make their decisions when future market movements are unknown. It is in such cases that techniques for dealing with uncertainty become relevant. Some salient features of uncertainty will be outlined in later chapters and some useful techniques of how to deal with it will be explained.

The Two Main Areas of Statistics

Descriptive Statistics

Descriptive statistics is dealt with in the first few chapters of this book. Its aim is to simplify or summarise a set of data by outlining its main features. It achieves this by

the use of illustrations, as for example in graphs and diagrams. It also uses numerical techniques to assign a representative value, for example the mean, or by measuring the relationship between two variables as with a correlation coefficient.

The variables (numerical values) used can be classified as **DISCRETE** or **CONTINUOUS**. Discrete variables are the consequence of counting, for example, the number of students attending a university or the number of respondents in a survey who claim to own their own house. Continuous variables can, on the other hand, take any value within a continuum. Their value is limited only by the precision of the instrument used to measure them. For example, height and weight are examples of continuous variables.

Data can be **CROSS-SECTIONAL**. Cross-sectional data provide a picture of a situation at one point in time. In contrast, **TIME-SERIES** data present a number of observations of the variable in question measured at different points in time.

Inferential Statistics

Chapters concerning descriptive statistics are followed by those discussing inferential statistics. They outline the key concepts of inferential statistics whose aim is to provide us with some appropriate conclusions about a population using sample data drawn from it.

Chapter 1
Collecting Data

1.1 Introduction

In our introduction we discussed the fact that statistics is a means by which we can describe and derive some conclusions from data. In order to fulfil these aims we first need to collect data. There are many ways of obtaining data. The five main sources are described below.

A rich source is data that has already been published by industrial, governmental or individual sources. We can obtain data by designing experiments or by observation. Data can be obtained by a survey or panel survey and by longitudinal studies. Lastly, information technology offers vast opportunities as a source of data. For example ATM machines (bank cash machines) record every transaction. Bar codes used in the retail trade provide information about sales. Holiday bookings provide an up-to-date record of holiday sales. These are all potential sources of data.

Why do we collect data? There are many reasons, the most important of which is as a support for scientific study and for measuring the performance of an economy or a business. In market research data collection helps decision processes. For example before an investment into the production of a new good, market research into the demand for that particular good can be crucial. As a part of the decision process the existing facts, the reliability of the current information and the requirement for new information need to be carefully assessed. If the data is 'biased' or misleading it can damage the decision making process. Data collection can also be a means of satisfying political, social or personal curiosities.

D. Özdemir, *Applied Statistics for Economics and Business*,
DOI 10.1007/978-3-319-26497-4_1

1.2 Sources of Data

We will now examine the previously mentioned data sources in more detail. We must distinguish between established (secondary) data sources, which provide required statistical knowledge and are obtained through compilation, and primary data sources which are generated by data collection and in which the required statistical knowledge is not available.

1.2.1 Data Published by Industrial, Governmental or Individual Sources

This section discusses secondary (compiled) data sources. Most countries have their own national data collection institute. In Hungary, Hungarian Central Statistical Office, Eurostat in European Union countries and in Turkey the most important source of external data is the Turkish Statistical Institute (TUIK). Apart from these there are also a number of other sources of statistics. The aim of the following is to briefly outline some of these sources and to provide information regarding their web addresses.

National and Organizational Data Collection Institutes: EUROSTAT; It is the statistical office of the European Union. EUROSTAT's task is to provide the European Union with statistics at European level that enable comparisons between countries and regions. http://ec.europa.eu/eurostat/web/main

The Country Central Bank web sites: The Central Banks usually have an excellent electronic data source in their web sites and they are usually good in financial data. European Central Bank (ECB) is also good for European monetary data; https://www.ecb.europa.eu/stats/html/index.en.html, Turkish Central Bank website; www.tcmb.gov.tr, Bank Indonesia (The Central Bank of Indonesia); http://www.bi.go.id/web/en,

OECD Statistics: is a good source of general OECD country statistics. www.oecd.org/statistics/

World Bank data sources: http://data.worldbank.org/ The source is usually good at global social data

Turkish Statistical Institute (TUIK); www.tuik.gov.tr The TUIK is the main and the most substantial data provider in Turkey. The TUIK publishes a number of statistics.[1] Under general publications they have the *Monthly Bulletin of Statistics* which has been published since 1952. It includes information on national accounts, industry, building construction, finance and banking, foreign trade, transporting, prices and indexes, environment etc. The web site provide very good data but if needed it is possible to purchase further data for a small fee.

[1] See their web site at www.tuik.gov.tr/yayin/publists.htm for more detailed information.

The TUIK also has other statistical publications. For example: historical surveys, such as Statistical Indicators 1923–2015 and Women in Statistics; provincial and regional statistics, the results of income distribution surveys, consumption expenditures and National income; and data concerning agriculture and building construction. It also has information about prices and indices and foreign trade, Labour force and household labour force survey results, trade and services information, transportation (including road, sea and air statistics), data about finance and municipalities, the environment, population, demographics, justice, tourism, elections, R&D and country statistics.

The other data sources are the Economist Intelligence Unit (EIU), *Penn World Tables; World Currency Yearbook, International Financial Statistics (IFS) service of the International Monetary Fund* are to name a few more.

If the data you are looking for is related to any organization or ministry not listed above you are advised to look for the resources and web sites of that particular organization.

Before going on to collect primary data you are always strongly advised to collect as much secondary data as possible.

1.2.2 Experiments

Before conducting any experiment some advance (pilot) tests are required. These tests will be examined in more detail in future chapters when we will be considering hypothesis testing and ANOVA etc. For example if we would like to determine the effectiveness of new dandruff shampoo it is first necessary to find out who uses this type of product (target population) and then to experiment how effective the shampoo is on people selected from this population. The aim is to determine whether the product it is in fact effective.

1.2.3 Observation

Observational studies, which have become popular in recent years, have a variety of forms. They form a crucial role in researches such as those into animal behavior, astronomy, geology and biology. Where experimentation and surveys are impractical observation becomes increasingly necessary.

In the economics and business arenas they are more useful when you intend to collect information in a group setting to assist in decision making. For example the organizational or industrial behavior of these groups can be observed for decision making.

1.2.4 Sources Created by Information Technology

Nowadays most retail outlets use bar codes when they sell their goods. These bar codes automatically record inventory information as soon as a good is purchased. Similarly banks provide credit cards from which they can have access to data on short-term loans. ATM machines instantly record all banking transactions on bank balances, which are another source of data. Holidays booked through travel agencies provide data about travel and tourism. Similarly electronic passport controls on borders can provide data on the people entering or leaving a country.

1.2.5 Surveys, Panel Surveys and Longitudinal Studies

These studies have no influence over the behavior of the people being surveyed, who are merely asked questions about their attitudes, beliefs, behaviors and other characteristics. These characteristics, which are recorded by a surveyor, are called random variables. They may be classified into two types, categorical and numerical. For example, the question 'do you currently do any sport?' is categorical because the answer is restricted to 'Yes' or 'No'. The answer to the question 'how many magazines do you read per month?' is a numerical random variable because the answer is a random number. Numerical random variables can be continuous or discrete. If the answer to the question is a consequence of counting (4 or 5 per month) then it is discrete. Continuous data take any value within a continuum. For example if we want to look at how long it takes to complete a task the answer in hours, minutes and seconds, can be anything along a continuum, hence it is a continuous random variable.

An important part of primary data collection is to determine the target population. Since considering a whole population may not be feasible due to reasons of cost or size, determination of a sample may then be necessary. The population includes all observations of interest. For example, if we want to know the average graduation degree of all economics graduates in Turkey in a particular year, then the population is all economics graduates in Turkey during that year. As can be seen from this example, counting all economics graduates in a country might prove difficult. Therefore we take some part of this population, (a sample), look at their average graduation degrees and try to make inferences about the whole population from this sample. So a sample is a representative part of the population of interest. However our sample selection must be carefully considered. Not every sample is useful. The sample must represent and be able to tell us something about the population. Returning the above example: If we only included Bosphorous University and Yaşar University economics graduates degree averages in our sample, this would not be representative of the whole population. It is therefore desirable to include the results of students from other universities.

If we take our whole population as a sample then obviously we do not need to select a sample. This type of data collection is called census or complete enumeration of the identified population. The population census is the best example of these types of survey. However, a census is usually a very costly practice and often not possible.

What is the ideal size of a sample? For example, if, when predicting the results of a national election, we take 1000 people's opinions about the parties they are going to vote for, then is this sample size adequate? There are no hard and fast rules about how large or small a sample should be. Obviously the larger a sample is the more accurate the information it provides us with will be. However, a more crucial consideration is the way in which the sample is selected. An appropriately selected small sample will tell us more than a badly selected large sample.

One way of ensuring that a sample is appropriately chosen is to randomly select people and ask them their opinions. However this practice presents us with further difficulties. For instance, if we have chosen 1000 people and 10 of them live in Kars, 3 in Edirne, 4 in İzmir, 3 in Bodrum and the remainder live in Istanbul. Unless we have selected our sample whilst giving some consideration to different subgroups, that is differences in socioeconomic status, educational background, age and so on, our results will be biased. Obviously controlling our sample selection with respect to such variables is not an easy task.

Selecting a Sample:[2] Broadly speaking there are two types of sample, the probability sample and the non-probability sample. The probability sample is a sample where sample subjects are chosen on the basis of known probabilities. There are four main types of probability sample, the simple random sample, the systematic sample, the stratified sample, and the cluster sample.

Panel Survey: This type of survey utilizes the same respondents who are asked a series of questions at different periods of time. Panel surveys have become a more popular tool in recent economic research studies. They can be used for observing political situations before elections, for example opinion polls. They are also suitable, for use as part of a business investigation, for the assessing the impact of an advertising campaign.

There are two types of problem associated with such surveys. 'Panel mortality and 'panel conditioning'. The first problem arises when a panel member leaves the panel thus causing the panel to become less representative. Panel conditioning is the situation whereby panel members may become influenced by the nature of the inquiry and thus change their opinions.

Longitudinal Studies: This type of study is a form of panel study but it is applied to a group of people or cohort over a longer period of time. For example we may wish to look at the lifetime earnings pattern of a cohort. Due to the problem of panel mortality a longitudinal study requires a larger initial group or cohort.

[2] For all these sampling methods see the first four references at the end of this book.

Longitudinal or panel studies are more commonly undertaken in the Developed countries. The application of this type of survey is rather poor in developing countries.

Having clarified the type of sample we wish to use, we now need to determine the questions that we are going to ask in a survey.

1.3 Questionnaire Design

The problem of what questions to ask and how they are to be used are the two most important considerations when designing a survey questionnaire. If the questions are biased or the interviewer makes mistakes in recording the answers to the questions, the outcome will not be of any use.

Whichever method we use, the preparation of questions needs careful consideration and one must have clearly determined objectives.

In general the procedure for questionnaire design may be outlined under the following four steps. The first step is to determine what we wish the questionnaire to tell us and to decide on its length. The second step is to determine the method. The third step is formulating the questionnaire. Lastly we must test the questionnaire setting.

As was stated above, we must first be absolutely clear about what we wish our questionnaire to tell us. That is to think about what information we wish to learn. When thinking about the questionnaire's length we must keep in mind that this will have some effect on the response rate. People will be less inclined to fill out a long questionnaire. We should try to include absolutely essential questions, but we must make sure that the required information can be obtained with the fewest responses possible. We must also ensure that there is no replication of information between questions. That is that different questions are asking the same thing but in a different form.

Determination of method refers to the way in which we wish to apply our questionnaire to our target population or sample. There are a variety of methods currently in use. For example, personal interviews, postal surveys and telephone interviews. As far as response rate is concerned postal surveys have the lowest uptake compared to the other two methods. However postal survey is the cheapest method. Telephone interviews are probably the most expensive method by which to conduct a survey.

Once we have determined which method we are going to use to conduct our survey our next task is to formulate the questions:

As we mentioned earlier, the questions should be short, clear and not include any ambiguities. Consider the following questions:

1. 'Do you do sport?'
2. 'How long ago did you graduate?'

Both questions have some ambiguities. The first question does not make it clear which types of sport it includes. It also doesn't distinguish frequency i.e. once a year, once a week etc.

If we were interested in the frequency of physical training than perhaps we could ask; 'how often do you do a sport?'

The second question could be asked in more precise manner. For example:

2. State your graduation date:

Month	Day	Year

Once a questionnaire has been prepared, it should be tested on small group of people. Their opinion may help to improve the questionnaire. This type of testing, which is often called 'pilot testing', can help to improve both the clarity and the length of the questionnaire. It may be useful to add some questions at the end of the questionnaire so that respondents completing the pilot test can comment on any ambiguities that they may have noticed in the questions and their estimated time taken for completion of the questionnaire. Their recommendations may be useful.

We will now consider an example survey questionnaire. The following questionnaire is taken from the Turkish Statistical Institutes' (TUIK) Employment and Earnings Survey. The aim of this survey was to gather and compare data on weekly actual working hours, overtime hours and employees' earnings with respect to economic activity, geographical region, establishment size and other characteristics of their places of work and to observe the changes in these variables with respect to time.

The questionnaire was designed according to the data needs of the users and producers of data in the labour market, specialist opinion and International standards.

Prior to its application the questionnaire was piloted to ensure appropriate length and usability.

There are three parts to the questionnaire. The first part (A) concerns the establishment worked for. It requires information about its name, address, main activity, legal status, and so forth. The second part (B) concerns information about the person who fills in the questionnaire. The third part (C) comprises two separate parts. It first asks about the establishment's employment information and secondly for information on its gross payments to employees.

For this type of survey a reference period is usually stipulated. In this case the reference period was January 1–June 30, 2015. The survey was then applied a second time with a given reference period of July 1–December 31, 2015.

The TUIK used two methods for this survey: by mailing and by interview.

For the sample selection stratification was carried out. This was done according to branch of economic activity, region and establishment size.

For reference a copy of the TUIK Employment and Earnings Survey Questionnaire now follows.

1.4 Employment and Earning Survey Questionnaire[3]

PART A - GENERAL ESTABLISHMENT INFORMATION

1. Name of the establishment : .....

2. Address of the establishment :
   Province : .....
   District : .....
   Street : .....
   Door number : ..... Telephone no: .....
   Post code : ..... Fax no : .....

3. Main activity of the establishment : .....
   A few examples for the activity :
(Specify clearly the main activity of the establishment between 1 January 2015 and 30 June 2015 (building construction, shoe manufacturing, hotel management, restaurant management) and give a few examples to clarify this activity. In case of having more than one activity, indicate the activity which occupies most of the employees. If it is not possible to clarify activities on this basis, indicate the activity which has largest gross sale amount.)

4- Legal position of the establishment :

Individual Ownership	□1	Collective Company	□3
Limited Company	□5	Cooperative	□7
Simple Partnership	□2	Limited partnership	□4
Joint Stock company	□6	Other (specify)	□8

5- Legal status of the establishment;
Private □ 1 Public □2 ≥ Question 7

6- Is the establishment registered in Social Security institutions?
Yes □1 No □2

7- Is the establishment following directly or indirectly the results of collective bargaining?
Yes □1 No □2

[3] This questionnaire is taken directly from the TUIK's Employment and Earnings Survey as an example. For further details see the original reference.

PART B - INFORMATION ABOUT PERSONS WHO FILLED IN THIS QUESTIONNAIRE

8- Person who gave information about the establishment:
Name and Surname : ..... Position : .....
Address : ..... Telephone no : .....
Fax no : .....
Date : ...../...../20.....

PART C - EMPLOYMENT INFORMATION AND EARNINGS OF EMPLOYEES
(CAUTION! Please read the whole questionnaire before beginning to fill it in. Questions 1 and 2 refers to 30 June 2015, all other questions refer to the period 1 January 2015 - 30 June 2015.)

I. Employment Information	Total employees	Employees work with collective bargaining agreement
1- Number of owners and partners on 30 June 2015	.....	.....
2- Number of unpaid family workers on 30 June 2015	.....	.....
3- Total number of paid employees in the period January 2015-30 June 2015	.....	.....
4- Average weekly normal working hours of employees (hour) (Excluding overtime hour)	.....	.....
5- Total actual working hours of all employees (employee-hour) in the period 1 January 2015-30 June 2015	.....	.....
6- Total overtime working hours of employees (employee-hour) in the period 1 January 2015-30 June 2015	.....	.....
7- Total number of paid days of employees (employee-day) in the period 1 January 2015-30 June 2015	.....	.....

8- Total number of days paid to the employees for days not worked (employee-day) in the period 1 January 2015-30 June 2015

8.1- Annual paid leave (employee day) ....

8.2- National and religious festivals (excluding weekends) (employee-day) ....

8.3- Weekends (employee day) ....

8.4- Other days paid but not worked (for instance birth, illness etc.) (employee day) ....

8'- Total number of days paid but not worked (8.1+8.2+ 8.3+8.4) ≥ ....

II. Gross Payments to Paid Employees

9- Total gross wages and salaries to employees in the period 1 January 2015-30 June 2015 (Gross wages and salaries before deducting the social security contribution payments will be written. Advance payments will not be included).

9.1- Basic gross wage and salary payments to employees in the period 1 January 2015-30 June 2015

9.1.1- Gross payments to employees for the time worked in the period 1 January 2015-30 June 2015 ....

9.1.2- Gross payments to employees for the time not worked in the period 1 January 2015-30 June 2015 ....

TOTAL BASIC GROSS WAGES AND SALARIES (9.1.1+9.1.2) ....

9.2- Other gross payments to the employees in the form of wages in the period 1 January 2015-30 June 2015

9.2.1- Dirty and dangerous work, financial responsibilities premiums ....

9.2.2- Regional bonuses ....

9.2.3- Shift work payment ....

9.2.4- Night work payment ....

9.2.5- Other payments (production/quality premium. Encouragement wage system payments. Payments like cashier, foreign language, responsibility premium etc.) ....

TOTAL OF OTHER GROSS PAYMENTS IN THE FORM OF WAGES ((9.2.1)+(9.2.2)+(9.2.3)+(9.2.4)+(9.2.5)) ≥ ....

9.3- Total gross payments to employees for overtime work in the period 1 January 2015-30 June 2015 ....

9.4- Total gross payments to employees due to wage differences arising from collective agreement in the period 1 January 2015 -30 June 2015 ....

TOTAL WAGES AND SALARIES ((9.1)+(9.2)+(9.3)+(9.4) (A)

....

10- Total gross bonus payments to employees in the period 1 January 2015-30 June 2015. (Bonuses in the form of wages will be placed in 9.2.5, not here) (B)

....

11- Total gross premium payments to employees in the period 1 January 2015-30 June 2015. (Premiums in the form of wages will be placed in 9.2.5, not here) (C). ....

12- Social allowances in cash and in kind paid to the employees in period 1 January 2015 - 30 June 2015.

	Put a check (X) for monthly allowances		
12.1- Family child education allowances	□	.....	.....
12.2- Marriage, birth and death allowance	□	.....	.....
12.3- Fuel allowance	□	.....	.....
12.4- Residence allowance (in cash)	□	.....	.....
12.5- Food allowance	□	.....	.....
12.6- Clothing allowance	□	.....	.....
12.7- Transportation allowance	□	.....	.....
12.8- Health allowance	□	.....	.....
12.9- Cultural allowance	□	.....	.....
12.10- Allowances for religious festivals	□	.....	.....
12.11- Allowances for leaves	□	.....	.....
12.12- Other social allowance	□	.....	.....

TOTAL SOCIAL ALLOWANCES IN CASH AND IN KIND (12.1)+(12.2)+(12.3)+(12.4)+(12.5)+(12.6)+12.7)+ (12.8)+(12.9)+(12.10)+(12.11)+ (12.12) (D) ....

TOTAL GROSS PAYMENTS (A)+(B)+(C)+(D) ....

13- Total social security contribution paid by employees ....
(Sll premium paid by employee. Payments to the retirement fund paid by employees. Payments to the saving fund paid by employees. Seal and income tax)

As was already mentioned, the above survey questionnaire, taken from SIS sources, was applied twice. The questionnaires were first mailed in October 2015 and again in May 2015. For those establishments not responding to or not completing the questionnaires properly after the second posting, interviewers were sent to complete the survey.

Once all the questionnaires had been completed the data were computer processed by the Institute's data processing center. The results were then published.

There are a number of software packages available for data handling. The most popular ones used in statistics are, SPSS, EXEL, R and STATA. For regression and modelling analysis there is an even larger choice, for example, Eviews, Matlab, and GAMS to name but a few.

As part of our daily lives we are constantly being presented with the results of surveys. We can see these results when we listen to the news or a TV program and when we read a newspaper or a magazine.

In our earlier discussion we said that it is very important to avoid errors when conducting a survey. One way of eliminating potential errors is to use random sampling probability methods. However even when a survey uses random probability sampling methods there is still some potential for errors. The aim of a good survey is to minimize these potential errors. Potential errors to consider are those of measurement, sampling, coverage or selection bias, selection bias and non-response.

Measurement errors come from inaccuracies in the recorded responses. The reasons for this could be due to weaknesses in the questionnaire's wording, respondents' poor efforts or the influence of interviewer.

Sampling errors arise from the chance differences from sample to sample. Coverage error is due to the exclusion of certain groups of individuals from the sample. This error leads to the problem of selection bias. Hence the sample's representation of population may be questionable.

The non-response error is self-explanatory. It results in not all the opinions from a sample of individuals being taken. This situation creates a problem known as the non-response bias.

As can be imagined, even the results of what seems to be a perfect survey include some errors.

In our next chapter we will consider the presentation of numerical information obtained from these resources.

1.5 A Review of This Chapter

In this chapter we outlined ways by which we can obtain data. Firstly we focused on the secondary data sources. The second part of the chapter outlined questionnaire design, how surveys are conducted and their possible problems. The TUIK questionnaire was included as a typical example of a survey that has been carried out.

1.6 Review Problems for Data Collecting

Q.1.1. Find a source, such as a newspaper article or textbook, that uses statistics.
   (a) Are the data from the entire population or just from a sample? What is a sample?
   (b) Classify each as being of discrete or continuous data and state reasons why you consider them to be informative or misleading

Q.1.2. Visit a data collection Institutes' web site, whose address was provided earlier, and download some data on a topic of your interest.

Q.1.3. Explain the meaning of population and sample. Why is it essential to clearly define the population when conducting a survey?

Q.1.4. Which of the survey methods discussed might have the lowest response ratio and why?

Q.1.5. Construct a simple questionnaire and criticize its layout, question ordering and wording.

Chapter 2
Data Presentation: Graphs, Frequency Tables and Histograms

2.1 Introduction

Illustrations are generally not used enough as a tool for analysing data and it is often the case that, on their own, large batches of raw data can appear overwhelming to the observer. If a body of information is presented using a simple diagram or graph then it is, as a rule, more readily understandable. In general diagrams are much easier on the eye, being more visually attractive. They are also easier on the brain, in that they are less difficult to comprehend at a glance. Another advantage is that they may also reveal aspects of data that might otherwise be overlooked. Their disadvantage lies in the fact that they are not always precise and often don't allow for further statistical analysis.

2.2 Simple Graphs and Charts

2.2.1 Bar Charts

Bar charts are a frequently used form of graph. They enable quick visual comparisons of facts and figures. Their use is most appropriate in situations where the data set of interest is **discrete** (as a result of counting) and contains only a few variables.

As the following example illustrates, the composite bars may be drawn horizontally (along the graph's 'Y' axis), or, as is the more conventional choice, vertically (along the 'X' axis).

A sample of 100 university students were asked to state in what way they preferred to spend their Saturday evening. The results were as follows:

The online version of this chapter (doi:10.1007/978-3-319-26497-4_2) contains supplementary material, which is available to authorized users.

D. Özdemir, *Applied Statistics for Economics and Business*,
DOI 10.1007/978-3-319-26497-4_2

Table 2.1 Preferred Saturday evening leisure pursuits of 100 students

Preferred leisure pursuit	Number of students
Cinema	31
Musical concert	23
Watching television	19
Theatre	5
Restaurant	20
Studying	2
Total	100

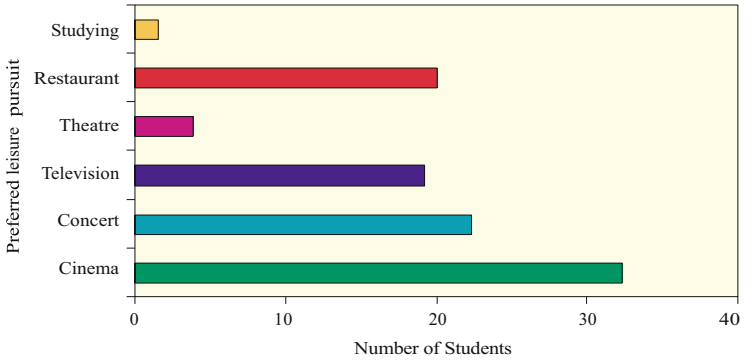


Fig. 2.1 Preferred leisure pursuit Horizontal bar chart of preferred Saturday evening leisure pursuits of 100 students

These results of Table 2.1 were then plotted as horizontal and vertical bar charts (Figs. 2.1 and 2.2).

When plotting bar charts there are several important features that must be kept in mind. Firstly, the length of each bar is directly proportional to the frequency. Secondly, the individual bars must be of uniform width and equidistant from each other. Axes must be kept to scale and must be labelled clearly. It is often useful to first organise the data to be illustrated in to a table.

The following example, which utilises average maximum and minimum monthly temperatures for Istanbul, demonstrates how it is possible to illustrate more than one outcome for a variable on the same axes of a bar chart (Table 2.2; Figs. 2.3 and 2.4).

If, as might be the case with the above data, we would like to present all of our information simultaneously, without having to resort to two graphs, we can use what is called a **Multiple bar chart**. Such a chart enables us to directly compare our two (or more if that is the case) sets of data.

Note. To distinguish between the two sets of data it is important to provide a key insert.

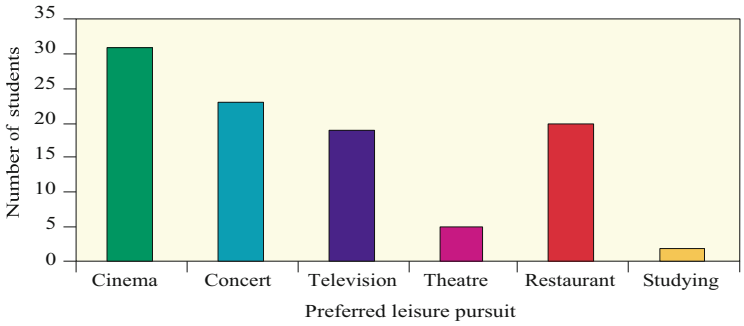


Fig. 2.2 Vertical bar chart of preferred Saturday evening leisure pursuits of 100 students

Table 2.2 Average max. and min. monthly temperatures for Istanbul (centigrade)

Month	Jan.	Feb.	March	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Max. temp.	9	10	11	17	21	27	29	30	27	20	15	10
Min. temp.	3	2	4	8	11	17	19	20	17	13	9	5

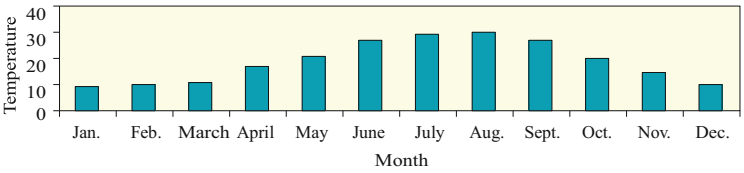


Fig. 2.3 Bar chart of the average maximum monthly temperatures for Istanbul (centigrade)

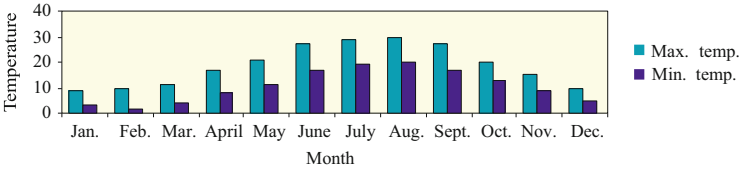


Fig. 2.4 Multiple bar chart for average min. and max. monthly temperatures for Istanbul (centigrade)

2.2.2 Pareto Charts

This vertical chart is a slight variation on the conventional bar chart. It differs in that the bars, whose heights are still directly proportional to the frequency, are arranged in order of decreasing height from left to right, that is starting with the tallest on the

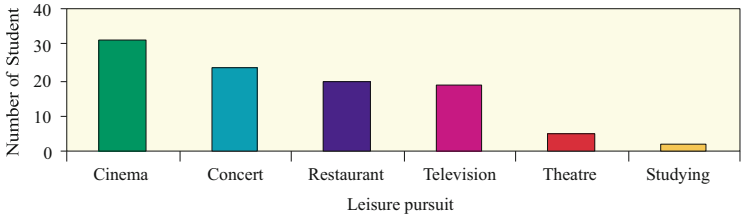


Fig. 2.5 Pareto chart of the preferred Saturday evening leisure pursuits of 100 students

left. The figure below is a Pareto chart constructed using our student data from Table 2.1 (Fig. 2.5).

2.2.3 Pie Charts

Pie charts are a popular form of illustration. They are easy to construct and don't lend themselves easily to misinterpretation. They provide a quick sense of the percentages by which a variable is distributed between different categories. The total area of the chart is equivalent to 100 %. The area of each segment is proportional to its respective variable's percentage share of the whole sample. If the chart is to be drawn by hand then it will be necessary to use a protractor. In order to do this it is first necessary to calculate each individual segment's share of the circle's 360°. The segments are conventionally labelled with their corresponding percentage of the total. It must be noted that as the number of categories to be represented becomes larger, the more difficult the chart becomes to interpret.

Example of How to Draw a Pie Chart

Refer back once again to the data in Table 2.1 which shows the preferred Saturday evening recreational pursuits of 100 university students (Fig. 2.6).

Preferred pursuit	No. students	No. students as % of total	No. degrees in the chart
Cinema	31	31 %	$(360/100) \times 31 = 111.6$
Music concert	23	23 %	$3.6 \times 23 = 82.8$
Watching television	19	19 %	$3.6 \times 19 = 68.4$
Theatre	5	5 %	$3.6 \times 5 = 18$
Restaurant	20	20 %	$3.6 \times 20 = 72$
Studying	2	2 %	$3.6 \times 2 = 7.2$
Total	100	100 %	360°

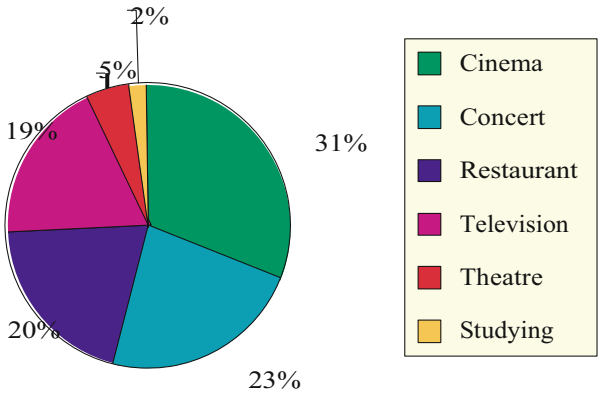


Fig. 2.6 Pie chart of preferred Saturday evening leisure pursuits of 100 students

2.2.4 XY Graphs

These are often used for time series data, with time represented on one of the axes. The below data is taken from the State Institute of Statistics, Statistical Indicators (1923–2006). It provides information on per capita GNP growth for Turkey during the Republican period (1923–2006). If we observe the data in its raw form it is not immediately obvious what the usual trend of per capita economic growth is. In order to see this trend more clearly we can construct an XY chart.

ch2data 1.xls

If we represent years on the 'X' axis and growth per capita on the 'Y' axis then we can construct a typical XY chart (as shown below). XY charts are a simple means by which data can be represented graphically. In this example the chart provides information on the economic growth of Turkey from 1923 to 2006 (Fig. 2.7).

As can be appreciated, economic growth on a yearly basis is seen more clearly on the line diagram as opposed to from the figures presented in Table 2.1. If we closely examine the diagram we will see that some years show negative growth and some years show high positive growth. Since 1954 a more stable growth path is observed that is neither high nor low.

2.3 Histograms

The distribution of measurement on a **continuous scale** (as the result of using a measuring device) is presented by the use of a histogram. Consider the table below which gives the times (to the nearest minute) it takes 50 university students to travel from their home to campus on an average day (Table 2.4).

Table 2.3 GNP per capita growth rates at producers' prices, 1923–2006

Years	GNP growth	Years	GNP growth	Years	GNP growth	Years	GNP growth
1923		1944	−5.9	1965	0.6	1986	4.4
1924	12.5	1945	−16.1	1966	9.2	1987	7.5
1925	10.5	1946	29.2	1967	1.6	1988	−0.7
1926	15.8	1947	1.9	1968	4	1989	−0.6
1927	−14.6	1948	14.1	1969	1.7	1990	6.8
1928	8.7	1949	−7	1970	1.8	1991	−1.6
1929	19.2	1950	6.9	1971	4.4	1992	4.4
1930	0	1951	10	1972	6.5	1993	6.2
1931	6.7	1952	8.8	1973	2.3	1994	−7.8
1932	−12.8	1953	8.2	1974	0.7	1995	6.2
1933	13.5	1954	−5.6	1975	3.3	1996	5.3
1934	3.8	1955	5	1976	6.8	1997	8.7
1935	−5.1	1956	0.7	1977	0.9	1998	2.3
1936	21.1	1957	4.4	1978	−0.8	1999	−7.4
1937	−0.2	1958	1.6	1979	−2.5	2000	1.4
1938	7.7	1959	1.1	1980	−4.8	2001	−11.1
1939	4.2	1960	0.5	1981	2.3	2002	6.4
1940	−6.8	1961	−0.6	1982	0.6	2003	4.2
1941	−11.6	1962	3.6	1983	1.7	2004	8.2
1942	4.6	1963	7	1984	4.5	2005	7.2
1943	−10.8	1964	1.5	1985	1.7	2006	4.6

Source: SIS Statistical indicators (1923–2006)

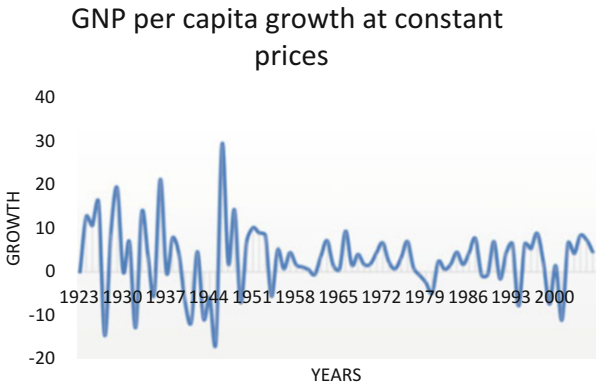


Fig. 2.7 XY chart for the Republican years of Turkish economic growth

Before constructing a histogram it is necessary to condense the above data in to more manageable chunks, that is by compiling a **frequency table**. The data is first grouped into subintervals of similar outcome known as **classes**. The number of observations falling into each class is termed the **class frequency**. For any one class a **cumulative frequency**, which is the sum of frequencies in that class and all those preceding it, can also be determined if desired. The number of classes formed is a

Table 2.4 Times taken to travel from home to campus (to nearest min.) of 50 students

21	25	120	55	66	40	30	10	85	47
41	09	35	17	90	21	29	16	60	19
18	08	34	39	18	111	62	43	18	28
04	98	11	84	21	46	06	33	25	15
12	28	102	118	69	09	115	67	80	101
13									

matter of personal preference, between 5 and 15 classes is an ideal figure. The range of each class (the outcomes which it spans) is termed **class width**, it is determined using the simple formula below.

$$\text{Class width} = \frac{\text{Largest data value} - \text{smallest data value}}{\text{Desired number of classes}}$$

For ease of usage the result is then rounded to the nearest whole number.

If, for example, we wanted to construct 12 classes using our data from example 3 then our class width would be as follows:

$$\text{Class width} = 10(\text{approx.})$$

$$\text{Class width} = \frac{120 - 4}{12} = 9.6$$

The lower limit (smallest value) and upper limit (largest value) of each class are respectively called the **lower class limit** (LCL) and the **upper class limit** (UCL).

The class width is the difference in value between the lower class limit of any particular class and the lower class limit of the class immediately above or below it.

The centre value of each class is called the **midpoint** or **class mark**.

As far as is possible the class widths should be equal in value, although this is not always the case as will be discussed later.

ch2data2.xls

Using the above principles a frequency table of the data shown in Table 2.5 has been constructed above. Note that the cumulative frequency is a useful tool for double-checking that your data count is accurate.

A histogram can be used to represent the above frequency table diagrammatically. Histograms are similar to bar charts but differ in that the **individual bars always touch** and that the **width of each bar is representative of class width**. That is if a class width were to double in size in comparison to the others then so too would the width of the corresponding bar. If the class widths are all the same then the height of the bars is proportional to the frequency and the area of each bar is proportional to the frequency in each class.

Table 2.5 Times (mins.) taken by 50 university students to travel from home to campus on an average day

Class (mins. taken)		Class boundary	Freq.	Cumulative frequency	Class mark
LCL	UCL				
1	10	10.5	6	6	5.5
11	20	20.5	9	15	15.5
21	30	30.5	9	24	25.5
31	40	40.5	5	29	35.5
41	50	50.5	4	33	45.5
51	60	60.5	2	35	55.5
61	70	70.5	4	39	65.5
71	80	80.5	1	40	75.5
81	90	90.5	3	43	85.5
91	100	100.5	1	44	95.5
101	110	110.5	2	46	105.5
111	120	120.5	4	50	115.5

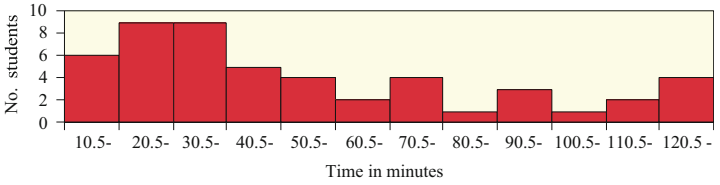


Fig. 2.8 Histogram of times taken (mins.) by 50 students to travel from home to campus on an average day

As can be seen from the above data there is a gap between the upper class limit of one class and the lower class limit of the next. It has already been stated that the bars of a histogram must touch. To overcome this we calculate **class boundaries**. These class boundaries represent the midpoint between one class and the next. The boundaries go on to form the upper and lower limits of our bars and are shown on the axis.

As was already implied above, it is not always the case that class widths are equal. This might occur if on constructing a frequency table it is found that a number of class intervals have very low frequencies. In such a case the frequencies of the classes are sometimes summed and the class width increased accordingly. If this occurs then the area of each bar of the histogram is no longer representative of the frequency. Remember that the widths of the individual bars are drawn in proportion to class width. Thus the histogram can no longer be relied upon to give an accurate picture of the shape of the distribution. The simple example below attempts to explain this concept diagrammatically.

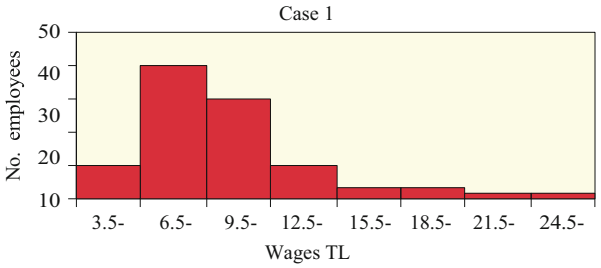


Fig. 2.9 Histogram of ave. hourly wages (TL) of 100 employees (Case one)

Fig. 2.10 Histogram of ave. hourly wages (TL) of 100 employees. (Case two)

Consider the frequency table below which shows the average hourly wages, to the nearest 1 Turkish Lira, (TL) received by a random sample of 100 employees of a company. This will be referred to as case one.

As can be seen from the table the higher wage earners (from 13 TL to 24 TL) are fewer in number. If the last four classes are joined a new Frequency table can be obtained. (Case two).

If the histograms for the two frequency tables are now drawn it is apparent that in case two (unequal class widths) a false impression is given of the proportion of staff earning a higher wage. That is, the area representing class interval 12.5–24.5 is no longer proportional to the class frequency and gives the impression that a larger number of employees earn a higher wage than is actually the case.

2.3.1 Frequency Density

One way of overcoming the problem, illustrated above, is to construct a histogram of the **frequency density**. This compensates for variations in class widths and is derived as follows:

Table 2.6 Average hourly wages (TL) of 100 company employees. (Case one)

Hourly wage (TL)		Freq.	Class boundary	Cumulative freq.
LCL	UCL			
1	3	10	3.5	10
4	6	40	6.5	50
7	9	30	9.5	80
10	12	10	12.5	90
13	15	3	15.5	93
16	18	3	18.5	96
19	21	2	21.5	98
22	24	2	24.5	100

Table 2.7 Average hourly wages (TL) of 100 company employees. (Case two)

Hourly wage (TL)				
LCL	UCL			
1	3	10	3.5	10
4	6	40	6.5	50
7	9	30	9.5	80
10	12	10	12.5	90
13	24	10	24.5	10

Table 2.8 Frequency densities for average hourly wages (TL) or 100 employees

1. Case one					
Hourly wage (TL)		class width	freq.	freq. density	class boundary
LCL	UCL				
1	3	3	10	3.3	3.5
4	6	3	40	13.3	6.5
7	9	3	30	10	9.5
10	12	3	10	3.3	12.5
13	15	3	3	1	15.5
16	18	3	3	1	18.5
19	21	3	2	0.7	21.5
22	24	3	2	0.7	24.5

$$\text{Frequency density} = \frac{\text{frequency}}{\text{Class width}}$$

For the above data (Tables 2.6 and 2.7) the frequency densities are calculated below; firstly for case one, where class widths are equal, secondly for case two where the last four class widths are grouped.

If the histograms are now redrawn it can be seen that by calculating the frequency density a more accurate picture is given in the second case. It must be remembered that the height of each bar now indicates the frequency density.

Table 2.9 Freq. densities for ave. hourly wages of 100 employees. (Unequal class widths)

2. Case two					
Hourly wage (TL)		class width	freq.	freq. density	class boundary
LCL	UCL				
1	3	3	10	3.3	3.5
4	6	3	40	13.3	6.5
7	9	3	30	10	9.5
10	12	3	10	3.3	12.5
13	24	12	10	1.2	24.5

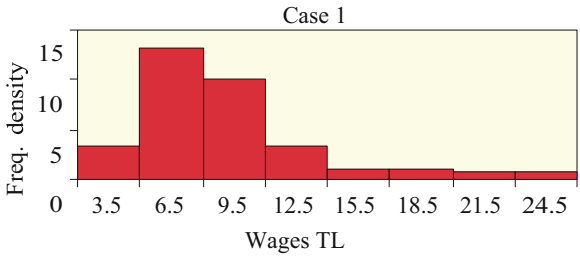


Fig. 2.11 Frequency density for Case I

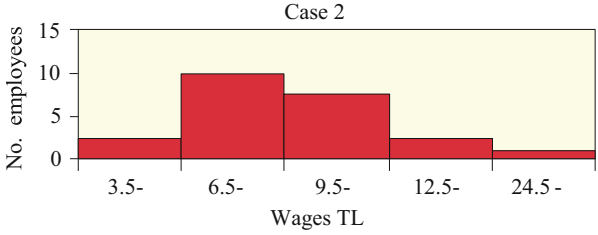


Fig. 2.12 Frequency density for Case II

2.3.2 Frequency Polygon

When observing a histogram it can often appear as if the changes in frequency from class to class are rather abrupt. A smoother, more continuous looking curve can be obtained by constructing a **frequency polygon**. This is achieved by joining the mid points of each consecutive bar of the histogram. To the left and right of the histogram the line of the curve is extended to include zero. This type of curve is particularly useful if two separate distributions are being compared on the same axes. As an example our wages data from Table 2.6 has been plotted to form a frequency polygon.

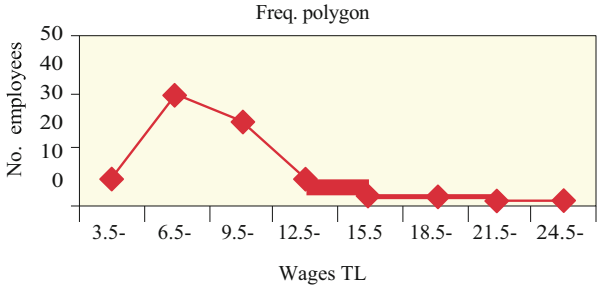


Fig. 2.13 Freq. polygon of ave. No. employees hourly wages (TL) of 100 employees

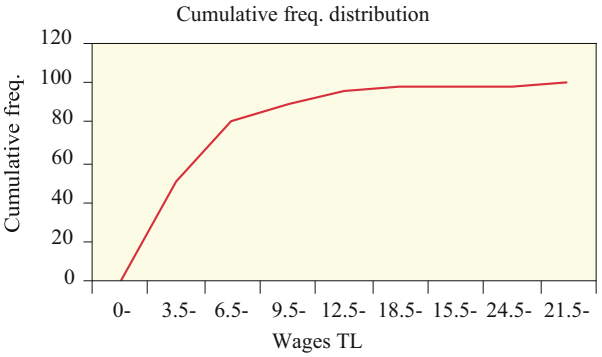


Fig. 2.14 Cumulative frequency distribution of ave. hourly wages (TL) of 100 employees

2.3.3 Cumulative Frequency Distribution

The cumulative frequency and how it is obtained has already been discussed. It can be represented diagrammatically using what is called an **ogive**. Firstly the upper class boundaries of the frequency table are marked along the 'X' axis. The corresponding cumulative frequencies are plotted along the 'Y' axis. The lower class boundary to the left of the distribution begins with a cumulative frequency of zero. The individual plots are joined using straight lines. Cumulative frequency distributions are useful for calculating the number of scores above or below a certain level. Below is an example drawn from the data in Table 2.6.

2.3.4 Relative Frequency Distribution

This is constructed in exactly the same way as the cumulative frequency distribution except that on the 'Y' axis the cumulative frequency is replaced with the **relative frequency**. The relative frequency converts the frequency of each class interval into a percentage of the total frequency. The formula used to obtain it is as shown below:

$$\text{Relative frequency} = \frac{\text{frequency}}{\text{total frequency}} \times 100$$

2.4 A Review of This Chapter

This chapter firstly examined the simplest forms of data interpretation. That is, graphs frequency tables and histograms. For our section on graphs we firstly looked at bar charts and then went on to explain Pareto charts, pie charts and line graphs or XY diagrams. These were described using simple data and we also used some real data examples. In the second part of this chapter we looked at more complicated graphs: histograms, frequency density, and frequency polygon.

2.5 Review Problems for Graphical Presentation

Q.2.1 The data below refer to the monthly quantity of anchovy (hamsi) landed at Black Sea ports, and the average price paid each month for the fish

		Tonnes landed		Price (TL per tonne)
	2014	2015	2014	2015
January	4321	2951	576	794
February	3816	4316	620	631
March	4047	5179	608	602
April	4954	5553	629	646
May	6171	7449	635	602
June	6144	7313	605	604
July	4998	5396	645	583
August	3856	5070	749	669
September	4352	3727	686	752
October	3696	2925	711	834
November	4461	4029	683	769
December	4082	2490	649	908

ch2data3.xls

(a) Draw time series graphs of both the quantity landed and the price of anchovy.
(b) Calculate the percentage change to June 2000 in tonnes landed from, (1) June 1999 (2) December 1999, (3) May 2000.

(c) Plot a scatter diagram with quantity landed on the x axis and price on the y axis. Any comments?

Q.2.2 The data below refer to GNP growth rates at producers' prices. 1923–2006

1923		1944	−5.9	1965	0.6	1986	4.4
1924	12.5	1945	−16.1	1966	9.2	1987	7.5
1925	10.5	1946	29.2	1967	1.6	1988	−0.7
1926	15.8	1947	1.9	1968	4	1989	−0.6
1927	−14.6	1948	14.1	1969	1.7	1990	6.8
1928	8.7	1949	−7	1970	1.8	1991	−1.6
1929	19.2	1950	6.9	1971	4.4	1992	4.4
1930	0	1951	10	1972	6.5	1993	6.2
1931	6.7	1952	8.8	1973	2.3	1994	−7.8
1932	−12.8	1953	8.2	1974	0.7	1995	6.2
1933	13.5	1954	−5.6	1975	3.3	1996	5.3
1934	3.8	1955	5	1976	6.8	1997	8.7
1935	−5.1	1956	0.7	1977	0.9	1998	2.3
1936	21.1	1957	4.4	1978	−0.8	1999	−7.4
1937	−0.2	1958	1.6	1979	−2.5	2000	1.4
1938	7.7	1959	1.1	1980	−4.8	2001	−11.1
1939	4.2	1960	0.5	1981	2.3	2002	6.4
1940	−6.8	1961	−0.6	1982	0.6	2003	4.2
1941	−11.6	1962	3.6	1983	1.7	2004	8.2
1942	4.6	1963	7	1984	4.5	2005	7.2
1943	−10.8	1964	1.5	1985	1.7	2006	4.6

Source: TUIK Statistical indicators, Turkey

ch2data4.xls

Draw an XY chart similar to the one in the text and compare the two diagrams. Can you comment about the effect of population growth on long term economic growth in Turkey?

Q.2.3 Construct the histogram and ogive for the age distribution of cars in 1990 from the data below

Age distribution of cars			
Age		Age	No of cars
less than 2 years old		1	4127
2–4 years old		3	3958
4–6		5	3425
6–8		7	3152
8–10		9	2359
10–12		11	1648
12–14		13	755
14+		15	805
		Total	20,229

Source: UK Transport Statistics (1992)

ch2data5.xls

Q.2.4 The distribution of household income is shown in the table below. Draw a bar chart and histogram of the data and comment

Income groups TL		Total Income TL
	Number of households	
0–49,999	321,746	10,674,326
50,000–99,999	1,254,863	99,023,453
100,000–149,999	1,799,098	222,434,428
150,000–199,999	1,695,101	294,254,201
200,000–249,999	1,330,912	295,933,661
250,000–299,999	987,626	268,315,943
300,000–349,999	767,705	247,523,158
350,000–399,999	539,023	200,770,779
400,000–449,999	471,176	199,308,071
450,000–499,999	304,610	143,795,912
500,000–599,999	508,190	275,493,366
600,000–699,999	312,648	200,143,170
700,000–799,999	190,017	141,024,361
800,000–899,999	144,870	122,320,401
900,000–999,999	87,654	82,538,197
1,000,000–1,499,999	186,600	219,570,936
1,500,000–1,999,999	74,114	126,035,614
2,000,000–4,999,999	62,535	174,767,006
5,000,000–9,999,999	7502	43,815,423
10,000,000–24,999,999	1570	20,681,050
TOTAL:	11,047,560	3,388,423,456

Source: TUIK Income survey results, 1987 Table 2.1

ch2data6.xls

2.6 Computing Practical for Graphs and Charts in Economics

Although we will mainly be using Excel for these practical sheets it is also useful to know how to draw figures without data. We will first try to use the software called Paint, one of the Windows Accessories. Begin by clicking on the Paint icon on Applications.

2.6.1 Using Paint

Drawing a Line

(a) Click on the line tool.
(b) Click on the drawing space where you want the line to begin, and holding the button drag the mouse to produce your line. When you release the button, your line is fixed.
(c) For a vertical, horizontal or 45° line, press shift and keep it depressed until you have completed step (b) above.

To Remove an Item

(a) Click the eraser, and move it over what you want to rub out.
(b) If you just want to start again, click File and choose New.

To Draw a Curve

(a) Click on the Curve tool.
(b) Click and drag from where you want the curve to start to where you want it to finish. It will look like a line at this stage. Release the mouse button.
(c) Move the mouse to where you want to shape the curve, then click and drag to bend the curve to shape.
(d) When you are satisfied with the shape of the curve, click its end point to fix its shape.
(e) If the shaping process goes wrong in stage (c) you can click the right mouse button to start again.
(f) Instead of step (d), you can repeat step (c) to make the curve bend in a second direction. As you release the mouse button after the second shaping the curve will be fixed automatically.

To Add Text

(a) Click the text tool (A).
(b) Click in the drawing space where you want to position the text.
(c) Type your text.

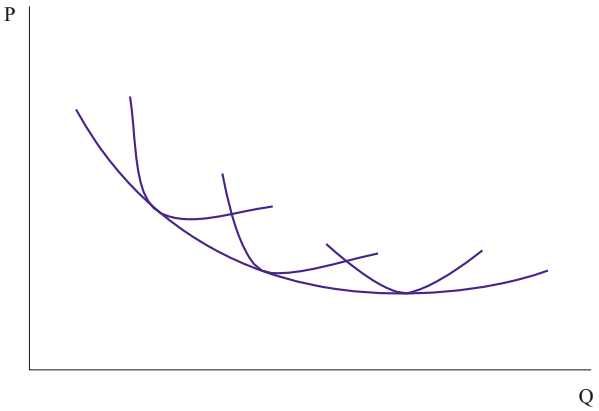
To Save Your Work: Choose **File**, **save**, type a name and click OK.

To Copy Your Drawing to the Clipboard

(a) Click the select tool to define a rectangle to cut.
(b) Click and drag to mark the rectangle to be cut in dotted lines.
(c) Choose Copy in the menu bar, then Copy.

To Close Paint: Double click the control box at the top left of the window.

To Bring your Paint drawing into another application: Open the application (say Word for Windows). Choose Copy, then Paste.



Here is a diagram from Paint:

Notice that curves and lines in Paint are not completely smooth. This is because Paint is really a painting package and not a drawing one.

Nevertheless, it is quick and easy to use, and available on any machine that runs Windows. You may find it useful for economics assignments.

C.1.1
Draw the figure above in Paint. You may like to practice some other diagrams.

2.6.2 Using Excel

Open a new Excel workbook. Type your name and a title (e.g. Sales Figures) at the top of the spreadsheet, then type *Year* in cell A3 and enter the years 1991 through to 1994 in cells A4: A7. Type *Sales* in cell B3 and enter 135, 201, 174 and 235 in cells

B4: B7. These sales figures are in Euros. Select cells B4: B7, click the currency button and click the Decrease decimal button twice.

C.1.2

(a) Select the Year and Sale figures in cells A4: B7. Click the Recommended Charts button. Next, click on the chart title and type a title for your chart.
(b) Prepare a bar chart (horizontal bars) with Year on the vertical axis (1991 on top and 1994 on the bottom) and Sales on the horizontal axis.
(c) Similarly, prepare a pie chart showing annual sales as a proportion of total sales during the four year period and a line chart showing Year on the horizontal axis and Sales on the vertical axis. Save your work.

C.1.3

The following data show GDP per capita growth rates, at current prices, of Turkey.

1923	–	**1948**	–	**1973**	22.7	**1998**	–
1924	23.6	**1949**	−6.7	**1974**	32.3	**1999**	46.9
1925	23.8	**1950**	4.8	**1975**	26.3	**2000**	57.1
1926	6.0	**1951**	17.0	**1976**	24.5	**2001**	42.2
1927	−12.6	**1952**	11.9	**1977**	25.3	**2002**	44.0
1928	8.4	**1953**	13.3	**1978**	45.8	**2003**	28.2
1929	24.3	**1954**	−0.7	**1979**	71.1	**2004**	21.5
1930	−25.2	**1955**	17.0	**1980**	79.7	**2005**	14.7
1931	−14.2	**1956**	12.6	**1981**	47.4	**2006**	15.4
1932	−17.5	**1957**	28.7	**1982**	29.5	**2007**	9.8
1933	−4.8	**1958**	16.0	**1983**	29.3	**2008**	11.3
1934	4.6	**1959**	21.5	**1984**	54.3	**2009**	−1.2
1935	5.5	**1960**	3.9	**1985**	55.6	**2010**	13.6
1936	27.0	**1961**	3.3	**1986**	42.4	**2011**	16.4
1937	4.7	**1962**	13.3	**1987**	43.1	**2012**	7.8
1938	3.1	**1963**	12.8	**1988**	69.2	**2013**$^{(r)}$	9.3
1939	5.7	**1964**	4.2	**1989**	72.1	**2014**	10.4
1940	14.7	**1965**	4.3	**1990**	68.9		
1941	23.1	**1966**	15.8	**1991**	57.2		
1942	104.7	**1967**	8.7	**1992**	70.3		
1943	47.4	**1968**	–	**1993**	77.9		
1944	−28.3	**1969**	9.0	**1994**	91.7		
1945	−19.0	**1970**	9.7	**1995**	97.2		
1946	22.8	**1971**	21.0	**1996**	87.1		
1947	7.6	**1972**	16.6	**1997**	95.9		

Source: TUIK Statistical Indicators 1923–2014

Ch2data7.xls

(a) Use Excel and enter the data above in your spreadsheet. Draw a time-series graph of Per capita economic growth. Comment upon the main features of the series.

(b) Draw time-series graphs of the change in per capita growth, the (natural) log of growth, and the change in the ln. Comment upon the results.

Print out the answers for these three questions.

Chapter 3
Measures of Location

3.1 Introduction

In the second chapter we discussed graphical techniques. Their advantage is that they provide a quick overview of data. However, they are limited in that they are not very precise and do not allow for a further analysis. The following chapter introduces some simple numerical techniques that allow us to make a further summary on data by computing an **average**.

Averages are considered as **measures of location** because they attempt to summarize data using one figure, which acts as a kind of central data value. There is more than one such a typical summary value or average. We will consider the three main ones, which are:

1. The **MEAN**—'arithmetic average'
2. The **MEDIAN**—'middle observation of an ordered distribution'
3. The **MODE**—'most frequent value'

3.2 Arithmetic Mean

The arithmetic mean is the most common used measure of location. The median and mode are only used in special circumstances, which we will define later. Basically, the mean is simply an arithmetic average. That is, the data is summed and then divided by the total number of observations. The mean of a population is denoted by the symbol μ, a sample by $\bar{x}$. N denotes the number of observations in a population, n, in a sample. $\sum$ denotes the sum. The following formulae are for the calculations of the arithmetic mean in both a population and sample respectively.

The online version of this chapter (doi:10.1007/978-3-319-26497-4_3) contains supplementary material, which is available to authorized users.

D. Özdemir, *Applied Statistics for Economics and Business*,
DOI 10.1007/978-3-319-26497-4_3

$$\textbf{Population mean}\ \ \mu = \frac{1}{N}(X_1 + X_2 + X_3 + \cdots + X_n),\ = \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$\textbf{Sample mean}\ \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Depending on the type of data available, the formula used to calculate the mean can vary slightly as the following examples demonstrate.

Depending on the type of data available, the formula used to calculate the mean can vary slightly as the following examples demonstrate.

3.2.1 Un-Tabuled Data (Ungrouped Data)

Here the mean is a simple average. For example, in a class consisting of 13 students the statistics examination grade for each student is as follows:

23 %, 72 %, 28 %, 42 %, 50 %, 55 %, 60 %, 52 %, 48 %, 35 %, 62 %, 67 %, 56 %.

To calculate the mean examination grade of the class, the individual grades are first summed to produce a total. This total is then divided by the number of observations, which in this case is 13 (n = 13).

$$\sum_{i=1}^{n} X_i = 650\ n\ =\ 13$$

Thus the average class grade is 50 %

3.2.2 Tabulated (Grouped Data)

For this the following formula may be used:

$$\overline{X} = \frac{\sum_{i=1}^{C} f_i X_i'}{\sum_{i=1}^{C} f_i}$$

Where:

f_i = Class frequency.
X_i' = Midpoint of the class interval.

C = Number of class intervals.

This type of mean is also called a **weighted mean.** As the individual X values are unknown in the grouped data (that is they are not represented on the frequency table) X', which denotes the mid-point of each class interval, is used. gives the total number of observations which is equal to the sum of the individual frequencies.

Example

Consider the grouped data below.

ch2data5.xls

In order to calculate the mean the data in Table 3.1 can be represented as in Table 3.2 below:

By incorporating the data from table two into the grouped data formula, we are able to calculate the average car age in that country in 1990.

$$\sum fi = \mathrm{N} = 20229$$
$$\sum_{i=1}^{C} fiX'i = 116439$$

For this the following formula may be used:

$$\overline{X} = \frac{\sum_{i=1}^{C} f_i X_i'}{\sum_{i=1}^{C} f_i} = \frac{116439}{20229} = 5.7560$$

Thus it can be concluded that the mean car age in 1990 is 5.6 years.

3.3 Median

The Median is the middle value of an ordered distribution. By 'ordered' we mean that the data is arranged according to the value starting from the smallest figure upwards.

Table 3.1 Age distribution of cars in the U.K. in 1990 (thousands)

Age distribution of cars		
Age	Age	No of cars
less than 2 years old	1	4127
2–4 years old	3	3958
4–6	5	3425
6–8	7	3152
8–10	9	2359
10–12	11	1648
12–14	13	755
14+	15	805
Total		20,229

Source: UK Transport Statistics 1991

Table 3.2 Calculations for the age distribution of cars

Upper age limit	Midpoint X'	Cars f	fX'
0		0	0
2	1	4127	4127
4	3	3958	11,874
6	5	3425	17,125
8	7	3152	22,064
10	9	2359	21,231
12	11	1648	18,128
14	13	755	9815
16	15	805	12,075
		20229	116,439

3.3.1 Calculating the Median for Ungrouped Data

Referring to the statistics examination grades, which are given once again:

23 %, 72 %, 28 %, 42 %, 50 %, 55 %, 60 %, 52 %, 48 %, 35 %, 62 %, 67 %, 56 %.

If these are arranged in order of increasing value:

$$\underbrace{23,\ 28,\ 35,\ 42,\ 48,\ 50,}_{\text{6 below}}\ \mathbf{52},\ \underbrace{55,\ 56,\ 60,\ 62,\ 67,\ 72}_{\text{6 above}}$$

We may observe that the median examination grade is 52 %

In the above case, the location of the middle value presents no difficulty as we have an odd number of observations. In the case of an even number of observations, the median is obtained by calculating the average value of the two most central values of the ordered distribution.

For example, if another student's mark was included in the above example, for instance 65 %, making the total number of grades 14 and the ordered distribution would be as below:

$$\underbrace{23, 28, 35, 42, 48, 50,}_{\text{6 below}} \ \mathbf{52,\ 55},\ \underbrace{56, 60, 62, 65, 67, 72.}_{\text{6 above}}$$

Thus the median is the average of the two middle values 52 and 55.

$$\text{Median} = \frac{(52 + 55)}{2} = 53.5\,\%$$

Compared to the mode, the median is more stable. However, it does not tell u anything about the range of values above or below it. For example, consider the following two sets of data:

a)	53	54	**55**	56	92
b)	53	54	**55**	56	57

The median is 55 in both groups 'a' and 'b' above. However, the range of values in each group differs.

In group a, the range of values is larger, 53–92. In group b, the range of values is smaller, 53–57.

This difference in range affects the mean value of each group. In group a, the mean value is 62.

In group b, the mean value is 55.

It can be seen from this observation that the median, unlike the mean, is less affected by extreme outlying observations.

3.3.2 Calculation of the Median for Grouped Data

There are two methods we can use to determine median, namely the graphical method and by calculation. The first step in both cases is to calculate cumulative frequencies.

Let us use our household income data again. The cumulative frequencies are calculated in the below data. It can be noted that 321,746 respondents have less than TL 50,000,000 worth of income, 1,576,609 respondents have less than TL 100,000,000 worth of income and so on (Table 3.3).

ch2data6.xls

Table 3.3 Income distribution data for Turkey

Income groups TL	Number of households	Total Income TL
0–49,999	321,746	10,674,326
50,000–99,999	1,254,863	99,023,453
100,000–149,999	1,799,098	222,434,428
150,000–199,999	1,695,101	294,254,201
200,000–249,999	1,330,912	295,933,661
250,000–299,999	987,626	268,315,943
300,000–349,999	767,705	247,523,158
350,000–399,999	539,023	200,770,779
400,000–449,999	471,176	199,308,071
450,000–499,999	304,610	143,795,912
500,000–599,999	508,190	275,493,366
600,000–699,999	312,648	200,143,170
700,000–799,999	190,017	141,024,361
800,000–899,999	144,870	122,320,401
900,000–999,999	87,654	82,538,197
1,000,000–1,499,999	186,600	219,570,936
1,500,000–1,999,999	74,114	126,035,614
000,000–4,999,999	62,535	174,767,006
5,000,000–9,999,999	7502	43,815,423
10,000,000–24,999,999	1570	20,681,050
TOTAL:	11,047,560	3,388,423,456

Firstly, the class interval containing the median respondent should be identified. Secondly, the respondent's position within the interval should be determined.

3.3.3 Find the Appropriate Class Interval

The number of observations $N = 11{,}047{,}560$ and the middle value $N/2 = 5{,}523{,}780$ need to be found (If we use the graphical method, we divide the area under the curve into two.) If we look at the cumulative frequencies;

0–149,999	3,375,707
0–199,999	5,070,808
0–249,999	6,401,720

The nearest class interval to the median is 200,000–249,999.

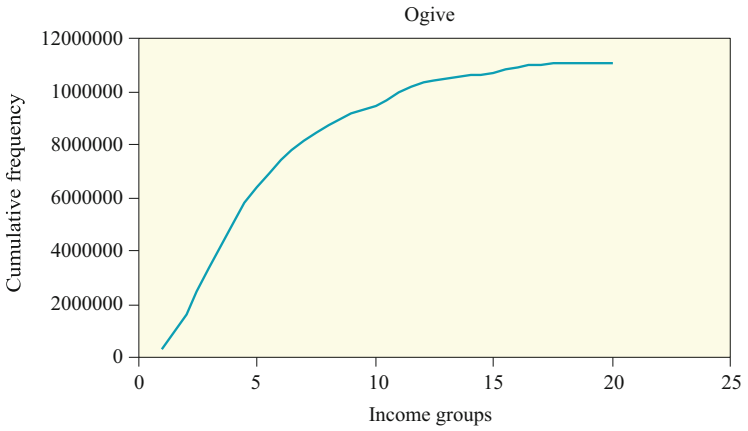


Fig. 3.1 Cumulative frequency Ogive and median

The Graphical Method

Firstly, we plot the cumulative frequency against the upper boundary of the corresponding interval. This graph is called an **Ogive** (Fig. 3.1).

In this case, the median corresponds to the value of the 5523780th observation, which can be read from the Ogive as 216791.6

Median by Calculation

We use the formula below to obtain the exact figure.

$$Median = X_L + (X_u - X_L)\left\{\frac{\frac{N+1}{2} - F}{f}\right\}$$

Where

x_L= the lower limit of the class interval containing the median.
x_u = the upper limit of this class interval
N = the number of observations (using N + 1 rather than N in the formula is only important when N is relatively small)
F = the cumulative frequency of the class intervals up to (but not including) the one containing the median.
f = the frequency for the class interval containing the median.

So the median in our example can be found by substituting these values into the formula:

$$= 200,000 + (249,999 - 200,000)\left\{\frac{5,523,780 - 5,076,808}{1,330,912}\right\}$$

$$= 216,796.6$$

3.4 Mode

The **Mode** is the value or property that occurs most frequently in the data. If there are equal numbers of outcome, there may not be a mode in a distribution. The mode is use-ful if you are interested in the most commonly used value in a distribution. It is easy to calculate as an average, but as it can be very sensitive to changes in a distribution, it is not stable. In the example below if the highest outcome is 4 instead of 3, then the mode becomes 4. Compared to the arithmetic average, its value changes a lot.

3.4.1 A Simple Example with Ungrouped Data

If we throw a dice 10 times, let us say the outcome is:

Number on dice	Outcome
1	1
2	1
3	**3** → 3 is the most frequent outcome and is defined as the mode
4	2
5	2
6	1

3.4.2 Grouped Data

For grouped data the mode can be determined by graph or by calculation.

3.4.3 The Mode by Calculation

The formula is:

Table 3.4 Weekly food expenditures (Euro)

Expenditure on food	f (Number of respondents)	F (Cumulative frequency.)
less than 5 Euro	2	2
>5 but under 10	8	10
>10 but under 15	11	21
>15 but under 20	**15**	**36**
>20 but under 30	12	48
>30 but under 40	6	54
40 Euro or more	3	57

$$Mode = x_L + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} xi$$

where:

x_L = the lower boundary of the modal group
f_M = the (scaled) frequency of the modal group
f_{M-1} = the (scaled) frequency of the pre-modal group
f_{M+1} = the (scaled) frequency of the post modal group
i = the width of the modal group.

Consider the following table of weekly food expenditures in Euros (Table 3.4):

ch3data1.xls

The data is firstly scaled so that the class intervals are equal in size (see Table 3.5)

Applying the formula to the above data provides the modal weekly spending on food:

$$Mode = x_L + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} xi$$

$$Mode = 15 + ((15 - 11)/(2(15) - 11 - 6))x5 = 16.54\, Euro$$

3.4.4 The Mode by Graphical Method

The construction of a histogram for the above data (Table 3.6) is as follows:

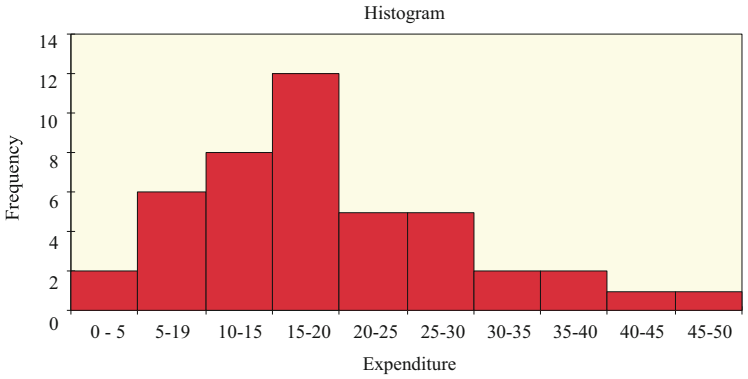
Table 3.5 Adjusted class intervals for weekly food expenditures Euro

Expenditure on food Euro	Frequency f	Mid point x	f*x	Cumulative freq (F)
0–52	2	2.5	5	2
5–10	8	7.5	60	10
10–15	11	12.5	137.5	21
15–20	**15**	**17.5**	**262.5**	**36**
20–25	6	22.5	135	42
25–30	6	27.5	165	48
30–35	3	32.5	97.5	51
35–40	3	37.5	112.5	54
40–45	1.5	42.5	36.75	55.5
45–50	1.5	47.5	71.25	57

Note: Bold typing indicates the modal group

Table 3.6 Frequency density

Income groups TL	Number of households	Total Income TL
0–49,999	321,746	10,674,326
50,000–99,999	1,254,863	99,023,453
100,000–149,999	1,799,098	222,434,428
150,000–199,999	1,695,101	294,254,201
200,000–249,999	1,330,912	295,933,661
250,000–299,999	987,626	268,315,943
300,000–349,999	767,705	247,523,158
350,000–399,999	539,023	200,770,779
400,000–449,999	471,176	199,308,071
450,000–499,999	304,610	143,795,912
500,000–599,999	508,190	275,493,366
600,000–699,999	312,648	200,143,170
700,000–799,999	190,017	141,024,361
800,000–899,999	144,870	122,320,401
900,000–999,999	87,654	82,538,197
1,000,000–1,499,999	186,600	219,570,936
1,500,000–1,999,999	74,114	126,035,614
000,000–4,999,999	62,535	174,767,006
5,000,000–9,999,999	7502	43,815,423
10,000,000–24,999,999	1570	20,681,050
TOTAL:	11,047,560	3,388,423,456



The mode can be estimated from the above histogram. Firstly, identify the tallest block on the histogram and then join the corner points. The point of intersection locates the mode which is 17 Euro in this example.

3.4.5 Another Example

Let us consider our income data once again. The mode is defined as the level of income occurring with the greatest frequency. The level of income means a class interval, corrected for width.

ch2data6.xls

By using the frequency densities (Frequency density = frequency 3/4 class width) it can be seen that the highest frequency density is shown by the third class interval, from 100,000 to 149,999. The formula for the mode can be used after adjusting the class intervals as in the first example above.

3.5 Other Measures of Location

We will briefly talk about two other measures of location. These are the geometric mean and the harmonic mean, both of which have limited uses.

3.5.1 The Geometric Mean

It is defined as '*the nth root of the product of n numbers*'. It is useful when averaging percentages or index numbers.

Example

The following five percentages show the amount of time spent for sleeping within a 24 h period:

$$36\,\%\;\; 32\,\%\;\; 26\,\%\;\; 40\,\%\;\; 28\,\%$$

36 %	32 %	26 %	40 %	28 %

Multiplying these numbers makes: $36 \times 32 \times 26 \times 40 \times 28 = 33,546,240$

The geometric mean can be found by taking the fifth root of 33,546,240 and equals to 32 %, which is less than the arithmetic mean (32.4 %). In this example, the arithmetic mean is an overestimate of the amount of time spent for sleeping in a 24 h period.

3.5.2 Harmonic Mean

It is defined as '*the reciprocal of the arithmetic mean of the reciprocals of the data*'. It is used in the case of ratio data. For example, output per hour or kilometers per liter of fuel.

Example

The following data indicates the amount of overtime (hours) worked by employees at a company during 1 month:

22	30	8	16	25

The reciprocal of each number (1 divided by each number) is:

0.04545	0.03333	0.125	0.0625	0.04

The average of the reciprocals is:

$$(0.04545 + 0.03333 + 0.125\; + 0.0625 + 0.04)/5\; = \mathbf{0.061256}$$

The reciprocal value of this figure is (1/0.061256) = **16.32**

The arithmetic mean is 20.2 h but the harmonic mean is 16.3.

3.6 Comparison

The mean, median and mode provide easy summary measures for numerical information. The mean gives the center of gravity, the median divides the distribution into two and the mode gives the highest point of the distribution. The mode is less reliable, but useful if we are interested in the most commonly occurring value. The median is more stable, but it does not indicate the range of values of the distribution.

3.7 A Review of This chapter

In this chapter, we have examined the main measures of location. We have focused mainly on the mean, the median and the mode. These measures have been examined using some different types of data.

3.7.1 Use the Mode

If you want to know which value occurs most frequently in a distribution.

3.7.2 Use the Median

If you want to divide a distribution in half.

3.7.3 Use the Mean

If you want each entry value in the data to be entered into the average.

3.8 Review Problems for Measures of Location

3.1. (a)'The arithmetic mean is the only average ever required' Write a brief, critical assessment of this statement.
(b)'The median provides a better description of typical earnings than the mean' Discuss

(c)The number of new orders received by a company over the past 25 working days were recorded as follows:

3	0	1	4	4
4	2	5	3	6
4	5	1	4	2
3	0	2	0	5
4	2	3	3	1

Determine the mean, median and mode.

3.2. Use the data in question 2.3 to calculate the mean, median and modal age of cars on the roads in 1990.

3.3. A survey of workers in a particular industrial sector produced the following table:

Weekly income (Euros)	Number (f)	Mid point (X)
0 but under 100	200	50
$100 < 150$	300	125
$150 < 200$	227	175
$200 < 400$	157	300
over 400	114	500

Use the data above, calculate the mean, find the appropriate class intervals for the median and define the modal group.

3.4. The data below refers to income before and after tax. For the data on incomes before tax, obtain the histogram or frequency polygon. Calculate the mean, median and modal income before tax. Which is the biggest and which is the smallest? If you didn't expect this result, then your calculations may be wrong. Which of these charts and statistics can be calculated for after tax income from the data given below? Explain.

Distribution of total incomes before and after tax

Range of total income before tax (lower limit) £	Number of tax payers £	Total income before tax £	Total income after tax £
3445	2.2	9600	9300
5000	3.9	24,500	22,700
7500	3.8	33,500	29,900
10,000	6.2	76,600	66,600
15,000	3.8	65,800	55,800
20,000	3.2	75,800	62,400

(continued)

Range of total income before tax (lower limit) £	Number of tax payers £	Total income before tax £	Total income after tax £
30,000	0.9	29,500	23,300
40,000	0.7	52,200	36,900
All ranges	24.8	367,400	306,800

Source: U.K. Inland Revenue Statistics

Chapter 4
Measures of Dispersion

4.1 Introduction

The last chapter considered several measures in which we can summarize a set of data using just one value. However, as was concluded, these single values do not say very much about the data itself and how it is dispersed (its spread). For instance, it is possible to take two separate sets of data with the same mean, but with great differences in distribution.

As a simple example, let us take the examination grades for two classes of students which have the same average grade but different extremes of attainment.

The Fig. 4.1 below explains, graphically, the two distributions with the same mean, but with different degrees of dispersion.

I. The students‘ examination grades are narrowly dispersed around the mean.
II. The grades are more widely dispersed with a larger number of students achieving lower and higher grades. In other words there are more extremes.

By using statistics, it is possible to make a comment about the dispersion of a set of data. In this chapter, we will discuss some statistical techniques used to measure dispersion, that is the range, quartiles, variance, standard deviation, coefficient of variation and coefficient of skewness.

The online version of this chapter (doi:10.1007/978-3-319-26497-4_4) contains supplementary material, which is available to authorized users.

D. Özdemir, *Applied Statistics for Economics and Business*,
DOI 10.1007/978-3-319-26497-4_4

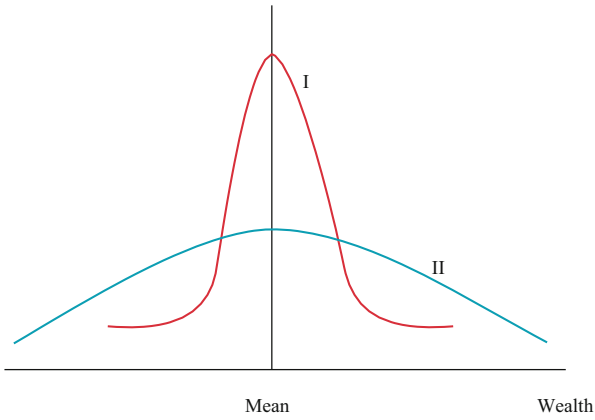


Fig. 4.1 Two distributions with different degrees of dispersion.

4.2 The Range

The range is the simplest measure of dispersion. It is calculated by obtaining the difference between the highest and lowest values in a set of observations. Let us again consider the statistics examination grades from Chap. 2. There are 13 observations (taken as percentages):

$$23, 72, 28, 42, 50, 55, 60, 52, 48, 35, 62, 67, 56$$

The lowest mark is 23 and the highest mark is 72. The range can be calculated by subtracting the lowest value from the highest value:

$$\begin{aligned} \text{Range} &= 72 - 23 \\ \text{Range} &= 49 \end{aligned}$$

This measure is considered an unstable and crude way of measuring the spread of a set of data because any change in one of these values, the highest or the lowest, regardless of the rest of the numbers, will alter the range. If these values are extreme, the change could be significant. Instead, it is sometimes more informative just to quote the lowest and highest values.

4.3 Quartiles

Quartiles are another measure of dispersion. In the last chapter, we showed how to calculate the median which represents a 'half way value'. Similarly, we can obtain 'quarter way values', which are called quartiles. For tabulated discrete or un-tabulated data, we simply count the ordered data set until we reach a quarter

of the way through the data. The value we reach is called the first quartile (Q1). Similarly, the three- quarters value is called the third quartile (Q3).

The method for computing quartiles:

- Rank the data in order of increasing value.
- Find the median. This is the 2nd quartile (Q2).
- The first quartile (Q1) is then the median of the lower half of the data. This is, the data falling below Q2 **but not including Q2**.
- The third quartile (Q3) is the median of the upper half of the data. This is, the data falling above Q2 **and including Q2**.

4.4 The Inter-quartile Range

The inter-quartile range is derived by calculating the difference between the third and the first quartiles (Q3–Q1). It tells us something about how the middle half of the data is dispersed. The higher its value is the more dispersed the data.

As an example, let us consider the weekly food expenditure data, which is similar to the one in Chap. 3 (Table 4.1):

ch4data1.xls

A Graphical Explanation

The first quartile (Q1) will correspond to

$$\frac{n}{4} = \frac{58}{4} = 14.5$$

The second quartile (Q2) will correspond to the median.

The third quartile (Q3) will correspond to $(3n)/4 = (3 \times 58)/4 = 43.5$ rd ordered value.

The determination of these quartiles can also be shown in the following cumulative frequency graph.

Figure 4.2 is a plot of the cumulative frequency against the upper range. The value of Q1 corresponds to the 14.5th on the cumulative frequency axis and is 11.59 The value of Q3 corresponds to 43.5rd on the cumulative frequency axis and is 26.25.

Calculation of the Quartiles

For calculations, we can adopt the median formula from our previous chapter.

Table 4.1 Weekly food expenditures (Euro)

Expenditure on food	f (Number of respondents)	F (Cumulative freq.)
less than 5 Euro	2	2
>5 but under 10	8	10
>10 but under 15	11	21
>15 but under 20	15	36
>20 but under 25	6	42
>25 but under 30	6	48
>30 but under 35	3	51
>35 but under 40	3	54
>40 but under 45	2	56
45 Euro or more	2	58

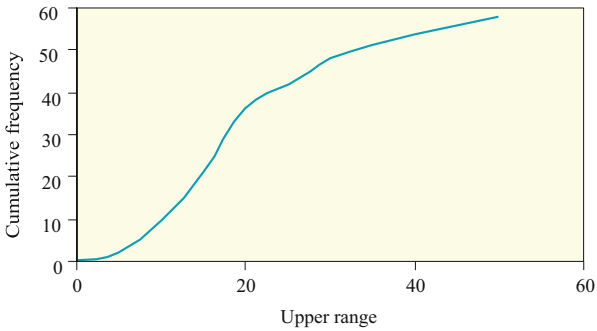


Fig. 4.2 Graphical representation of the quartiles.

$$Q1 = x_L + (x_u - x_L)\left\{\frac{O - F}{f}\right\} = 10 + (15 - 10)\left\{\frac{14.5 - 11}{11}\right\}$$

=11.59 Euro, where O is the order value of interest which is different for different quartiles. i.e. For Q1 it is 14.5 for Q3 it is 43.5

x_L = the lower limit of the class interval containing the quartile.
x_u = the upper limit of this class interval
F = the cumulative frequency of the class intervals up to (but not including) the one containing the quartile.
f = the frequency for the class interval containing the quartile.

$$\text{Q3} = x_L + (x_u - x_L)\left\{\frac{O - F}{f}\right\} = 25 + (30 - 25)\left\{\frac{43.5 - 42}{6}\right\} = 26.25$$

Thus the **interquartile range** is Q3 − Q1 = 26.25 − 11.59 = 14.66. The **quartile deviation** is the average difference: (Q3 − Q1)/2 = 7.33 Euro. In some cases, it could be better to take these two actual quartile values rather than the deviation because it may mislead us. For example, if the majority of the data is towards the

Table 4.2 Income frequencies.

Income groups	(Frequency)	Total Income	
(TL)	Number of households	(TL)	Cumulative frequency
0–49,999	321,746	10,674,326	321,746
50,000–99,999	1,254,863	99,023,453	1,576,609
100,000–149,999	1,799,098	222,434,428	3,375,707
150,000–199,999	1,695,101	294,254,201	5,070,808
200,000–249,999	1,330,912	295,933,661	6,401,720
250,000–299,999	987,626	268,315,943	7,389,346
300,000–349,999	767,705	247,523,158	8,157,051
350,000–399,999	539,023	200,770,779	8,696,074
400,000–449,999	471,176	199,308,071	9,167,250
450,000–499,999	304,610	143,795,912	9,471,860
500,000–599,999	508,190	275,493,366	9,980,050
600,000–699,999	312,648	200,143,170	10,292,698
700,000–799,999	190,017	141,024,361	10,482,715
800,000–899,999	144,870	122,320,401	10,627,585
900,000–999,999	87,654	82,538,197	10,715,239
1,000,000–1,499,999	186,600	219,570,936	10,901,839
1,500,000–1,999,999	74,114	126,035,614	10,975,953
2,000,000–4,999,999	62,535	174,767,006	11,038,488
5,000,000–9,999,999	7502	43,815,423	11,045,990
10,000,000–24,999,999	1570	20,681,050	11,047,560
TOTAL:	11,047,560	3,388,423,456	

lower end of the range, then the third quartile will be considerably further above the median than the first quartile below it is. Averaging these numbers as we did in the quartile deviation can disguise this difference.

Let us take the income data and examine the quartiles;

Obtaining the first quarter of the below data (Table 4.2) is as follows; The sample size is given as 11,047,560

- The first quarter of 11,047,560 is (11,047,560)/4 = 2,761,890.
- The person ranked 2,761,890 is in the 50,000–99,999 class.
- The adoption of the median formula is;

$$Q1 = x_L + (x_u - x_L)\left\{\frac{O - F}{f}\right\}$$

$$= 100000 + (149999 - 100000)\left\{\frac{2761890 - 1576609}{1799098}\right\} = 132940.32$$

Similarly

- three quarters of 11,047,560 is 8,285,670.

– the person ranked 8,285,670 is in the 350,000–399,999 class.
– using the formula provides:

$$Q3 = x_L + (x_u - x_L)\left\{\frac{O - F}{f}\right\}$$
$$= 350000 + (399999 - 350000)\left\{\frac{8285670 - 8157051}{539023}\right\} = 361930.5$$

The interquartile range is Q3 − Q1 = 361,930.5 − 132,940.32 = 228,990.18
If the value gets higher, this means the distribution is more spread out.


ch2data6.xls

4.5 The Variance

The variance is a better measure of dispersion because it makes use of all of the information in the sample. If the distribution is more dispersed, then the variance is larger. It is useful to compare two sets of data to see which of them has the greater degree of dispersion. The formula of variance is different for different data, i.e. a simple or a grouped data. It is denoted as 'sigma squared', σ^2 and has the following formula which is suitable for a simple data:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

where

– x is the observation and μ is the mean hence x_i – μ measures the deviation of x_i from the mean.
– N is the number of observations. Thus the sum of the squared deviations divided by N gives the variance. The above formula needs a slight alteration for a grouped data (Table 4.3):

$$\sigma^2 = \frac{\sum f(x_i - \mu)^2}{\sum f}$$

Let us consider the income data again with the mean deviations.

Fig. 4.3 Zero skewness

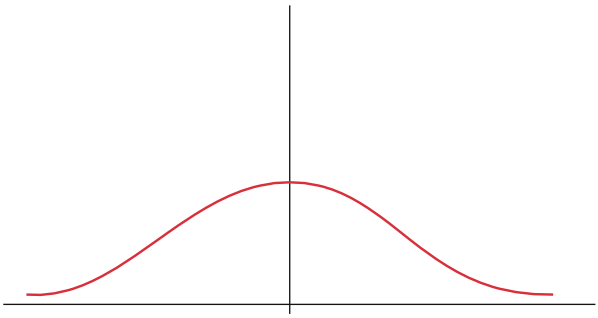


Fig. 4.4 Negative skewness

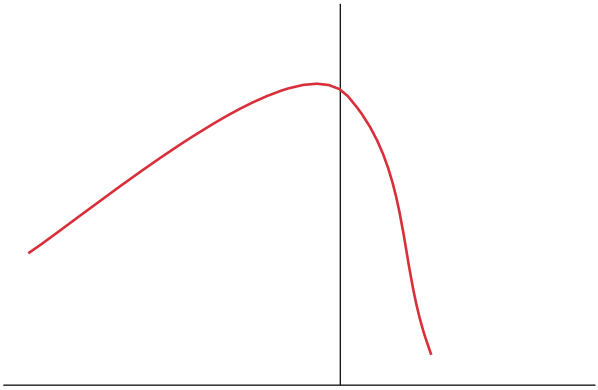
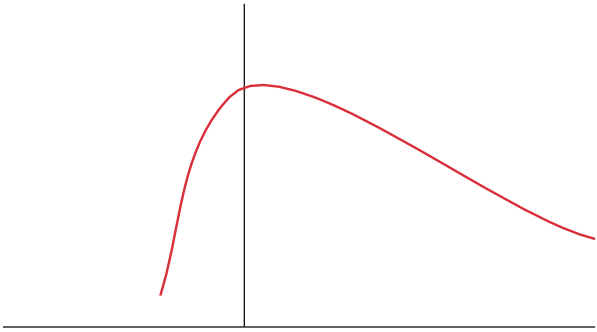


Fig. 4.5 Positive skewness



$$\mu = \text{Sum}(\text{f x})/\text{Sum}(\text{f}) = 314623,22$$

If we substitute the derived values into the variance formula above:

$$\sigma^2 = \frac{2,1967\text{E} + 18}{11047560} = 1,98841\,E + 11$$

Table 4.3 The variance of income

Income groups TL	Mid point (X)	(Freq.) f No. oh H. holds	Cumulative Frequency	f * X	Deviation	(x − mu)^2	f(x − mu) ^2
0–49,999	24,999,5	321,746	321,746	8,043,489,127	−289,623,72	8,3882E + 10	2,6989E + 16
50,000–99,999	74,999,5	1,254,863	1,576,609	94,114,097,569	−239,623,72	5,742E + 10	7,2054E + 16
100,000–149,999	124,999,5	1,799,098	3,375,707	2,24886E + 11	−189,623,72	3,5957E + 10	6,469E + 16
150,000–199,999	174,999,5	1,695,101	5,070,808	2,96642E + 11	−139,623,72	1,9495E + 10	3,3046E + 16
200,000–249,999	224,999,5	1,330,912	6,401,720	2,99455E + 11	−89,623,722	8,032,411,606	1,069E + 16
250,000–299,999	274,999,5	987,626	7,389,346	2,71597E + 11	−39,623,722	1,570,039,372	1,5506E + 15
300,000–349,999	324,999,5	767,705	8,157,051	2,49504E + 11	10,376,278	107,667,138	8,2657E + 13
350,000–399,999	374,999,5	539,023	8,696,074	2,02133E + 11	60,376,278	3,645,294,904	1,9649E + 15
400,000–449,999	424,999,5	471,176	9,167,250	2,0025E + 11	110,376,28	1,2183E + 10	5,7403E + 15
450,000–499,999	474,999,5	304,610	9,471,860	1,4469E + 11	160,376,28	2,5721E + 10	7,8347E + 15
500,000–599,999	524,999,5	508,190	9,980,050	2,66799E + 11	210,376,28	4,4258E + 10	2,2492E + 16
600,000–699,999	649,999,5	312,648	10,292,698	2,03221E + 11	335,376,28	1,1248E + 11	3,5166E + 16
700,000–799,999	749,999,5	190,017	10,482,715	1,42513E + 11	435,376,28	1,8955E + 11	3,6018E + 16
800,000–899,999	849,999,5	144,870	10,627,585	1,23139E + 11	535,376,28	2,8663E + 11	4,1524E + 16

900,000–999,999	949,999,5	87,654	10,715,239	83,271,256,173	635,376,28	4,037E + 11	3,5386E + 16
1,000,000–1,499,999	1,249,999,5	186,600	10,901,839	2,3325E + 11	935,376,28	8,7493E + 11	1,6326E + 17
1,500,000–1,999,999	1,749,999,5	74,114	10,975,953	1,29699E + 11	1,435,376,3	2,0603E + 12	1,527E + 17
2,000,000–4,999,999	3,499,999,5	62,535	11,038,488	2,18872E + 11	3,185,376,3	1,0147E + 13	6,3452E + 17
5,000,000–9,999,999	7,499,999,5	7502	11,045,990	56,264,996,249	7,185,376,3	5,163E + 13	3,8733E + 17
10,000,000–24,999,999	17,499,999,5	1570	11,047,560	27,474,999,215	17,185,376	2,9534E + 14	4,6368E + 17
TOTAL:		11,047,560		3,47582E + 12			2,1967E + 18

4.6 The Standard Deviation

The standard deviation is the most commonly used measure of dispersion because it is directly related to the mean, which is the most commonly used measure of location. It measures the differences from the mean.

Question
In what units is the variance measured? (Could the answer be the squared Turkish Liras !?. Surely not!).

Thus the square root of the variance is the standard deviation. The formula for the standard deviation for:

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

Or for a grouped data: $\sigma = \sqrt{\frac{\sum f(x_i - \mu)^2}{\sum f}}$

Hence the standard deviation of the income data is:

$$\sigma = \sqrt{1,98841\ E+11} = 560,91285$$

It is more difficult to interpret a single standard deviation number and it makes more sense if we have the standard deviation of another set to compare it with.

As we mentioned before, the formula for the standard deviation differs slightly depending on whether the entire population or a sample is used. The formula above is for an entire population.

4.7 The Variance and the Standard Deviation of a Sample

For a sample data, the sample variance is denoted by 'S^2' and the formula is:

$$S^2 = \frac{\Sigma (x_i - \overline{x})^2}{n-1}$$

The standard deviation is:

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

Since we do not know N and $\propto$ in a sample, we use the sample size n and the sample mean instead. However, our interest is the population variance not sample variance and the variation of the sample observations around tends to be smaller than the one around μ. Using $n-1$ rather than n in the formula compensates for this. Then the result is an unbiased (correct on average) estimate of the population variance. If the sample size is smaller, then the use of the correct formula becomes more important.

Example

This is an example for a **simple ungrouped data**. The following 10 observations indicate the number of marks obtained in a tutorial group's assignments in statistics.

40 52 61 39 54 57 59 39 53 42

Thus the mean, $\overline{x}$ is 49.6 mark.

Table 4.4 shows the means deviations, negative and positive

As you may notice, some of the differences are negative while the others are positive. i.e. the observation 40 is 9.6 units below the mean. Hence the sum of these differences is zero which will be avoided by taking the squares of these values for the calculation of variance.

For the calculation of the variance and the standard deviation, the following steps are useful.

(a) Calculate the mean, $\overline{x}$.
(b) Calculate the mean deviations, $(x-\overline{x})$.
(c) Square these differences, $(x-\overline{x})^2$.
(d) Sum the squared differences, $\sum (x-\overline{x})^{2.}$
(e) Average the squared differences to find variance.
(f) Take the square root of the variance to find standard deviation. The steps are shown below in our simple data:

Number of marks obtained (Table 4.5)

Thus the variance is $S^2(684.4)/10 = 68.44$ and the standard deviation is

$$S^2 = \sqrt{68.44} = 8.27$$

Let us take another example for a **grouped data**. The data below (Table 4.6) refers to incomes before and after tax in the U.K.

For a grouped data the formulas for sample variance and the standard deviation are slightly different;

ch3data2.xls

For the variance of a sample: $S^2 = \frac{\Sigma f(x_i-\overline{x})^2}{n-1}$

Table 4.4 Mean deviations

Mean = 49.6	
Observations 39, 39, 40, 42, 52, 53, 54, 57, 59, 61	
−10.6	2.4
−10.6	3.4
−9.6	4.4
−7.6	7.4
9.4	
11.4	

Table 4.5 Calculation of the standard deviation

X	x mean	(x − mean)^2
40	−9.6	92.16
52	2.4	5.76
61	11.4	129.96
39	−10.6	112.36
54	4.4	19.36
57	7.4	54.76
59	9.4	88.36
39	−10.6	112.36
53	3.4	11.56
42	−7.6	57.76
196	49.6	684.4

Table 4.6 Distribution of total incomes before and after tax (British Sterling)

Distribution of total incomes before and after tax.			
Range of total income before tax £	Number of tax payers	Total income before tax	Total income after tax (lower limit) £
	(Millions)	£	£
3.445	2.2	9.600	9.300
5.000	3.9	24.500	22.700
7.500	3.8	33.500	29.900
10.000	6.2	76.600	66.600
15.000	3.8	65.800	55.800
20.000	3.2	75.800	62.400
30.000	0.9	29.500	23.300
40.000	0.7	52.200	36.900
All ranges	24.8	367.400	306.800

Source: The UK Inland Revenue Statistics

For the standard deviation of a sample: $S^2 = \sqrt{\frac{\Sigma f(x_i - \bar{x})^2}{n-1}}$ (Table 4.7)

Table 4.7 Standard deviation

Pre-tax inc.	f- million	Total inc.	Total inc.	Pre-tax/f			
lover limit	Number of	before tax	after tax	X−(income	(x − mean)	(x − mean)^2	f*(x − mean)^2
£	taxpayers	(£ m)	(£ m)	class value)			
3445	2.2	9600	9300	4363.6364	−10,514.91	110,563,251.3	243,239,153
5000	3.9	24,500	22,500	6282.0513	−8596.491	73,899,661.43	288,208,680
7500	3.8	33,500	29,900	8815.7895	−6062.753	36,756,974.38	139,676,503
10,000	6.2	76,600	66,600	12,354.839	−2523.704	6,369,080.872	3,948,8301.4
15,000	3.8	65,800	55,800	17,315.789	2437.247	5,940,172.761	22,572,656.5
20,000	3.2	75,800	62,400	23,687.5	8808.9575	77,597,732.06	248,312,743
30,000	0.9	29,500	23,300	32,777.778	17,899.235	32,038,2623.2	288,344,361
40,000	0.7	52,200	36,900	74,571.429	59,692.886	3,563,240,646	2,494,268,452
							3,764,110,849
130,945	24.7	367,500					
	Mean =	**14,878.543**					

For the variance of the above data (income before tax) for : $S^2 = \frac{\sum f(x_i - \bar{x})^2}{n-1} = \sqrt{\frac{3764110849}{24.7-1} = 158823242.2}$

For the standard deviation of the above data (before income tax): $S = \sqrt{\frac{\sum f(x_i - \bar{x})^2}{n-1}} = 12602.51$

The following alternative formulae for calculating the variance and standard deviation would give the same results as the formulae given above: For the population variance using simple data $: \sigma^2 = \frac{\sum x^2}{N} - \mu^2$,

For the population variance using grouped data $: \sigma^2 = \frac{\sum fx^2}{\sum f} - \mu^2$,

For a sample variance and using simple data:$S^2 = \frac{\sum x^2 - n\bar{x}^2}{n-1}$,

Or for a grouped data: $S^2 = \frac{\sum fx^2 - n\bar{x}^2}{n-1}$,

The standard deviations can be obtained by square rooting these formulae.

4.8 The Coefficient of Variation

We have so far discussed absolute dispersion measures. Their values quite often depend on the units in which the variable is measured. If we measure two variables in different units, then comparing them could be difficult. For example, comparing the income levels of Turkey and the UK, one is measured in Turkish Liras and the other in pounds. A solution to this problem is to measure relative dispersion. The coefficient of variation is a relative measure of dispersion. It is defined as standard deviation divided by the mean.

$$\textbf{COEFFICIENT OF VARIATION} = (\boldsymbol{\sigma}/\boldsymbol{\mu})$$

For the income data, this would be: 560,91,285/314,623,22 = 0,00,178,280,817, which means the standard deviation is 0,0178% of the mean.

4.9 Measuring Skewness

Skewness of the data can be observed from a histogram which has one tail longer than the other hence it is skewed to the left or right. In the income data, the distribution is heavily skewed to the right. The measure of skewness is a numerical measure about the asymmetry of the data and based on the cubed deviations from the mean.

The use of skewness can be examined in a typical income data. For example, the gross income and after tax income would likely to have different coefficients of skewness. The coefficient of skewness after tax income would be reduced. You can examine this with the data above for pre-tax and post-tax income.

$$\textbf{THE COEFFICIENT OF SKEWNESS} = \frac{\Sigma f(x-\mu)^3}{N\sigma^3}$$

In the Turkish income data this is equal to: $= \frac{1.\ 31806E+25}{(11047560)\ *11560,\ 912853} = 6760561097$ positive coefficient of skewness indicates that it is skewed to the right.

HE Coefficient of Skewness (Pre-Tax UK Data)

$$= \frac{\sum f(x-\mu)^3}{N\sigma^3} = (1.5031\text{E}+14)(2.002\text{E}+12*24.7) \qquad = \mathbf{3.0403625}$$

It is positive and the data is skewed to the right.

Similarly, the coefficient of skewness of the data presented in the Table 4.6 can be calculated (Table 4.8).

4.10 A Review of This Chapter

A single value such as an average obtained from data can be misleading. Hence in this chapter, we have examined the measure of dispersion. We have looked into the range, quartiles, interquartile range, variance, standard deviation and skewness. For our purpose, we have used some real data as well as some simplified forms of data.

4.10.1 Review Problems for Measures of Dispersion

4.1. Use the data on the age distribution of the cars in 1990 given in question 2.3 to find the standard deviation and quartile deviation of car ages.

4.2. Use the data in question 3.3 to calculate both for before tax income and for after tax income, the standard deviation and the coefficient of variation. Comment.

4.3. The data below are taken from TUIK's earnings survey. It provides normal hours of work per week earnings by broad age group. Find the standard deviation and quartile deviation of broad age group.

Table 4.8 Calculations for the coefficient of skewness of pre-tax income

Pre-tax inc.	f- million	Total inc.	Total inc.	Pre-tax/f			
lover limit	Number of	before tax	after tax	X−(income	(x − mean)	(x − Mean)^3	f*(x − mean)^3
£	taxpayers	(£ m)	(£ m)	class value)			
3445	2.2	9600	9300	4363.6364	−10,514.91	−1.1626E + 12	−2.558E + 12
5000	3.9	24,500	22,500	6282.0513	−8596.491	−6.3528E + 11	−2.478E + 12
7500	3.8	33,500	29,900	8815.7895	−6062.753	−2.2285E + 11	−8.468E + 11
10,000	6.2	76,600	66,600	12,354.839	−2523.704	−1.6074E + 10	−9.966E + 10
15,000	3.8	65,800	55,800	17,315.789	2437.247	14,477,668,026	5.5015E + 10
20,000	3.2	75,800	62,400	23,687.5	8808.9575	6.83555E + 11	2.1874E + 12
30,000	0.9	29,500	23,300	32,777.778	17,899.235	5.7346E + 12	5.1611E + 12
40,000	0.7	52,200	36,900	74,571.429	59,692.886	2.127E + 14	1.4889E + 14
							1.5031E + 14
130,945	24.7	367,500					

Age	Earnings per Employee	Percentage of Employees (%)
12–14	44,894	0.3
15–19	55,789	5.3
20–24	86,393	9.6
25–34	152,698	39
35–54	222,725	44
55+	193,125	0.8
Unknown	128,331	1
Total		100

ch4data2.xls

4.10.2 Computing Practical For Calculation of Mean, Median, Mode and Standard Deviation

C.4.1.

Open a new Excel worksheet, type your name and a title at the top of the spreadsheet: Measure of Location and Dispersion in Excel. Use the data from Problem Sheet 2.3. Type upper age limits in cells A4:A5 and enter upper age limits from 0 to 16 in cells A6:A14 (e.g. similar to the table below)

(a) Complete the rest of the table in Excel and calculate the mean, median and modal age of cars on the road in 1990;

Use formula: Type '=(A6 + A7)/2' in Cell B7 and hit return. You have entered a formula for the average of these two values which should appear as the result 1 on cell B7. Now click back on B7 and point the fill handle at the lower right hand corner of cell B7. The cross should become darker, then press the left button of the mouse and drag it down to B14. This exercise should provide all mid point values in column B. Now select the cells C6:C14 and click on the Σ icon under the FORMULAS tool bar. Is the sample size 20,229? If not, then you have made an error. Check your steps. Now complete the columns D, E, F and G by entering the appropriate formulas [f*X, X- mean, F and f (X-mean)^2] similar to the exercise in the column B. (See also Hint 22 at the end of this sheet)

(b) Use the appropriate formula given in the chapter to find the standard deviation and quartile deviation of car ages.

Upper Age limit	Mid point (X)	Frequency (f) No. of Cars	f * X	(X-mean)	Cumulative Frequency (F)	f(X-mean) ^2
0		0				
2		4127				
4		3958				
6		3425				
8		3152				
10		2359				
12		1648				
14		755				
16		805				

ch2data5.xls

C.4.2.

Enter the data from the problem sheet 2.1(c). Be sure the values in your data set are entered in a single column on the worksheet, with a label (Orders or X) in the cell just above the first value. Construct a similar table (the above question 2.1). Enter in an empty column of the spreadsheet the mean deviations of the orders (e.g. X –) and $(X -))^2$ values. Use the formula $\sqrt{\left(\frac{\Sigma\left(X_i-\overline{X}\right)}{n}\right)}$ for standard deviation.

(a) Calculate the mean, median and mode.
(b) Determine the range, quartile deviation and standard deviation.

C.4.3.

Excel includes the FORMULAS Tool which provides measures of tendency, variability, and skewness. It is suitable for data without any time dimension, such as the data given in the question 2.1 (c) of the Problem Sheet 2. Now copy the same Orders (X) column into another column and make sure that the right hand columns are free for the Output range.

Now, From the **FORMULAS** menu, choose **More Functions** and **Statistical**. There are a large number of statistical tools available there.

Have you obtained the same mean and the standard deviation as you calculated previously? What are the values for the mean, median, mode, standard deviation and the skewness of the distribution? Save your results in an Excel file.

Print out the answers for these three questions and hand them in to your tutor in your Statistics tutorial.

Hints

1. Position the cell pointer first before entering data or beginning a command. Use the mouse to click on a cell. Often you need to select a range of cells. To do this click on the first one, hold the mouse button and drag to the end of the range, then release the button.
2. Excel uses a sophisticated system to distinguish what you enter as being either text or values. Only values can be used in calculations.
3. Enter one complete item per cell. If it doesn't seem to fit, you will find you can adjust the cell size.
4. (Optional) To enter your data in a formatted table,

   a) Select the row where you wish to enter column titles by positioning the mouse pointer on the appropriate row number. Notice that it should then be the shape of an outlined + sign, and click.
   b) Type your column headings pressing the Enter key at the end of each entry. Don't worry at this stage if some of them don't appear to fit in the cells.
   c) With the row still highlighted, click the Auto Format button. This will choose appropriate column widths for your titles, centre them in the cells and display them in Italics. If you don't want Italics, click the italic button to remove.
   d) Select successive rows in turn to enter data.

5. Do not leave blank rows, as this can lead to difficulties with some spreadsheet commands.
6. If hash signs, ####, appear when you enter a number, the column is not wide enough to display the value.
7. To widen a column, in the row of column letters position the mouse pointer on the right hand boundary of the column you want to widen, and watch for it to change shape. Then click and drag: you will see the column boundary move. Release the mouse button when you are satisfied with the width.

Correcting Mistakes

8. Typing mistakes can be corrected with the Backspace key.
9. If you wish to change something already entered in the spreadsheet, simply re-enter the correct item.
10. You can blank a selected cell or range of cells by pressing the delete key. The clear Menu will appear (From Edit). Select All if you want to remove what is entered and also the way it is formatted, and click OK.
11. Used a command you wish you hadn't ? Immediately use Edit Undo.
12. Power cuts and Network Failures

    Save your work regularly, using the File menu (click on it to open) or the Save file button which has a disk on it. You will have to supply a name the first time you save a spread sheet. Next time you may be happy to overwrite your first version. However, it is good practice to sometimes save operate versions of spreadsheets separately by using File, Save As.

Displaying Text and Numbers

13. You can alter the way text is played by selecting it with the mouse, then clicking the appropriate button in the toolbar.
14. To alter the way numbers are displayed select those you want to alter in the same way then click on Format and Number. Select the category and then the Format Code you wish and click OK.
15. If you want to put the data into a certain order use Data and then Sort Command. Select the data to be sorted and then point and click to use the command.

Positioning the Window

16. For maximum spreadsheet display area, click the maximum button at the top right of the spreadsheet.
17. Scroll bars hidden? Click on the spreadsheet title bar, and drag.

Charts

18. For simple charts, use the Recommended Charts button,
19. To alter the chart after it is completed double click on it. Notice that you are now offered a different options to make alterations on your chart.
20. To graph separated columns. To select these, select the first, then hold CTRL while selecting others.

Formulae

21 a) Start a formula by typing an = sign.
   b) Formulae must be typed without any spaces.
   c) It is important to use cell references in formulae so that they can be copied and so that they will automatically recalculate if changes are made.
   d) Cell references can be obtained by pointing at the appropriate cell.
   e) The arithmetic symbols used in formulae are: +, −, * (multiply), / (divide), ^ (exponentiation). Brackets may be used. The order of precedence of operations is the normal algebraic order.
   f) When formulae are to be copied it is important to distinguish between relative and absolute references. A relative address is considered relative to the cell in which the formula containing it has been entered. It is the usual, or default, form of address. If it is copied to other rows its row number will change, and similarly the column letter will change if it is copied to other columns. An absolute reference has its column letter and row number preceded by $ signs. It does not change when copied. When you are pointing at a cell reference, it can be made absolute by pressing the function key F4.
   g) When a value is to be used in several calculations, and that value might be changed, it should be entered in a single cell, e.g. the percentage change. If you do it correctly.

      i. You can see what value has been used

ii. The value displayed is used in the formulae, being referenced by its cell reference.
iii. When a new value is typed in to the cell in question, all the formula which contain a cell reference to this cell immediately recalculate.
iv. To get this to work properly you need to understand the difference between relative and absolute addresses.

h) To make formulae easier to read, you can name cells and use these names in formulae instead of cell addresses.
e.g. Suppose you have entered 5% in cell and the name Change in the cell above. Select these cells and choose Formula Create Names, Top Row, OK. This will give the value 5% the name Change, which you can type in to a formula. Names are assumed to be absolute addresses, which is often convenient when copying.

i) A formula entered to perform a calculation, say, for one type of formula may be copied to perform for other data's calculation. This is best done by first positioning the pointer on the formula to be copied, then pointing to the fill handle at the lower right hand corner, clicking and dragging to cover the cells you wish to fill.

j) The easiest way of summing a column (row) of numbers is to select the cell below (to the right of) the numbers then click the Auto sum button, S, A formula will appear in the Formula Bar. If you are happy with it, click to accept it.

k) To average set of numbers you can use a function. Select the cell where the answer is to be displayed, then choose Formula, Paste Function, Statistical and Average. Select the range you want to average, delete in the formula offered number 2, ... and enter the formula. Excel offers a very wide variety of built in functions.
To do arithmetic, you either type arithmetic signs between the values or cell addresses, OR you use a function—NOT BOTH! e.g. If you use the Average function, the list or range of values you wish to average should be in the parentheses. Values in a list are separated by commas, ends of a range by a full stop (.) or colon (:). You do not need arithmetic signs within the parentheses of a function.

i. If your formulae are correct, it should be possible to enter a new value for one of the items used in their calculation and watch the formulae recalculate. Try it and see!

Layout

22. If you move part of your table about so you can print it, you need to ensure you give identifiers to the rows that you move. One way would simply to copy the original identifiers.
Aim to present information in a form which would be useful to other people.

Printing

23. If you click the Print button, Excel will print your entire spreadsheet. It is usually best to plan first what you want to print. If your spreadsheet will not all fit on one page, you should decide how it should be split.

    The method is to select the cells you want to print then choose Options, Set Print Area. You can preview before printing with the File, Print command will provide a preview. If you are satisfied, choose Print, otherwise choose Close and proceed to make the changes you wish.

Closing

24. To Close a spreadsheet, click on the close button.

Exiting

25. To exit Excel, click on the File then Close button.

Chapter 5
Index Numbers

5.1 Introduction

In this chapter we will continue the theme of data description (data presentation, using indices). In our last chapter when we discussed skewness we mentioned the difficulty of using two different measures. A similar difficulty may appear in the case of changes in economic, business and social variables over time. Index numbers can solve this problem by providing a simple summary of change by aggregating the information available and making a comparison to a base figure. This base figure could be an arbitrary year or the starting figure. Index numbers are not concerned with absolute values but rather the movement of values, i.e. they allow us to distinguish between nominal and real values. This is another way of summarizing information. Examples of index numbers are the RPI (Retail Price Index), the exchange rate, the poverty line and real income. We can have indexes for prices, quantities, expenditures, etc.

5.2 A Simple Interpretation of an Index Number

Consider an example of the number of crimes committed in a particular area:

Year	Number of crimes
1	127
2	142
3	116

We could calculate the percentage increase from one year to the next: from year 1 to year 2 there is an increase of 11.8 % and from year 2 to year 3 there is a decrease of 18.3 %. This is an accurate description, but rather messy and does not directly

D. Özdemir, *Applied Statistics for Economics and Business*,
DOI 10.1007/978-3-319-26497-4_5

Table 5.1 Index numbers for base year 1

Year	No of Crimes	Calculation	Index (Year 1 = 100)
1	127	$\frac{127}{127} \times 100$	100
2	142	$\frac{142}{127} \times 100$	112
3	116	$\frac{116}{127} \times 100$	91

compare years 1 and 3. We need a numerical measure for monitoring the change in crime. This is called an **Index number**.

An index is a number which compares the values of a variable at any time with its value at another fixed time (based period).

$$\textbf{Index for period} = [(\textbf{value in period})/(\textbf{value in base period})] \times (\textbf{100}).$$

(It is multiplied by 100 to set the base value as equal to 100)

Consider the Table 5.1: column 4 has the index numbers

The index numbers in column 4 of the Table 5.1 allow a direct comparison with the first period; year 1 to year 2 there is a 12 % increase (i.e. (12/100) and year 1 to year 3 there is a 9 % decrease (i.e. (9/100). However, the choice of year one is arbitrary. Similarly, we could choose year 3 as a base year.

Table 5.2 takes the year 3 as a base year rather than year one as in the first table. We can see that from year 1 to year 3 there is a 9 % decrease and from year 2 to year 3 there is a 22 % decrease. Note that,

- Relative values are preserved.
- The choice of base year depends on purpose.
- We could use any value as base value, e.g. 1 or 1000 but it is conventional to use 100.

Consider another example

We can use the data on unemployment flows to calculate a simple index number (Table 5.3).

Beware of the definition changes again. The Index doesn't add information (it actually hides some), but it makes things clearer and easier to compare.

The Reference year is the one for which the value of index = 100. Changing the reference year does not matter: informational content is the same and this is only for convenience's sake.

5.3 A Simple Price Index

Price indices are a very common application of index numbers. Let us take an example of an item. Prices of this item change monthly as you can see in Table 5.4.

Clearly, a change in the index from the base year to the current year is a percentage change.

Table 5.2 Index numbers for base year 2

Year	No of Crimes	Calculation	Index (Year 3 = 100)
1	127	$\frac{127}{116} \times 100$	100
2	142	$\frac{142}{116} \times 100$	122
3	116	$\frac{116}{116} \times 100$	100

Table 5.3 Unemployment and index number

	Unemployment (000)	Index 1	Index 2
1982	2770	100.0	91.4
1983	2984	107.7	98.5
1984	3030	109.4	100.0
1985	3179	114.8	104.9

Table 5.4 Price indices

	Price	Calculation	Index
January	20	(20/20) × 100	100
February	22	(22/20) × 100	110
March	25	(25/20) × 100	125
April	26	(26/20) × 100	130
May	27	(27/20) × 100	135

5.4 Changing the Base Year

We have seen that the base period is arbitrary: it is usually chosen for convenience. Most data sources (official statistical tables) update the base period from time to time because of changing circumstances or the index becoming too large.

- Changing circumstances are important because it sometimes means we can no longer compare them with earlier periods.
- If the index becomes too large, change will become less intuitive. e.g. (11,000/10,000) implies a 10 % change but it is easier to see the change from 100 to 110.

When you collect time-series data, you often find that each volume of your data source only contains observations for a limited number of years. We may need to use several volumes (different issues) and also the base period often varies between volumes. We need to know how to construct an index in order to chain it up with other indices.

5.5 To Chain Indices

We need to know how to construct an index to chain it up with other indices. Let us take an example and say we have two indices: Index 1 and Index 2 as shown in the Table 5.5 below.

Table 5.5 Chaining index 1 to index 2

Period	Index 1	Index 2
1	X1	
2	X2	
3	X3	Y3
4		Y4
5		Y5

Table 5.6 A numerical example of chaining two indices

Year	Index 1	Index 2	Calculation	Chain index 1	Calculation	Chain index 2
1	100			100	100.(100/220)	45.45
2	138			138	138.(100/220)	62.73
3	162			162	162.(100/220)	73.64
4	196			196	196.(100/220)	89.09
5	220	100		220		100
6		125	125.(220/100)	275		125
7		140	140.(220/100)	308		140
8		165	165.(220/100)	363		165

Remember that the indices preserve % change between periods. Now, let us say I want to use index X and need to find the corresponding value for periods 4 and 5. i.e. X4 and X5 which are unknown. We need to connect Y into the X index. If the relative values are preserved, then the case must be that: (X4/X3) = (Y4/Y3) and (X5/X3) = (Y5/Y3). From the above equations,

$$X4 = Y4^{*}\frac{X3}{Y3} \text{and } X5 = Y5^{*}\frac{X3}{Y3}$$

Thus we have a constant multiplicative factor (X3/Y3).

- If $X3 < Y3 \Rightarrow X4 < Y4$ and vice versa.

As this example shows, you need at least one overlapping observation: If not, then you cannot find the multiplicative factor. Table 5.6 below provides a numerical example of chaining indices.

5.6 Indices of Real Variables Versus Nominal Variables

We are often interested in a real rather than a nominal variable, such as a real wage. By convention wages are not set in real terms but set in nominal terms. Hence when we obtain data on wages, we have to work out ourselves what the real wage is. A text book definition of the real wage is (W/P). What is P? If you only consume one good, e.g. In a Robinson Crusoe economy with coconuts being the only good, then Pc implies a real wage in terms of coconuts. P then has a definite meaning.

Table 5.7 Hourly after-tax wage and CPI

Year	W	CPI-1	CPI-2	CPI	(W/CPI)	w (index)
1	4.20	150		84.3	0.0498	100.0
2	4.50	165		92.7	0.0485	97.4
3	4.80	178	100	100	0.0480	96.4
4	5.25		104	104	0.0505	101.4
5	5.50		107	107	0.0514	103.2

Usually in real economies, we use a composite price index for P, for example CPI. In this case (W/P) has no direct meaning (i.e. it is not a measure of a quantity). P is usually represented as an index. An example of this is presented in the Table 5.7.

Sometimes we are not interested in changes in a single variable, but in a combination of different variables. The CPI - aggregate price- can be given as an example.

There are two rather simple indices of this nature.

(1) Simple aggregate index $\Rightarrow \frac{\Sigma P_t}{\Sigma P_0} 100$

(2) Mean price relative index $\Rightarrow \frac{\Sigma\left(\frac{P_t}{P_0}\right)}{n} 100$

where 'price relative' $= \frac{\Sigma P_t}{\Sigma P_0}$

A simple example:

T	Bread	Milk	Tea
0	550	280	720
1	620	320	740

For the simple aggregate price index we sum the prices:

$$\Sigma P_0 = 550 + 280 + 720 = 1550$$
$$\Sigma P_t = 620 + 320 + 740 = 1680$$

Hence the simple aggregate index is: $I\mathrm{o} = 100$

Hence **the increase in prices is 8.4 %**

$$I_0 = \frac{\Sigma P_1}{\Sigma P_0} 100 = \frac{1680}{1550} 100 = 108.4$$

The mean price relative index can be calculated as follows:

Bread $= (620/550) = 1.127$

Milk $= (320/280) = 1.143$

Tea $= (740/720) = 1.028$

$$\text{Io} = 100 \quad \text{The Formula} : \frac{1}{n}\sum\left(\frac{Pt}{P0}\right)100$$
$$I1 = \left(\frac{1,127 + 1,143 + 1.028}{3}\right)100 = 109.9$$

Thus the mean relative price index calculations indicate a 9.9 % change in the price. It is easy to define and use but there are some problems, such as:

- Average Price Index (API) depends on the units used. (i.e. price per what?).
- Both indices do not consider the relative importance of each good, (i.e. gold vs bread).

Thus a good index should take account of both: a) Quantities consumed, b) price of the goods.

5.7 Weighted Indices

Weighted indices take account of the relative importance of each variable. For example you may want to look at a household's weekly shopping basket and compare cost over time. The total cost is: TC = f(Px Q) That is $\Sigma P_i Q_i$ for each could define a weighted index.

This, basically, can be defined as the current cost of the shopping basket divided by its cost in a base year and can be expressed with the following formula:

$$\frac{\Sigma P_{it} Q_{it}}{\Sigma P_{i0} Q_{i0}}$$

there is a simple problem with this in that it will indicate both price and quantity change. We therefore, need to keep the quantity constant if we want to see the change in prices. There are two widely used indices used for this purpose which are the base- weighted index and the current-weighted index.

5.7.1 Base Weighted Index (Laspeyres Price Index)

Assume that quantity does not change. The base year cost of $\Sigma P_0 Q_0$ the basket is $\Sigma P_t Q_0$ equal to and the current cost of the same basket is. The Laspeyres index formula is:

$$\textbf{Laspeyres} = \frac{\Sigma P_t Q_0}{\Sigma P_0 Q_0} \cdot 100$$

Table 5.8 A simplified example of a student price index

	1999	2000	2001	Quantities per month
	P0	P1	P2	Q0
Books	10.00	11.00	12.00	2 books
Food	1.50	1.55	1.70	50 refectory meals
Drink	1.00	1.05	1.05	20 cans of coke
Travel	0.80	0.70	0.75	25 trips to campus

Table 5.9 Calculations of the Laspeyres index

	P0 * Q0	P1 * Q0	P2 * Q0
Books	20.00	22.00	24.00
Food	75.00	77.50	85.00
Drink	20.00	21.00	21.00
Travel	20.00	17.50	18.75
Totals	135.00	138.00	148.75
Index	100.00	102.22	110.19

The **positive** side of this index is that it gives direct cost comparisons of original bundles of goods.

The **negative** side of this index is that it assumes the composition of the bundle has not changed (eg. $\Delta P \rightarrow \Delta Q$ etc.). No substitution takes place. Even if goods become relatively more expensive it assumes that the same quantities are bought. Hence, for example, the index tends to overshoot inflation.

Example Let us take a simplified example and say, during the term time, students consume four goods: books, food, drink and travel. The basic data are as follows (Table 5.8):

If the price of books increases by 20 % (over two years) and food increases by 13.2 %, etc. What is the price increase? Firstly start with constructing (2000) 'shopping basket'.

Constructing the Laspeyres index (Table 5.9):

Interpretation: a 10.19 % income increase is needed between 1999 and 2001 to maintain living standards.

Hence we see that 'prices' for students rose by 2.22 % in 2000, and by a further 7.8 % (110.19/102.22) in 2001, making a 10.19 % increase over 2 years. Hence a 10.19 % grant increase is needed over 2 years to maintain a constant standard of living.

5.7.2 Current-Weighted Index (Paasche Price Index)

Why do we use the 1999 basket? By this, we are using base year weights with 1999 as a base year. Paasche uses current year weights. It assumes that the current shopping basket was used in the base period. The Paasche index shows the lower

rates of price inflation because students switch their purchases to goods whose relative price is falling. This assumes tastes don't change. The cost of today's basket is equal to ($\Sigma P_t Q_t / \Sigma P_0 Q_t$) (today's bundle divided by the base period). Thus the Paasche index formula is:

$$\textbf{Paasche} = \frac{\Sigma P_t Q_t}{\Sigma P_0 Q_t}.100$$

The **Advantage** of the Paasche index is that it gives an accurate picture of change in the cost of current purchase.

The **disadvantage** of this index is that it changes the base of the calculation each period. (It does not give a direct comparison of price over time. i.e. assume that the substitution of goods have become relatively cheaper). It also takes more time to calculate, because the weight has to be updated in each period. (Hence this index is less widely used compared to the other indices).

An example of Paasche index using current year weights (Table 5.10):

Hence the calculations (from the Table 5.11):

$$\frac{\Sigma P_0 Q_1}{\Sigma P_0 Q_1}.100 = \frac{136.50}{136.50} = 100.00$$

$$\frac{\Sigma P_1 Q_1}{\Sigma P_0 Q_1}.100 = \frac{138.50}{136.50} = 101.28$$

$$\frac{\Sigma P_2 Q_2}{\Sigma P_0 Q_2}.100 = \frac{125.00}{115.00} = 108.70$$

Interpretation: an 8.70 % increase is required between 1999 and 2001 to maintain living standards.

Table 5.10 Quantities for Paasche index

Quantities	Q1	Q2
Books	1	1
Food	55	40
Drink	20	25
Travel	30	25

Table 5.11 Paasche calculations

	P0 * Q1	P1 * Q1	P0 * Q2	P2 * Q2
Books	10.00	11.00	10.00	12.00
Food	82.50	85.25	60.00	68.00
Drink	20.00	21.00	25.00	26.25
Travel	24.00	21.00	20.00	18.75
Totals	136.50	138.25	115.00	125.00
Index	100.00	101.28	100.00	108.70

Table 5.12 Real income

	Income	Index (Paasche)	Price index	Real Income	Index of Real Income
1999	135.00	100.00	100.00	135.00	100.00
2000	138.00	102.41	101.28	136.50	101.11
2001	125.00	92.59	108.70	115.00	85.19

5.8 Using a Price Index to Deflate

Either price index can be used to deflate expenditure (or income). Start with expenditure in current prices or in nominal terms. Deflating by a price index (Paasche price index here) gives expenditure in constant prices or in real terms.

Deflating by a general price index gives a series in real terms. If a specific price index is used (e.g. health services), then we get a series in volume terms. That is what we have here since we have used a specific student price index.

5.8.1 Deflating Students' Income

Assume (optimistically?) expenditure = income, so the income each year is

1999	135.00 ($=\Sigma P_{0*}Q_0$)
2000	138.25 ($=\Sigma P_{1*}Q_1$)
2001	125.00 ($=\Sigma P_{2*}Q_2$)

Then we have (Table 5.12):

The 'real' income (or expenditure at constant 1999 prices) measures how well off students are in each of the three years.

5.9 Quantity Indices

'Real' income is really an index of the quantities of goods that students buy. It can also be obtained as follows (Tables 5.13 and 5.14).

Hence

Thus the roles of prices and quantities are reversed in the calculation of the quantity index. The Paasche quantity index is derived using current prices as weights.

Table 5.13 Quantities and initial price

	Qo	Q1	Q2	P0
Books	2	1	1	10.00
Food	50	55	40	1.50
Drink	20	20	25	1.00
Travel	25	30	25	0.80

Table 5.14 Laspeyres quantity index calculations

	P0*Q0	P0*Q1	P0*Q2
Books	20.00	10.00	10.00
Food	75.00	82.50	60.00
Drink	20.00	20.00	25.00
Travel	20.00	24.00	20.00
Totals	135.00	136.50	115.00
Index	100.00	101.11	85.19

Table 5.15 Data series in index numbers

	1999	2000	2001
Books	100	110	120
Food	100	103.3	113.3
Drink	100	105	105
Travel	100	87.5	93.75

5.10 Using Expenditure Shares

Sometimes individual data series are already in index number terms.

Example We can't then use quantities as weights. We have to use expenditure shares. We will now do the Laspeyres as an example (Table 5.15):

1999 expenditure shares:

Books	Food	Drink	Travel
14.8	55.6	14.8	14.8

(e.g. 55.6 = 75/135 * 100)

2000 Laspeyres price index

$$\frac{110}{100}*14.8+\frac{103.3}{100}*55.6+\frac{105}{100}*14.8+\frac{87.5}{100}*14.8=102.22$$

as given before. It is similar to the 2001 index.

Formula of Laspeyres:

$$P_1^n=\sum\left(\frac{Pn}{P0}\right)S$$

where So is the expenditure share on good I.

$$= \sum \frac{PioQio}{PoQo}$$

Formula for Paasche:

$$P_p^n = \left(1/\Sigma\left(\frac{Po}{Pn}\right)Sn\right).100$$

Quantity indices are the same as the ones for price indexes, but this time the roles are reversed.

5.11 The Price Indices in Practice

5.11.1 The Retail Price Index

The Retail Price Index covers a range of goods and services bought by a typical household. The RPI represents the changing cost of a large 'basket' of goods and services reflecting the full range of things that people buy.

Retail Price Statistics are published in Turkey annually. These statistics take annual average prices of selected items. They cover annual average retail prices of selected items covered by the consumer price indices including food, beverages and tobacco, clothing and footwear, homes, furniture and furnishings, health, transportation, entertainment and culture, education, hotels, cafes and restaurants and miscellaneous goods and services expenditure.

5.11.2 Consumer Price Index (CPI)

The latest price indices in Turkey take the 1994 as a base year (ie 1994 = 100). The CPI is published in monthly figures by the State Institute of Statistics. Table 5.16 is taken from the SIS source dated 3rd September 2000. The index in August 2000 is about 3024,4. This clearly shows the high inflation ratio since 1994 in Turkey. For example, the consumer price index for the urban places is calculated using the Laspeyres formula which is based upon a chosen basket of commodities consumed in a given time period. Besides the weights, the average commodity prices of commodities are determined. For this the current year prices of commodities included in the matrix are multiplied by the pre-determined weights and divided into base year prices.[1]

[1] See the urban places consumer price index concepts, methods and sources by SIS for further details.

Table 5.16 The weights used for the CPI for urban Places

Categories (Major groups)	Weights
Food, beverages and tobacco	31.09
Clothing and footwear	9.71
Housing, water, electricity, gas and other fuels.	25.80
Health	2.76
Transport	9.30
Leisure, entertainment and culture	2.95
Education	1.59
Hotels, cafes and restaurants	3.07
Miscellaneous goods and services	4.38
Total	100.00

Weights from the Household income and Consumption Expenditure Survey are calculated annually by a chain linked Laspeyres index (Laspeyres within years, Paasche between years). For every good, a number of prices is obtained to get some idea of average. The table below shows the weights of major groups used for the 1994 = 100. Consumer Price Index for urban places.

The weights may change from year to year. These weights will only reflect the expenditure of a proportion of households accurately. Some households, obviously spent more on some items than the weights indicate but the above weights describe most of the households reasonably well. For example, if the 'price' of food, beverages and tobacco increased by 10 % in 1990, the overall impact would be 3.109 %. (a weight of 31.09 out of 100).

5.11.3 Wholesale Price Index (WPI)

Let us look at the Wholesale Price Index 3rd August 2000. Again the base year is taken as 1994 = 100. A rise in general indices was realized in 1994 = 100 based Wholesale Prices Index on a month-to-month base by 0,9 %, on December of the previous year base by 20,9 %, on the same month of the previous year base by 48,9 % and on an average of twelve months base by 58,8 %. These figures show that the CPI increase is greater than the WPI increase. This may indicate that the higher inflation is not due to an increase in the cost of production. In August 2000, 0,9 % rise in general prices is formed by 0,2 points from Public Sector and 0,7 s point from Private Sector. The result alerted the government which led the government sources to blame the private sector for the persisting higher inflation later on that year (Tables 5.17 and 5.18).

Table 5.17 Index numbers and rate of change of the CPI for urban settlement (%), 1994 = 100

	Year	Jan.	Feb.	Mar.	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
	1994	64,5	68,7	72,7	89,7	97,1	99,0	102,1	105,0	111,5	119,6	127,2	136,1
	1995	145,7	152,8	159,7	169,0	174,6	179,1	184,6	192,6	207,4	220,5	230,9	239,6
	1996	259,5	271,2	286,4	305,6	319,4	327,5	334,5	350,4	371,9	396,0	416,5	430,7
Index	1997	456,0	481,8	507,8	541,4	566,8	583,1	619,6	658,0	706,1	764,9	815,6	857,5
	1998	919,4	960,0	1001,3	1048,0	1084,7	1111,1	1148,4	1193,8	1274,0	1351,1	1409,1	1455,4
	1999	1525,3	1573,7	1637,5	1717,2	1767	1825,2	1894,9	1974,6	2092,8	2225,2	2318,7	2456,6
	2000	2575,9	2671,3	2749,3	2813,2	2875,6	2895,1	2960,1	3024,4				
Rate of change on moth -to-month base	1994		6,5	5,8	23,4	8,2	2,0	3,1	2,8	6,2	7,3	6,4	7,0
	1995	7,0	4,8	4,5	5,8	3,3	2,5	3,0	4,3	7,6	6,3	4,7	3,8
	1996	8,3	4,5	5,6	6,7	4,5	2,5	2,1	4,8	6,1	6,5	5,2	3,4
	1997	5,9	5,7	5,4	6,6	4,7	2,9	6,3	6,2	7,3	8,3	6,6	5,1
	1998	7,2	4,4	4,3	4,7	3,5	2,4	3,4	4,0	6,7	6,1	4,3	3,3
	1999	4,8	3,2	4,1	4,9	2,9	3,3	3,8	4,2	6,0	6,3	4,2	5,9
	2000	4,9	3,7	2,9	2,3	2,2	0,7	2,2	2,2				
Rate of change on december of the previous year base	1995	7,0	12,2	17,3	24,1	28,2	31,5	35,6	41,5	52,3	62,0	69,6	76,0
	1996	8,3	13,2	19,5	27,5	33,3	36,7	39,6	46,2	55,2	65,3	73,8	79,8
	1997	5,9	11,9	17,9	25,7	31,6	35,4	43,9	52,8	63,9	77,6	89,4	99,1
	1998	7,2	12,0	16,8	22,2	26,5	29,6	33,9	39,2	48,6	57,6	64,3	69,7
	1999	4,8	8,1	12,5	18,0	21,5	25,4	30,2	35,7	43,8	52,9	59,3	68,8
	2000	4,9	8,7	11,9	14,5	17,1	17,8	20,5	23,1				
Rate of change on same month of the previous year base	1996	125,8	122,4	119,6	88,4	79,8	80,9	80,8	83,4	86,0	84,3	81,5	76,0
	1996	78,1	77,5	79,3	80,8	82,9	82,9	81,2	81,9	79,3	79,6	80,4	79,8
	1997	75,7	77,7	77,3	77,2	77,5	78,0	85,2	87,8	89,9	93,2	95,8	99,1

(continued)

Table 5.17 (continued)

	Year	Jan.	Feb.	Mar.	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
	1998	101,6	99,3	97,2	93,6	91,4	90,6	85,3	81,4	80,4	76,6	72,8	69,7
	1999	65,9	63,9	63,5	63,9	63,0	64,3	65,0	65,4	64,3	64,7	64,6	68,8
	2000	68,9	69,7	67,9	63,8	62,7	58,6	56,2	53,2				
Rate of change on averages of the twelve month base	1995												89,1
	1996	86,0	83,2	80,9	80,5	80,8	81,0	81,0	81,0	80,5	80,2	80,1	80,4
	1997	80,0	79,9	79,6	79,3	78,9	78,5	79,1	79,8	80,9	82,3	83,9	85,7
	1998	88,0	89,9	91,4	92,6	93,4	94,0	93,6	92,6	91,4	89,6	87,3	84,6
	1999	81,5	78,5	75,8	79,5	71,3	69,5	68,2	67,1	66,1	65,3	64,8	64,9
	2000	65,2	65,8	66,1	66,0	65,9	65,2	64,3	63,1				

Source: The TUIK CPI Monthly Bulletin

Table 5.18 Index numbers and the rate of change of the WPI for urban settlement (%), 1994 = 100

		Year	Jan.	Feb.	Mar.	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Index	Total	1997	435.8	462.8	490.7	517.9	544.8	563.4	593.1	624.6	663.7	708.0	747.6	787.7
		1998	839.1	877.4	912.7	949.3	980.2	995.5	1020.7	1045.3	1101.2	1146.8	1185.7	1215.1
		1999	1258.6	1301.0	1352.9	1424.4	1469.9	1496.5	1556.0	1606.8	1700.8	1780.1	1852.7	1979.5
		2000	2094.0	2179.3	2246.8	2300.5	2339.5	2346.4	2370.5	2393.0				
	Public	1997	437.5	459.4	470.2	481.4	512.6	539.0	605.0	654.3	697.8	728.1	764.4	807.9
		1998	827.0	834.1	843.4	858.1	881.8	913.4	940.9	970.9	1034.1	1055.8	1081.8	1095.9
		1999	1131.9	1164.9	1230.7	1314.3	1394.3	1482.6	1638.4	1739.5	1871.4	1964.9	2090.5	2386.2
		2000	2479.9	2547.1	2619.5	2647.5	2680.4	2732.2	2770.9	2792.6				
	Private	1997	435.3	463.8	496.9	528.9	554.5	570.7	589.5	615.7	653.4	701.9	742.5	781.6
		1998	842.8	890.4	933.6	976.8	1009.9	1020.2	1044.8	1067.8	1121.4	1174.2	1217.0	1251.0
		1999	1296.8	1342.0	1389.8	1457.6	1492.7	1500.7	1531.1	1566.8	1649.3	1724.4	1781.0	1856.8
		2000	1977.6	2068.4	2134.4	2195.9	2236.7	2230.1	2249.7	2272.5				
Rate of Change on moth-to-month base	Total	1997	5.6	6.2	6.0	5.5	5.2	3.4	5.3	5.3	6.5	6.7	5.6	5.4
		1998	6.5	4.6	4.0	4.0	3.3	1.6	2.5	2.4	5.3	4.1	3.4	2.5
		1999	1.6	3.4	4.0	5.3	3.2	1.8	4.0	3.3	5.9	4.7	4.1	6.8
		2000	5.8	4.1	3.1	2.4	1.7	0.3	1.0	0.9				
	Public	1997	7.0	5.0	2.4	2.4	6.5	5.2	12.2	8.1	6.6	4.3	5.0	5.7
		1998	2.4	0.9	1.1	1.7	2.8	3.6	3.0	3.2	6.5	2.1	2.5	1.3
		1999	3.3	2.9	5.6	6.8	6.1	6.3	10.5	6.2	7.6	5.0	6.4	14.1
		2000	3.9	2.7	2.8	1.1	1.2	1.9	1.4	4.8				
	Private	1997	5.2	6.5	7.1	6.4	4.8	2.9	3.3	4.4	6.1	7.4	5.8	5.3
		1998	7.8	5.6	4.9	4.6	3.4	1.0	2.4	2.2	5.0	4.7	3.6	2.8
		1998	3.7	3.5	3.6	4.9	2.4	0.5	2.0	2.3	5.3	4.6	3.3	4.3
		2000	6.5	4.6	3.2	2.9	1.9	0.3	0.9	1.0				

(continued)

Table 5.18 (continued)

		Year	Jan.	Feb.	Mar.	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Rate of change on december of the previous year base	Total	1997	5.6	12.2	29.0	25.6	32.1	36.6	43.8	51.4	60.9	71.6	81.2	91.0
		1998	6.5	11.4	15.9	20.5	24.4	26.4	29.6	32.7	39.8	45.6	50.5	54.3
		1999	3.6	7.1	11.3	17.2	21.0	23.2	28.1	32.2	40.0	46.5	52.5	62.9
		2000	5.8	10.1	13.5	16.2	18.2	18.5	19.8	20.9				
	Public	1997	7.0	12.3	15.0	17.7	25.3	31.8	47.9	60.0	70.6	78.0	86.9	97.5
		1998	2.4	3.2	4.4	6.2	9.1	13.1	16.5	20.2	28.0	30.7	33.9	35.6
		1999	3.3	6.3	12.3	19.9	27.2	35.3	49.5	58.7	70.8	79.3	90.8	117.7
		2000	3.9	6.7	9.8	11.0	12.3	14.5	16.1	17.0				
	Private	1997	5.2	12.1	20.1	27.9	34.1	38.0	42.5	48.9	58.0	69.7	79.5	89.0
		1998	7.8	13.9	19.4	25.0	29.2	30.5	33.7	36.6	43.5	50.2	55.7	60.1
		1999	3.7	7.3	11.1	16.5	19.3	20.0	22.4	25.2	31.8	37.8	42.4	48.4
		2000	6.5	11.4	15.0	18.3	20.5	20.1	21.2	22.4				
Rate of change on same month of the previous year base	Total	1997	78.0	78.6	77.0	72.8	74.6	75.7	80.7	83.4	85.4	87.5	88.4	91.0
		1998	92.5	89.6	86.0	83.3	79.9	76.7	72.1	67.4	65.9	62.0	58.6	54.3
		1999	50.0	48.3	48.2	50.0	51.0	50.3	52.4	53.7	54.4	55.2	56.3	62.9
		2000	66.4	67.5	66.1	61.5	59.2	56.8	52.3	48.9				
	Public	1997	88.2	88.6	76.0	63.2	68.3	71.6	87.8	92.0	95.0	93.7	92.8	97.5
		1998	89.0	81.6	79.4	78.3	72.0	69.5	55.5	48.4	48.2	45.0	41.5	35.6
		1999	36.9	39.7	45.9	53.2	58.1	62.3	74.1	79.2	81.0	86.1	93.2	117.7
		2000	119.1	118.7	112.8	101.4	92.2	84.3	69.1	60.5				
	Private	1997	75.1	75.8	77.2	75.7	76.4	76.9	78.6	80.8	82.5	85.6	87.0	89.0
		1998	93.6	92.0	87.9	84.7	82.1	78.8	77.2	73.4	71.6	67.3	63.9	60.1
		1999	53.9	50.7	48.9	49.2	47.8	47.1	46.5	46.7	47.1	46.9	46.3	48.4
		2000	52.5	541	53.6	50.7	49.8	48.6	46.9	45.0				

Rate of change on averages of the twelve months base	Total	1997	76.9	78.0	78.7	78.6	78.5	78.3	78.7	79.2	79.8	80.4	81.0	81.8
		1998	83.2	84.1	84.8	85.4	85.4	85.1	84.0	82.2	80.2	77.8	25.0	71.8
		1999	68.2	64.8	61.7	59.2	57.0	55.2	53.9	53.1	52.4	52.1	52.1	53.1
		2000	54.6	56.3	57.8	58.7	59.3	59.7	59.4	58.8				
	Public	1997	84.1	86.4	86.5	84.5	83.1	82.2	83.1	83.9	84.7	85.0	85.0	85.5
		1998	85.7	85.1	85.0	85.7	85.4	84.5	81.1	76.7	72.3	67.9	63.5	58.4
		1999	54.3	51.1	49.0	47.8	47.4	47.6	49.7	52.6	55.7	59.4	64.0	71.2
		2000	78.3	84.9	90.4	94.0	96.2	97.1	95.5	92.5				
	Private	1997	74.8	75.6	76.5	76.9	77.2	77.2	77.4	77.8	78.4	79.1	79.8	80.7
		1998	82.4	83.8	84.7	85.3	85.5	85.3	84.8	83.8	82.6	80.7	78.5	75.9
		1999	72.4	68.9	65.5	62.6	59.8	57.4	55.1	53.2	51.5	50.1	48.9	48.1
		2000	48.2	48.6	49.1	49.2	49.4	49.5	49.4	49.2				

Source: The TUIK WPI monthly bulletin

5.12 A Review of this Chapter

In this chapter, we have examined the index numbers. Firstly, we looked at a simpleinterpretation of an index number and then we practiced chaining t wo different base year indices. Secondly, the weighted indices namely the Laspeyres and Paasche price indices were examined. Finally, we looked at price indices in practice.

5.12.1 Review Problems for Index Numbers

5.1. The following data show the gross trading profits of companies in TL billions.

1997	1998	1999	2000	2001	2002
51760	59190	63902	64415	68073	67969

(a) Turn the data into an index number series with 199 as the reference year.
(b) Transform the series so that 2000 is the reference year.
(c) What percentage increase has there been in profits between 1997 and 2002? Between 2000 and 2002.

5.2. Here are some fictional data on meals eaten that are not made at home. Figures are in Turkish Liras (TL).

	2010		2011		2012	
Items	Price (TL)	Quantity 000 s	Price (TL)	Quantity 000 s	Price (TL)	Quantity 000 s
Restaurant meals	6	200	7	195	7.5	180
Canteen meals	3	250	3.8	145	3.8	170
Fastfood	2.5	100	2.4	310	2.5	285

(a) Construct base weighted (Laspeyre) price indices for 2011 and 2012 for ‘meals eaten that are not produced at home’ with 2010 as the base year.
(b) Construct current weighted (Paasche) price indices for 2011 and 2012 for ‘meals eaten that are not produced at home’ with 2010 taken as 100.
(c) Why are the Laspeyre price indices for 2011 and 2012 greater than the Paasche indices?

5.3. A student consumes two goods, food and books. The price (P) and quantity (Q) consumed for two periods are given in the table below:

	Books		Food	
Period	P	Q	P	Q
1	15	3	1.5	25
2	18	4	2	30

Calculate the rate of inflation, based on both Laspeyres and Paasche price index between the two periods.

5.4. A petrol station recorded the following pattern of petrol sales:

Petrol sold (gallons)	No. of vehicles
0-5	75
6-10	193
11-15	281
16-20	106
21-25	24
26-30	7

(a) Identify and correct a weakness in the method of recording the quantities sold. Explain any assumptions you have made.
(b) Calculate the mean quantity of petrol sold per vehicle.
(c) What is the median quantity of petrol sold?
(d) Calculate the standard deviation of petrol sales
(e) Estimate the quantity of petrol sold on a normal day when 140 vehicles would be expected to call.

Chapter 6
Inequality Indices

6.1 Introduction

In this chapter we will continue to look at indices, i.e. inequality indices. Equality is an important issue. Most of us have some idea of 'fairness'—large inequalities tend to offend our senses. We will learn a method to describe and measure an inequality, which is a distributional inequality. We will focus on two types of measures; the Lorenz curve and the Gini coefficient.

6.2 The Lorenz Curve

It is invented by K. Lorenz around 1905 as a way of illustrating inequality. It is usually used for Income and wealth. The starting point is to have a sample of observations.

Lets use an imaginative simplified income data as an example. Table 6.1 shows the distribution of Income in this simplified form.

ch6data1.xls

The Lorenz curve is a way of graphically presenting the whole distribution.

What percentage of households have which percentage of the income?

We need to obtain the percentages of the income and the percentages of the households. Then we are able to compare the exact measure.

The online version of this chapter (doi:10.1007/978-3-319-26497-4_6) contains supplementary material, which is available to authorized users.

D. Özdemir, *Applied Statistics for Economics and Business*,
DOI 10.1007/978-3-319-26497-4_6

Table 6.1 A simplified monthly income data

Range (TL) Monthly Income	Mid point x	Numbers of Households f
0–	40	950
80–	105	1800
130–	190	1300
250–	350	610
450–	800	400
800–	900	200
1000–	1250	180
1500–	1.75	80
Total		5520

TL Turkish Liras

The first step is to calculate the total income for each range. Then obtain the percentages. Table 6.2 shows these calculations.

Columns are obtained by the following formulae: From left to right,

Column 4 = column 2 × column 3
Column 5 = column 3 ÷ 5520.
Column 6 = column 5 cumulated.
Column 7 = column 4 ÷ 1552500
Column 8 = column 7 cumulated

The table indicates a substantial degree of inequality. For example the richest 1.45 % of households earn more than 1500TL per month and the poorest 17 % earn nearly 20 times less. From the table the 17.21 % poorest people earn 2.45 % of the total income. Since 0 % of households earn 0 % of income, and 100 % of households earn 100 % of income, the curve must run from the origin up to the opposite corner. The Lorenz curve also must lie below the (perfect equality line) line because households are ranked from poorest to richest. On the Fig. 6.1 as we move to the right we encounter successively richer individuals and cumulative income grows faster, thus the curve must be concave. Then if we plot the cumulative % incomes (column 6) against the cumulative % households (column 8). (For excel plotting we need to take the cumulative f column twice so that we can represent perfect equality straight line). We obtain the Lorenz curve for this data.

If there is a perfect equality then the curve would be on the same line as the straight line in the middle (perfect equality line). However normally, income or wealth are not distributed equally.

The further away from the line the Lorenz curve is, the greater the degree of inequality.

Table 6.2 The Lorenz Curve calculations

Range (TL) Monthly Income	Mid point x	Numbers of Households f	fx	%f	% Cumulative	% Income	Cumulative Income
0-	40	950	38000	17.21	17.21	2.45	2.45
80-	105	1800	189000	32.61	49.82	12.17	14.62
130-	190	1300	247000	23.55	73.37	15.91	30.53
250-	350	610	213500	11.05	84.42	13.75	44.28
800-	900	200	180000	3.62	95.29	11.59	76.49
1000-	1250	180	225000	3.26	98.55	14.49	90.98
1500-	1,750	80	140000	1.45	100.00	9.02	100.00
Total		5520	1552500	100.00		100.00	

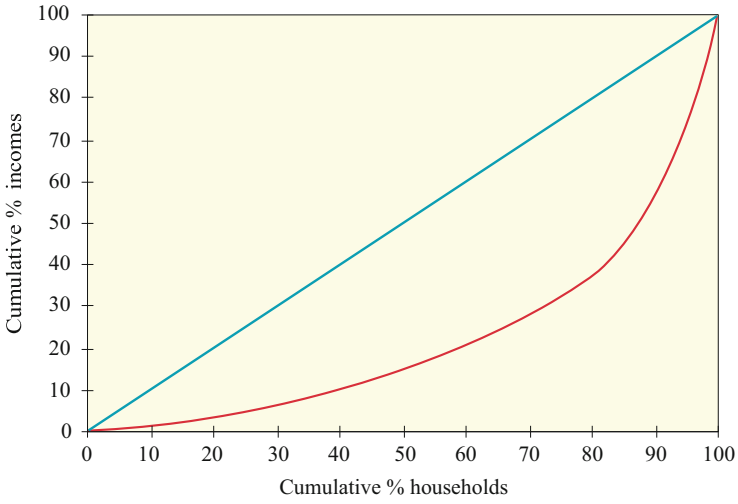


Fig. 6.1 Lorenz curve

6.2.1 The Distribution of Income and the Lorenz Curve in Turkey

In this part we will apply, the income data you have been familiar with from our previous Chapters into our analysis. This data is obtained from SIS sources. There are two surveys conducted in most recent years; the first one was in 1987 and the recent one was in 1994. We will first look into the earlier income distribution survey that is 1987 survey.

The income range is chosen by the TUIK sources in this example. We will firstly derive the Lorenz curve from this data and later on we will compare these two survey results within the framework of the Lorenz curve (Table 6.3).

Table 6.3 The TUIK household income survey results for Turkey

Income groups TL	Number of households	Total Income TL
0–49999	321746	10674326
50000–99999	1254863	99023453
100000–149999	1799098	222434428
150000–199999	1695101	294254201
200000–249999	1330912	295933661
250000–299999	987626	268315943
300000–349999	767705	247523158
350000–399999	539023	200770779
400000–449999	471176	199308071
450000–499999	304610	143795912
500000–599999	508190	275493366
600000–699999	312648	200143170
700000–799999	190017	141024361
800000–899999	144870	122320401
900000–999999	87654	82538197
1000000–1499999	186600	219570936
1500000–1999999	74114	126035614
2000000–4999999	62535	174767006
5000000–9999999	7502	43815423
10000000–24999999	1570	20681050
TOTAL:	11047560	3388423456

ch2data6.xls

One of the problems of working with real data is the complications and this complications become worst with increased numbers due to a high inflation in Turkey. Here again we will follow the same process as we did in our previous simplified data example.

In this curve calculations we can see a substantial degree of inequality. For example the richest 0.01 % of households earn more than 10.000 TL per month and the poorest 2.91 % earn 200 times less. From the table the 2.91 % poorest people earn 0.23 % of the total income. Similarly plotting the calculations into a Lorenz curve gives the Fig. 6.2 (Table 6.4).

As we mentioned earlier the Lorenz curve never equal to the line because in practice we have not seen a perfect equality case even in the socialist countries which had a main goal for equality. One way to get an idea about equality is to compare two cases. For this aim we will firstly look into the later Income Survey results in Turkey than we will try to look into another country example.

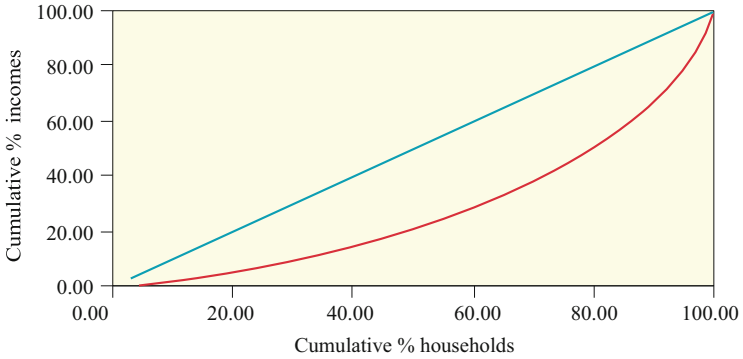


Fig. 6.2 Income distribution in Turkey, 1987

6.2.2 *Comparative Investigations of the Last Two Income Distribution Surveys in Turkey, 1987 and 1994*

In this part instead of looking into very details of the calculations we will take the two surveys income and population percentages (Table 6.5).

ch6data2.xls

A similar Lorenz curve calculations are carried out in EXCEL and the result is presented in Table 6.6.

If we plot these two Lorenz curves, one belong to 1987 and the other belong to 1994 data in the same diagram we will have the opportunity of seeing the progress of income distribution the country. It can clearly be seen that the distribution gets worst after seven year later. There are obviously number of reasons for this change but our aim is not to discuss the reasons in this book. It is rather to see the technical part of calculations and obtaining these curves. Plotting the cumulative percentage of household income groups and the percentage income cumulative columns in excel gives us the following diagram with two Lorenz curves.

The Lorenz curve in 1987 is closer to the perfect equality line which indicates that the distribution of income was more equal as opposed to 1994 distribution.

6.2.3 *Another Country Example with the Effect of Taxes on Income Distribution*

We will now consider a European country's income distribution, the UK. The income data here is taken from UK Inland Revenue statistics in 1992–1993. The tax payers are in millions and the income is in pounds sterling. From economics

Table 6.4 Excel Lorenz curve calculations for Turkish data

Income groups (TL)	Mid point (X)	Frequency f, No. of household	fX	% f	Cumulative % f	% Income	Cumulative Income (%)
0–49999	24999.5	321746	8043489127	2.91	2.91	0.23	0.23
50000–99999	74999.5	1254863	94114097569	11.36	14.27	2.71	2.94
100000–149999	124999.5	1799098	2.24886E + 11	16.29	30.56	6.47	9.41
150000–199999	174999.5	1695101	2.96642E + 11	15.34	45.90	8.53	17.94
200000–249999	224999.5	1330912	2.99455E + 11	12.05	57.95	8.62	26.56
250000–299999	274999.5	987626	2.71597E + 11	8.94	66.89	7.81	34.37
300000–349999	324999.5	767705	2.49504E + 11	6.95	73.84	7.18	41.55
350000–399999	374999.5	539023	2.02133E + 11	4.88	78.71	5.82	47.37
400000–449999	424999.5	471176	2.0025E + 11	4.26	82.98	5.76	53.13
450000–499999	474999.5	304610	1.4469E + 11	2.76	85.74	4.16	57.29
500000–599999	524999.5	508190	2.66799E + 11	4.60	90.34	7.68	64.97
600000–699999	649999.5	312648	2.03221E + 11	2.83	93.17	5.85	70.81
700000–799999	749999.5	190017	1.42513E + 11	1.72	94.89	4.10	74.91
800000–899999	849999.5	144870	1.23139E + 11	1.31	96.20	3.54	78.46
900000–999999	949999.5	87654	83271256173	0.79	96.99	2.40	80.85
1000000–1499999	1249999.5	186600	2.3325E + 11	1.69	98.68	6.71	87.56
1500000–1999999	1749999.5	74114	1.29699E + 11	0.67	99.35	3.73	91.29
2000000–4999999	3499999.5	62535	2.18872E + 11	0.57	99.92	6.30	97.59
5000000–9999999	7499999.5	7502	56264996249	0.07	99.99	1.62	99.21
10000000–24999999	17500000	1570	27474999215	0.01	100.00	0.79	100.00
TOTAL:		11047560	3.47582E + 12	100.00			

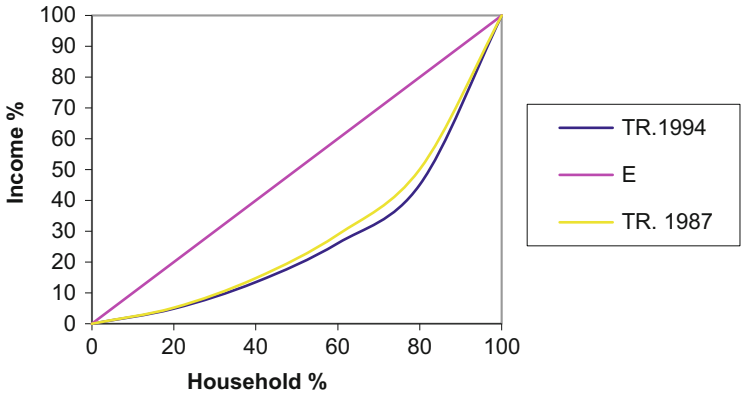


Fig. 6.3 Lorenz curves in Turkey 1987 and 1994

Table 6.5 Comparing income distribution between 1987 and 1994

Lowest to Highest % Household	Disposable Income % 1994	Disposable Income 1987 %
First% 20	4.9	5.2
Second% 20	8.6	9.6
Third% 20	12.6	14.1
Fourth% 20	19	21.2
Fifth% 20	54.9	49.9

Table 6.6 Comparing two Lorenz curves calculations in Turkey

Lowest to Highest % Household	Disposable Income % Turkey, 1994	Disposable 1987 TR Income	Cum.H.H	Cum.Inc. TR.1994	Cum. Inc. TR.1987
			0	0	0
First%20	4.9	5.2	20	4.9	5.2
Second%20	8.6	9.6	40	13.5	14.8
Third%20	12.6	14.1	60	26.1	28.9
Fourth%20	19	21.2	80	45.1	50.1
Fifth% 20	54.9	49.9	100	100	100

theory we know that a higher income tax tend to help more equal distribution of income. However not every forms of tax helps income distribution for example poll tax or VAT may increase the unequal distribution of income. We consider the income tax in this example. Table 6.7 is taken from the UK Inland revenue statistics for 1992–1993.

Distribution of total incomes before and after tax, 1992–1993.

Table 6.7 Distribution of total incomes before and after tax

Range of total income before tax (lower limit) Pounds	Number of tax payers	Total income before tax	Total income after tax
	(Millions)	Pounds	Pounds
3.445	2.2	9.600	9.300
5.000	3.9	24.500	22.700
7.500	3.8	33.500	29.900
10.000	6.2	76.600	66.600
15.000	3.8	65.800	55.800
20.000	3.2	75.800	62.400
30.000	0.9	29.500	23.300
40.000	0.7	52.200	36.900
All ranges	24.8	367.400	306.800

Table 6.8 Calculations for the Lorenz curve. Effect of a progressive tax on income distribution

		%	%
Cumulative f	Cum. Inc.	Cum. Inc.	Cumulative
x	Equality	Inc.Bef. T.	Inc. aft.T.
0 %	0 %	0 %	0 %
891 %	891 %	261 %	303 %
2470 %	2470 %	928 %	1037 %
4008 %	4008 %	1839 %	2012 %
6519 %	6519 %	3924 %	4183 %
8057 %	8057 %	5714 %	6003 %
9353 %	9353 %	7777 %	8037 %
9717 %	9717 %	8579 %	8797 %
10000 %	10000 %	10000 %	10000 %

ch3data2.xls

If we apply the same methods and calculations for income before and after tax we would obtain the following cumulative % households and the cumulative % incomes (Table 6.8).

If we plot these columns than we can obtain the following Lorenz curve (Fig. 6.4).

The figure indicates that the taxes have a positive effect on equality, if these are the income tax. If the taxes differ from the income tax, such as council tax VAT etc. then the effect may be different.

The Lorenz curve is good to compare a few country or periods distributions however it may be not that practical, for example to compare large number of countries distribution of income for this reason we will look into another way of measuring the distribution of income. The Gini coefficient.

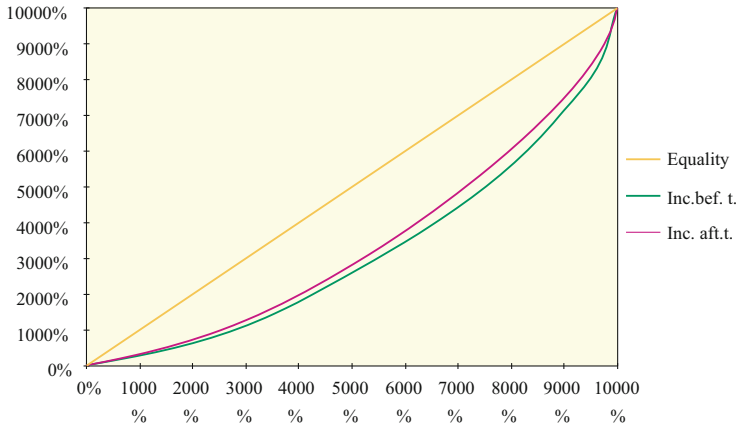


Fig. 6.4 Lorenz curve for after and before income tax

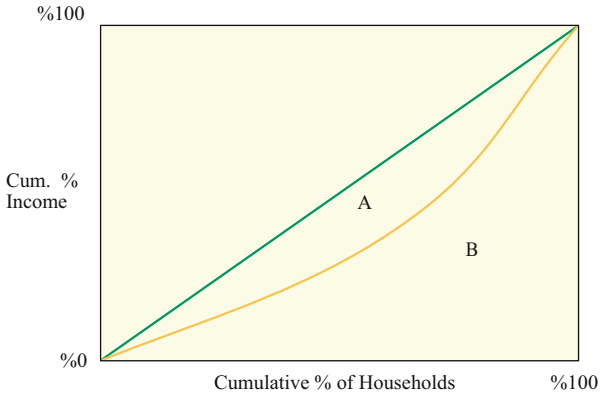


Fig. 6.5 A typical Lorenz curve and the Gini coefficient

6.3 A Numerical Measure of Inequality: Gini Coefficient

Summarizes what the Lorenz curve shows us in a single number. It can be derived directly from the Lorenz curve.

The Gini coefficient is the ratio of the area A to the sum of areas A and B.

$$Gini = G = \frac{A}{A + B}$$

Thus Gini coefficient is between zero and one. If $A = 0$ then $G = 0$ There is a total equality. If $B = 0$ then $G = A/(A + 0) = 1$ there is total inequality, i.e. one household is having all the income. These extreme cases are unlikely in the real life. It is usually between these two.

Table 6.9 Income distribution data for gini calculations.

Income Range	No. of Income Units	Total Incomes (TL)
2000–	5700	18400
4500–	6735	41480
8000–	5020	49000
12000–	4429	82700
Total =	21884	191580

Table 6.10 Income distributions and the calculations of the Lorenz curve

Income Range (TL)	Number of income units	Per cent income units	Cum % income	Total incomes (TL)	Per cent income	Cum per cent income
(1)	(2)	(3)	(4)	(5)	(6)	(7)
,000–	5,700	26.05	26.05	18400	9.60	9.60
4,500–	6,735	30.78	56.82	41480	21.65	31.26
8,000–	5,020	22.94	79.76	49000	25.58	56.83
12,000–	4,429	20.24	100.00	82700	43.17	100.00
Totals	21884	100.00		191580	100.00	

How do we obtain the values of the A and B?
Formula for B is;

$$B = \frac{1}{2}\{(x_1 - x_0)*(y_1 + y_0) + (x_2 - x_1)*(y_2 + y_1) + \ldots\ldots\ldots\ldots + (x_k - x_{k-1})*(y_k + y_{k-1})\}$$

where $k =$ Number of classes for income in the frequency table.

$x_0 = y_0 = 0$
$x_k = y_k = 0$ } The coordinates of the two end points of the Lorenz curve.

If we use percentages then the area A + B is 5000 because the multiplication of the two axes (100 × 100) divided by two gives this number, which is the area of the triangle on the Fig. 6.5. Then A = 5000−B. The Gini coefficient is:

$$G = \frac{A}{A + B} = \frac{A}{5000}$$

Example
Let us provide an example and calculate the Gini coefficient. The following data on the income distribution were given (Table 6.9).

ch6data3.xls

For calculation of the Gini coefficient, firstly we need to obtain the cumulative percentages, as we did in the Lorenz curve calculations. These calculations are presented in Table 6.10.

Table 6.11 Gini coefficient calculations in EXCEL

Gini (x1 − xo)	Gini (y1 + yo)	(x1 − xo)*(y1 + yo)
26.05	9.60	125.08
30.78	40.86	628.76
22.94	88.09	1010.34
20.24	156.83	1587.03
		B = 3351.20
A = 5000 − B = 1648.80		
Gini = (A/5000) = 0.33		

By using the formula for the Gini coefficient, firstly we calculate the values of B. These are presented in Table 6.11.

The Gini coefficient is equal to 0.33 or 33 %. In this method the calculated Gini coefficient is biased downwards because the method, wrongly, assumes that the Lorenz curve is made up of straight line segments connecting the observed points. Since the straight lines lie inside the actual Lorenz curve then the area B is over estimated. This means that the estimated coefficient is actually bigger than the obtained number. The true value of the Gini coefficient is slightly greater than 33 % in our example. This biasness will be higher when the observations are smaller and the more concave is the Lorenz curve is. The way to eliminate this problem of bias is to draw the Lorenz curve on graph paper and count squares. In this way you can draw a smooth line joining the observations and avoid the bias problem. The disadvantage of is that you can not use the computer!.

How do we know whether this is a high or low inequality?

One way to find out is to compare this number to the Gini coefficients of other countries or to look at trends in inequality over time. To look trends in inequality overtime is a difficult practice in Turkey because There aren't very many distribution surveys carried out in Turkey. The survey result of 1987 and 1994 provides gini coefficients but these surveys are not carried out in details for every year. Therefore we have limited Turkish data in practice. Gini coefficients for these two years are: 1987 = 0.40 and 1994 = 0.44. As the Gini coefficient increases so the inequality.

The another use of the Gini coefficient is to compare the Gini coefficient between countries. Table 6.12 presents the Gini coefficients for the number of countries in the world.

There is another simpler measure of the Gini coefficient.

6.4 A Simpler Formula for the Gini Coefficient

This method is applicable if the class intervals contain equal numbers of households, for example when the data is given for deciles or quintiles of the income distribution.

Table 6.12 Gini coefficients in the world (The averages are taken from the original data)

Low income	Economies		Industrial	Market	Econ.
Country	Year	Gini	Country	Year	Gini
Bangladesh	1981–1982	0.36	Spain	1980–1981	0.31
India	1975–1976	0.38	Ireland	1973	0.30
Kenya	1976	0.51	New Zealand	1981–82	0.37
Zambia	1976	0.50	Italy	1977	0.35
Sri Lanka	1980–1981	0.39	UK	1979	0.31
	Average	**0.43**	Belgium	1978–1979	
Middle income	Economies	(Lower mid Inc.)			
Indonesia	1976	0.41	Netherlands	1981	0.26
Philippines	1985	0.42	France	1975	0.34
Cote d'Ivoire	1985–1986	0.52	Australia	1975–1976	0.38
Egypt	1974	0.38	West Germany	1978	0.30
Thailand	1975–1976	0.40	Finland	1981	0.30
Peru	1972	0.54	Denmark	1981	0.32
Turkey	1973	0.47	Japan	1979	0.27
	Average	**0.45**	Sweden	1981	0.31
			Canada	1981	0.33
Upper-middle income					
Hungary	1982	0.27	Norway	1982	0.30
Brazil	1972	0.56	Switzerland	1978	0.29
Israel	1979–1980	0.32	United States	1980	0.33
Portugal	1973–1974	0.40			
	Average	**0.42**			

Table 6.13 The distribution of income by quintile

Quintile	1	2	3	4	5
% Income	4.9	8.6	12.6	19.0	54.9
Cumulative %	4.9	13.5	26.1	45.1	100

(Source; TUIK 1994 Income distribution survey results)

Let us reconsider Turkish 1994 distribution of income by quintiles. The Table 6.13 presents the 1994 distribution of income by quintiles. By using this table we can see that the area B is simplified by the formula below:

$$B = \frac{100}{2k}(y + 2y_1 + 2y_2 + \ldots\ldots\ldots\ldots + 2y_{k-1} + y_k) = \frac{100}{k}\sum_{i=0}^{i=k} y_i - 50$$

where $k =$ number of intervals, i.e. in the case of quintiles 5 but in the case of deciles 10. In this example $k = 5$.

The sum of the values for y appear in the final row of Table 6.13 and it is 189.6 (sum of the cumulative % row)

Thus the substitution of these values in to the formula above gives: $B = (100/5)$ $*(189.6 - 50) = 2792$ hence A $= 5000 - 2792 = 2208$

The Gini coefficient is G $= (2208\ /5000) = 0.4416$ or about 44 %.

Equality

To interpret the figures we need to consider the ideal value. Complete equality (Gini $= 0$) is not desirable because of incentive effects. We should also take account of stage of life cycle, particularly when looking at wealth distributions. For example a lecturer currently earn more than most students, though they will probably earn more over their lifetimes. The problem of age adjustment is even more severe for wealth; but it has been shown that even correcting for age differences, there is still great inequality in the distribution of wealth.

Better Measures of Inequality

The Gini coefficient is not ideal; perhaps no measure is. Rather than just inequality of income, we might worry more about the distribution of 'basic needs' e.g. food, housing etc. We wouldn't worry about an unequal distribution of diamonds (per se), for example.

6.5 A Review of This Chapter

In this chapter we have examined the inequality indices. These are the Lorenz curve and the Gini coefficient. For our purpose we mainly used the Turkish income data.

6.6 Review Problems for Inequality Indices

6.1. Use the data on incomes before and after tax given in question 2.3 to plot the Lorenz curve for pre and post-tax income. Comment.

6.2. The table below provides the quintile distribution of income in Turkey, 1987. Calculate the Gini coefficient from this data.
The distribution of income by quintile.

Quintile	1	2	3	4	5
% Income	5.2	9.6	14.1	21.2	49.9
Cumulative %	5.2	14.8	28.9	50.1	100

6.3 A household income survey results are given in the following table. Plot the Lorenz curve. Comment.

ch2data6.xls

Income groups (TL)	Mid point (X)	(Frequency) f, No. of households
0–49999	24999.5	321746
50000–99999	74999.5	1254863
100000–149999	124999.5	1799098
150000–199999	174999.5	1695101
200000–249999	224999.5	1330912
250000–299999	274999.5	987626
300000–349999	324999.5	767705
350000–399999	374999.5	539023
400000–449999	424999.5	471176
450000–499999	474999.5	304610
500000–599999	524999.5	508190
600000–699999	649999.5	312648
700000–799999	749999.5	190017
800000–899999	849999.5	144870
900000–999999	949999.5	87654

6.4 In the chapter on inequality, the following data on the income distribution were given, from which a Gini coefficient of 0.33 was calculated.

Income range (TL)	Number of income units	Per cent income units	Cum. % income	Total incomes (TL)	Per cent income	Cumulative per cent income
(1)	(2)	(3)	(4)	(5)	(6)	(7)
2,000–	5,700	26.05	26.05	18,400	9.60	9.60
4,500–	6,735	30.78	56.82	41,480	21.65	31.26
8,000–	5,020	22.94	79.76	49,000	25.58	56.83
12,000–	4,429	20.24	100.00	82,700	43.17	100.00
Total	21,884	100.00		191,580	100.00	

The actual post tax income distribution for that year was:

Income range (TL)	Number of income units	Per cent income units	Cum. % income	Total incomes (TL)	Per cent income	Cumulative per cent income
(1)	(2)	(3)	(4)	(5)	(6)	(7)
2,000–	5,700			16980		
4,500–	6,735			34970		
8,000–	5,020			39800		
12,000–	4,429			62430		
Totals	21884			154180		

Calculate the Gini coefficient for these data, and compare with that for the pre-tax distribution.

6.7 Computing Practical for Excel For Inequality Indices: The Lorenz Curve And The Gini C

C.6.1.
The data below refers to incomes before and after tax.

(a) Complete the table below in Excel for before tax and after tax.; Use the hints given in your Computing Practical Sheet 2.
Distribution of total incomes before and after tax, 1992–1993

Pre-tax inc.	f- million	Total inc.	Total inc.	% frequency	%	%	%
Lower limit	Number of	before tax	after tax		Cumulative frequency	Income	Cumulative
$	taxpayers	($ m)	($ m)	%f			Income
3445	2.2	9600	9300				
5000	3.9	24500	22500				
7500	3.8	33500	29900				
10000	6.2	76600	66600				
15000	3.8	65800	55800				
20000	3.2	75800	62400				
30000	0.9	29500	23300				
40000	0.7	52200	36900				
130945	24.7	367500					

(b) For the data on incomes before and after tax obtain the Lorenz curve;

On a separate space make four columns in the following order from left to right label the first column % cumulative frequency, call the second, which contains the same data as the first equality. Call the third column the percentage cumulative income before tax and the fourth column should have the percentage cumulative income after tax.

Select these columns including the titles, click the insert button and then choose option 5 in the XY (Scatter), then click 'ok'. Add legend and/or the titles for the axis then click on the finish icon. Now you have obtained an incomplete Lorenz curve, which needs axis scaling. In order to scale these axies, click on the Y axis and choose Axis options in vertical (value) axis. Adjust the scales of the Y axis by entering for 'Maximum = 100, Minimum = 0, Major unit = 10 and Minor unit = 1 in Axis option. Do the same scaling for the X axis. This exercise should provide you with two Lorenz curves in one figure. One for the income before tax and the other for the income after tax. Is inequality increasing after tax?. Comment. Save the work for a later printout.

C.6.2.
The following table shows the income distribution by quintile for various definitions of income:

Income measure				
Quintile	Original (%)	Gross (%)	Disposable (%)	Post-tax (%)
1 (bottom)	2.0	6.7	7.2	6.6
2	7.0	10.0	11.0	11.0
3	16.0	16.0	16.0	16.0
4	26.0	23.0	23.0	23.0
5	50.0	44.0	42.0	44.0

(a) Use the formula: $B = \frac{100}{2k}(y_0 + 2y_1 + 2y_2 + \ldots\ldots + 2y_{k-1} + y_k) = \frac{100}{k}\left(\sum_{i=0}^{i=k} y_i - 50\right)$
where k is number of intervals and calculate the Gini coefficient in EXCEL for each of the four categories of income (Use hint 22 from previous computing practical).

(b) For the 'original income' category, draw a smooth Lorenz curve on a piece of graph paper and calculate the Gini coefficient using the method of counting squares. How does your answer compare to that for part (a)? Show all your excel calculations. Print out the answers for these two questions.

6.8 Further Hints on Calculations

1. Remember (a)All your formulae must be able to recalculate automatically.
   (b) Save your work regularly.
2. In–built Functions. There are a number of functions in Excel. For these functions go to 'Formulas' on the tool bar and then choose insert Function, you will have a number of in built functions for a range of calculations, such as Mathematical, Statistical, Financial etc. For example you can calculate the mean, median, mode and standard deviation in this way. After entering your data in the spreadsheet using this method find the AVERAGE and STDEV functions from the statistical list and follow the instructions. (Note the STDEV function uses (n−1) as a divisor, the STDEVP function uses (n).

Chapter 7
Probability

7.1 Introduction

In this chapter we will discuss probability. In the previous chapters we examined the methods of description of data but it is not sufficient to have data or to know where the data can be found and describe it. Statistical inference involves drawing conclusions about a population from a sample of data, which is obtained from a population. If we want to understand sampling procedures and implications for statistical inference firstly probability theory needs to be examined because, for example, each member of the population has a particular probability of being in the sample. In many cases samples are randomly chosen from a population. In fact, if a second sample were selected it would almost certainly be different from the first. In simple random sampling, the probability of being sampled is the same for all members of the population.

Learning probability theory is the first step to measuring uncertainty. We do not live in a world of certainty. For example, it is not possible to say, with certainty, what the national lottery numbers will be this week. In fact, chance plays a part in every individual's life. For example, playing a lottery, buying life insurance or deciding about your degree: all involve some assessment of risk. In situations where all things can be known, data collection alone can provide a basis for decision making; for example say $A = 10$ and $B = 15$ then $A + B = 25$: the problem is deterministic. If $A = 10$ or 11 and $B = 15$ or 16 then $A + B = 25$ or 26 or 27: the outcome is subject to chance. In the presence of chance we cannot be sure of the outcome but it still makes a difference knowing whether the chance of a particular outcome is 1 in 10, 1 in 100 or 1 in a million. In this case the problem is probabilistic.

Let's look at an example to illustrate the relevance of probability theory: A test for AIDS is 99% successful, i.e. if you are HIV positive it will detect it in 99% of all tests, and if you are not, it will again be right 99% of the time.

D. Özdemir, *Applied Statistics for Economics and Business*,
DOI 10.1007/978-3-319-26497-4_7

Also assume that 0.5% of the population are HIV positive (in a 1992 sample of pregnant women in a London hospital 14/4000 were HIV positive, this is equivalent to 0.45% ~ 0.5%).

Say person A takes part in a random testing procedure, which gives a positive result.

Q. What is the probability that you are HIV positive?

The answer is 0.332. (probably much lower than expected). (This because all those who have had a positive test are either false positive (1% of 99.5% of the population) or true positives (99 % of 0.5 %) Here the group of false positives is bigger).

You will see later how to work this out that the assumption of 0.5% population with HIV is crucial.

This is in fact a good argument against random testing for HIV status, since a lot of people would face needless worry. Even if the tests were anonymous, the estimate of the number of HIV positive cases would be overestimated by ~200%.

Note that this implies only to random testing. If person A had some reason for taking the test then you should not be consoled by a positive test result by the above argument!

Example: If the people in A's risk category had a 20% HIV rate, the probability that you have HIV after a positive test is in fact ~96%.

In general we use probability theory as the basis for statistical inference - the drawing of conclusions about a population on the basis of a sample.

E. g.

- Is the estimated mean significantly different from zero?
- What confidence can we have in our regression estimates (i.e. slope)?
- What can we conclude about the incidence of HIV from a sample?

7.2 Basic Concepts and Definitions

7.2.1 Definitions

Experiment

An action that has a number of possible events/outcomes. (e.g. throwing a coin, drawing a lottery ticket, running a horse race).

Trial

One performance of experiment.

Outcomes

Simple (e.g. a six when a dice is thrown, heads, tails; Desert orchid winning) versus compound events (e.g. two sixes when a dice is thrown twice, Head(H), Tail(T))

Sample Space

A listing of all possible outcomes of a trial, in some form, of all the possible outcomes of a trial. E.g. heads, tails; all the cards in a pack.

Probability

The likelihood that an event will occur. Each outcome (point in the sample space) is associated with a probability (e.g. P (heads) = 0.5; P(ace of hearts) = 1/52).

In order to clarify the basic idea of the probability we will now look at a survey results. This survey is on about the people who play games of chance in Istanbul and has 580 respondents (250 male and 330 female). These games include the national lottery, betting on horse races (ganyan), sportoto and scratch cards. If we randomly select an individual, there would be 330 chances out of 580 that the person was female.

P(male) = (250/580) = 0.43

And read as the probability of a male or short: P(m). For female P(f) (Table 7.1).

The probability of selecting a person at random who does not play any chance games is: P(don't bet) = (280/580) = 0.48. The probability of selecting a female who does not use the betting shops is: P(female and don't bets) = (210/580) = 0.36. Probabilities of selecting a male and a female from this group is: P(m) = 0.43 and P (f) = 0.57.

Thus

P(m) + P(f) = 0.43 + 0.57 = 1 where male and female are abbreviated to 'm' and 'f'. The probability of not having a characteristic, for example not being female is:

Since

P(not f) + P(f) = 1. Or rearranging this equation: P(f) = 1- P(not f) = 1 - 0.43 = 0.57

Table 7.1 Respondents of the chance games

	Bet once or more than once per wk	Bet, but less than once per wk	Don't bet	Total
Male	110	70	70	250
Female	90	30	210	330
Total	200	100	280	580

This seems so obvious in this example of two probable outcomes case but in many cases it can be easier to find the probability of an event not happening than the probability of it happening.

7.3 Definitions of Probability

7.3.1 Classical Definition

Basically the case where a particular characteristic (e.g. being male) is divided by the total number of possible results. Let us call this event C, where each outcome is equally likely then:

$$\mathrm{P(C)} = (\text{no. of ways } C \text{ can } occur)/(\text{total number of outcomes})$$

For example if you throw a die the probability of a 4 is 1/6. If the die is weighted so that 5 appears each time it is thrown then the probability of a 4 is zero (P(4) = 0) where each outcome is not equally likely.

7.3.2 Objective Definition

What is P(heads) when throwing a coin? Most would say 0.5, but why?

An alternative way to look at probability would be to carry out an experiment to determine the proportion of outcomes in the long run. This is called the frequentist view. P = relative frequency. If a coin were tossed many times, approximately half of the outcomes would be heads and the other half is tail.

$$\begin{aligned}\mathrm{P(Heads)} &= \text{no. of heads}/\text{no. of tosses} \\ &= \text{no. of successes}/\text{no. of trials.}\end{aligned}$$

This would certainly overcome the problem involving the biased die given in the classical definition. P can be established by considering large samples. However this is not a problem-free option either. One must question how many times can you repeat the exercise or how long the long period is. For example experiments with an unbiased coin may not necessarily conclude that the probability of a head is 0.5, even after 10 million tosses. Another problem is that you may not be able to carry number of trials in certain situations.

For example, what is the probability of the TL going down (compared to other currencies) tomorrow? Here the frequentist view is of little use.

7.3.3 Subjective Definition

- P = degree of belief (Hence it can vary from individual to individual. What if not everybody agrees P(head) is 0.5? Suppose someone thinks it's 0.83?)
- belief is rational (One could argue that it's any rational degree of belief. But then the argument concerns what is rational).
- prior versus posterior beliefs.
- The Bayesian approach allows prior beliefs to be changed into posterior beliefs once evidence becomes known (The problem here though is to define the priors, especially when your prior belief is complete ignorance. Also probability tends to be a very mathematical analysis). Hence ignore the problem. P(heads) = 0.5.

7.4 Laws of Probability

There are four laws of probability for any event A or B;

1. Sum of probabilities of all outcomes in sample space equals 1.
   $\Sigma P_i = 1$

where is the probability of event i occurring
e.g. P(heads) + P(tails) = 0.5 + 0.5 = 1

2. Certainty
   If event A is certain to occur → P(A) = 1
   If event B is certain not to occur → P (B) = 0
3. $0 \leq P(A) \leq 1$
4. Complements:
   P(notB) = P(~B) = 1- P(B)
   e.g. P(not heads) = 1 - P (heads)

7.5 Probability Rules: Compound Events

The probability of a single event is important but most of the problems in economics have two or more events happening together. Hence we usually need to work out the probabilities of compound events. As a simple example, what is the probability of getting three heads in four tosses of coin? What is probability of drawing two aces from the pack in poker?

7.5.1 Mutually Exclusive Events

Addition Rule

Association with 'or' statement, such as the probability of getting two heads or two tails. E.g. what is probability of one head in two tosses?

$$\mathrm{P}(\text{one head}) = \mathrm{P}\big[\underset{\mathrm{A}}{(\mathrm{H\ and\ T})}\ \text{or}\ \underset{\mathrm{B}}{(\mathrm{T\ and\ H})}\big].$$

Hence equivalent to P(A or B).
The addition rule says

$$\mathrm{P}(\mathrm{A\,or\,B}) = \mathrm{P}(\mathrm{A} \cup \mathrm{B}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B}).$$

Now P(A) = P(H and T) = P(H1) * P(T2|H1) = 0.5 * 0.5 = 0.25.
Similarly for P(B), so P(A and B) = 0.25 + 0.25 = 0.5.

It is said that the events A and B are mutually exclusive if they cannot occur together. For example a clothing company produces jeans. If you buy a pair of jeans in your regular waist size without trying them on, the probability that the waist will be too tight is 0.30 and the probability it will be too loose is 0.10. In this case the waist can not be both too tight and too loose at the same time, so the events are mutually exclusive. The question is if you choose a pair of trousers at a random in your regular waist size, what is the probability that the waist will be too tight or too lose?

Since the event is mutually exclusive the probability is:

$$\mathrm{P}(\text{too tight or lose}) = \mathrm{P}(\text{too tight}) + \mathrm{P}(\text{too lose}) = 0.30 + 0.10 = 0.40.$$

So the general formula for a mutually exclusive event is:

$$\mathrm{P}(\mathrm{A\,or\,B}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B})$$

7.5.2 Non-Mutually Exclusive Events

The addition rule assumes A and B are mutually exclusive events, i.e. if one occurs the other cannot. A counter example is: what is the probability of drawing an ace or a spade from the pack?

P(ace) = 4/52, P(spade) 13/52, hence by addition rule the answer is 17/52. But from the sample space it is obvious there are only 16 different cards satisfying the criteria, 13 spades plus three other aces, hence the rule becomes:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = 16/52.$$

This avoids double counting A and B if they are not mutually exclusive, as in the case of aces and spades.

If the events A an B can occur together then the event is non-mutually exclusive.

For example

Dr. Adam Hepoku is in charge of the postgraduate programme. Records show that 40 % of the applicants need course work in Maths, 70 % need work in English, and 35 % need work in both areas. These events are not mutually exclusive because some students need both.

$$P(\text{needing maths and needing English}) = 0.35$$

What is the probability that a student selected at random needs maths or needs English?

Since the events are not mutually exclusive and the 35% appear in both maths and English, this sum needs to be subtracted, the formula is:

$$P(\text{needs maths or needs English}) =$$
$$P(\text{needs maths}) + P(\text{needs English}) - P(\text{needs maths and English})$$
$$= 0.40 + 0.70 - 0.35 = 0.75$$

7.5.3 Independent and Dependent Events

Multiplication Rule

Let us use coin tossing as an example. A coin is tossed twice; what are the probabilities of 0, 1, 2 heads? This is a compound event, made up of several simple events.

We can use the multiplication rule. This is associated with an 'and' statement, such as the probability of a head on the first toss and a head on the second.

$$P(H1 \text{ and } H2) = P(H1 \cap H2) = P(H1) * P(H2|H1) = 0.5 * 0.5 = 0.25.$$

This thus assumes that the events are independent, i.e. that the result of the first toss has no effect upon the second. This is true in this case.

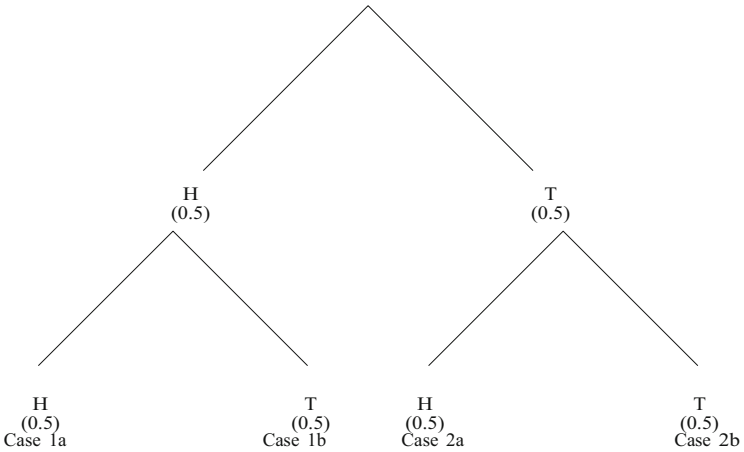
The events are independent if the outcome of one event has no effect on a subsequent outcome. For example, in a dice problem, the outcome of a 3 on the first die does not have any effect on the probability of getting a 3 on the second die, because these two events are independent. Suppose you are going to throw two fair dice. What is the probability of getting a 4 on each die? Since these two events are independent we should use the following formula:

$$P(A \text{ and } B) = P(A)*P(B)$$

Hence $$P(A \text{ and } B) = (1/6)*(1/6) = 1/36$$

A clear way to express these probability situations is to use a tree diagram. The best way to learn how to make a tree diagram is to see one. The below figure shows a simple tree diagram for independent events. Suppose you toss a coin twice. For the first throw P(H) = 0.5 and P(T) = 0.5. For the second throw these probabilities are the same. From the diagram we can see that for two throws of a coin:

$$P(\text{Case 1a}) = P(H,H) = 0.5*0.5 = 0.25, \quad P(\text{Case 1b}) = P(H,T) = 0.5*0.5 = 0.25$$
$$P(\text{Case 2a}) = P(T,H) = 0.5*0.5 = 0.25, \quad P(\text{Case 2b}) = P(T,T) = 0.5*0.5 = 0.25$$



A counter example would be drawing two aces from the pack (without replacement):

$$P(A1 \text{ and } A2) = P(A1)*P(A2|A1) = 4/52*3/51 = 12/2704.$$

If events A and B are independent, then P(A|B) = P(A|~B) = P(A), and so P (A and B) = P(A) * P (B).

In a card problem where 52 cards in a deck and 4 of them are aces, the probability of an ace on the first card is 4/52. If the first card is ace then the probability of an ace in the second card is 3/51. Hence these two events are not independent.

This is quite often referred to as a conditional probability.

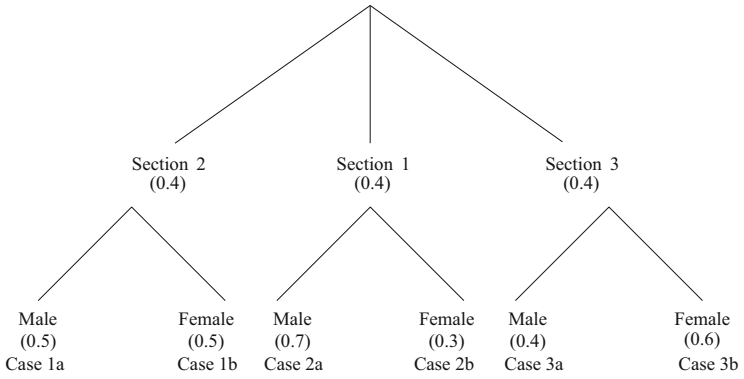
Example Q. What is the probability of drawing two aces from a well-shuffled deck if the first card is not replaced before the second card is drawn?

P(ace on 1st and ace on 2nd) = P(ace on 1st) * P(ace on 2nd, given ace on 1st) = (4/52) * (3/51) = 0.0045

A tree diagram example of independent events

Suppose we have three tutorial groups in a statistics course, section one of 17 male and 6 female students, section two of 20 male and 20 female students and section three of 8 male and 12 female students. In two stages we would like to select, first a section and then an individual. The probability of selecting a female student depends on which section you have selected in the first place. The tree diagram below shows these probabilities, which are derived from the numbers in each team.

$$\begin{aligned}
\text{Case1a} &= \text{P}\big(\text{Male}\big|\text{Section}\,2\big) = 0.20, \quad \text{Case1b} = \text{P}\big(\text{Female}\big|\text{Section}\,2\big) = 0.20,\\
\text{Case2a} &= \text{P}\big(\text{Male}\big|\text{Section}\,1\big) = 0.28, \quad \text{Case2b} = \text{P}\big(\text{Female}\big|\text{Section}\,1\big) = 0.12,\\
\text{Case3a} &= \text{P}\big(\text{male}\big|\text{Section}\,3\big) = 0.08, \quad \text{Case3b} = \text{P}\big(\text{Female}\big|\text{Section}\,3\big) = 0.12
\end{aligned}$$



How about if we have more than three sections, i.e. section 4 and 5 etc. Or in the case of coin tosses, If the tosses are more than 2, say 5 then how many branches would these diagrams have?

Tree diagrams can get out of hand. For eight tosses there are 256 end points () If there were 40 tosses, there would be an awful lot of branches to draw, and it would take just short of 70 000 years to draw, at the speed of one second for each branch! To derive these we use the idea of combinations and permutations.

7.6 Combinations and Permutations

Suppose a private university's board of the directors has 10 members. Three people, a president, vice president, and general secretary must be elected from the members. A group of candidates is a list of three people; one person for president listed first, one person for vice president listed second and one person for general secretary listed third. We need three names in each group and the order is also important. The size of the group is 10 (call this n) and the number of selected people in each group is 3 (call this r).

Each group of people consists of three candidates. For this we need to consider the number of **permutations** of 10 items arranged in groups of three. The formula is

$$P_{n,r} = \frac{n!}{(n-r)!} = P_{10,3} = \frac{10!}{(10-3)!} = \frac{3628800}{5040} = 720$$

There are 720 different possible groups of people

! = is read factorial
5! = 5 * 4 * 3 * 2 * 1 = 120

In the example above obviously the order is important. However, if the order is not important then in this case we would use **combinations.**

Let us say that the 3 members from the group of 10 on the board of directors of the university above will be selected to go to a convention. How many different groups of people are there? We need to look at the combinations.

The formula is:

$$C_{n,r} = \frac{n!}{\mathrm{r}!(n-r)!} = C_{10,3} = \frac{10!}{3!(10-3)} = 120$$

There are 120 different possible groups of 3 to go to the convention.

7.7 Expected Values

The concept of expected values are particularly important in economics. There is no doubt that you will be using these concepts in your microeconomics course.

Consider a simple game of chance. You throw a fair coin and if a head comes you win 1 Million TL and if it comes a tail you lose 2 million TL. If the game were repeated 10 times you would expect to win 10 times, that is 10 Million TL and be expected to lose 10 times, that is 20 Million TL. Your overall loss would be 10 Million TL or 500000 TL per game average. This average loss per game is called expected value.

The expected value of the winnings is: $1MTL\frac{1}{2} + -2MTL\frac{1}{2} = -500000TL$

Another example

You will toss a fair coin 10 times. If a head does not appear in 10 tosses you will win 1 Billion TL. The entry fee for the game is 5 Million TL. What is your expected gain? Would you play?

In a fair coin the probability of the chance of head or tail is 0.5. Here you have to consider the 10th probability which is 0.5^{10} the formula for this expected gain is then:

$$E(winnigs) = \left(0.5^{10}\right)*1BillionTL + \left(1 - 0.5^{10}\right)*(-5) = -4.0186$$

You have negative expected winnings. Playing this game depends on personal preferences. If you are a risk averse or risk neutral person then you wouldn't play but if you are a risk lover then you might play.

7.8 Bayes Theorem

Bayes' theorem is a way of looking at situations where the probabilities of their occurrence depend on the outcome of some previous event. These events are described as dependent. The final outcome is known and the probability that certain events occurred in the sequence where we are at the end of this sequence. This is called a posterior probability, which will almost always be different from the prior probability, which were known before the sequence began.

Bayes' Theorem itself can easily be derived from first principles.

$$\begin{aligned} &\mathrm{P}(\mathrm{A} \cap \mathrm{B}) = \mathrm{P}\big(\mathrm{A}\big|\mathrm{B}\big)*\mathrm{P}(\mathrm{B}) \\ &\rightarrow \mathrm{P}\big(\mathrm{A}\big|\mathrm{B}\big) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} \end{aligned} \tag{7.1}$$

This can be written as follows:

$$P(A \cap B) = P(B \cap A) = P(B \backslash A).\, PA \tag{7.2}$$

$$P(B) = P\big(B\big|A\big).\, P(A) + P(B \backslash \sim A).\, P(\sim A) \tag{7.3}$$

Putting Eqs. (7.2) and (7.3) into Eq. (7.1) gives Bayes' Theorem;

$$P\big(A\big|B\big) = \frac{P\big(B\big|A\big).P(A)}{P\big(B\big|A\big).P(A) + P\big(B\big| \sim A\big)P(\sim A)}$$

In Bayesian Language:
P(B|A)
$\mathrm{P}(\mathrm{B}|\sim \mathrm{A}) \Rightarrow$ Likelihood
P (A | B) $\Rightarrow$ Posterior odds
P (A) and P(B) $\Rightarrow$ Prior odds (odds = probability)
Thus

$$\text{Posterior Odds} = \frac{\mathit{Likelihood}*\text{Prior}P}{\sum(\mathit{Likelihood}*\text{Prior}P)}$$

So in general:

$$P(E_i|S) = \frac{P(S|E_i).P(E_i)}{\sum [P(S|E_i).P(E_i)]}$$

Example Let us consider the HIV example. The question was: What is the probability of being HIV positive when a positive test result has been received?

The given probabilities are:

The probability of being HIV positive is:

P (A) = HIV + 0.005
P (B) = positive test result.

Hence the probability of receiving a positive test result when probability of being HIV positive at first selection is: P (B | A) = 0.99.

The probability of receiving a positive test result when probability of not being HIV positive at first selection is: P (B | ~ A) = 0.01.

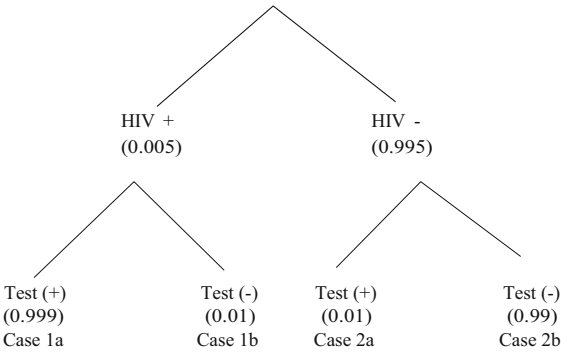
The probability of not being HIV positive: P (~ A) = 0.995

Formula:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B|A).P(A) + P(B|\sim A)P(\sim A)}$$

$$P(A|B) = \frac{0.99*0.005}{0.99*0.005 + 0.01*0.995} = 0.332$$

In a tree diagram:



P(case 1a) = 0.005*0.99 = 0.00495, P(case 2a) = 0.995*0.01 = 0.00995
P(Case 1b) = 0.005*0.01 = 0.00005, P(Case 2b) = 0.995*0.99 = 0.98505

$$\text{Original Q} : P(\text{HIV} + |\text{TEST}(+)) = \frac{0.00495}{0.00495 + 0.00995} = 0.332$$

i.e. Once that test is positive, what is P(HIV +)?

We could also ask;

If the test is negative, what is P(HIV −)?

$$P = \frac{0.98503}{0.00005 + 0.98503} = 0.9999$$

How many will Test (+)?

$$0.00995 \ + \ 0.00495 \ = \ 0.0149 \ = \ 1.49\%$$

How many will test (−)?

$$1 - 0.0149 \ = \ 0.9851 \ = \ 98.51\%$$

7.9 A Review of This Chapter

In this chapter we have examined probability theory. We began by discussing the basic concepts and definitions of probability. After examining the probability laws and rules, combinations and permutations were reviewed. Finally Bayes' theorem was examined with an example.

7.9.1 Review Problems for Probability

7.1. What are the probabilities of the following events:
- (a) drawing an ace at random from a pack of cards
- (b) drawing a spade at random from a pack of cards
- (c) obtaining a five on a roll of a die.
- (d) obtaining two fives with two throws of a fair die.
- (e) drawing two aces from a well-shuffled deck if the first card is not replaced before the second card is drawn.

7.2. When Kamil Bey drives from Istanbul to Ankara, the probability is 0.75 that at any given time he is going faster than the speed limit. If there are two radar traps between Istanbul and Ankara, what is the probability that Kamil Bey will be caught at least once (assume independence)?

7.3. A coin is tossed 10 million times. Let H be the number of heads obtained. Identify the following statements as either true or false:
- (i) it is very probable that H is very close (i.e. within a few thousand or so) to 5 million
- (ii) it is very probable that H/10,000,000 is very close to 0.5 (e.g. between 0.49 and 0.51)
- (iii) since heads and tails fall completely randomly and independently, all possible values of H are equally likely.
- (iv) heads and tails fall about equally often, so if the first ten tosses fall heads the eleventh is more likely to fall tails.

7.4. Sibel is 33, unmarried and assertive. She is a graduate in Political Science, and was involved in student union activities, anti-discrimination campaigns, etc. Which of the following is more probable in your opinion?
(a) Sibel is a bank clerk,
(b) Sibel is a bank clerk and is active in the feminist movement.

7.5. Which of the following events are independent?
(i) successive tosses of a fair coin
(ii) successive tosses of a biased coin
(iii) earthquake on successive days
(iv) earthquake on the 17th of August and on the first of September in Istanbul.
(v) a person smoking cigarettes.

How is the answer to (v) reflected in health insurance premiums?

7.6. A two-engine sailing boat can sail as long as at least one of its engines work and it's sails function. A one-engine boat sails as long as at least its sails works and there is some wind. The probability of an engine failure and no wind is 0.005. Would you feel happier in a two or one engine boat, and why? What assumption are you making in your calculation?

7.7. You toss a coin 15 times If a head does not appear in 15 tosses you win 2 Billion TL. The entry fee for the game is 10 Million TL. What are your expected winnings? Would you play?

7.8. What is the probability of inflation, I, or recession, R, if the probability of inflation is 0.6, the probability of recession is 0.4, and the probability of inflation and recession is 0.12.

Chapter 8
Probability Distributions

8.1 Introduction

In the previous chapter we discussed some basic probability concepts. In this chapter we will extend these ideas by looking at probability distributions. Probability distributions represent a pooling of knowledge about 'common' situations and they are relative frequency distributions.

In this chapter, we will look at the main three probability distributions, in more detail, namely the **Binomial distribution, the Poisson distribution and the Normal distribution.** These probability distributions are chosen because they occur more often and are well known.

A probability distribution lists, in some form, all the possible outcomes of an event and the probability associated with each one.

Binomial

Probability of number of successes (a particular outcome say H) in a series of trials.

Poisson

Probability of number of successes under certain circumstances. It is similar to the Binomial distribution.

Both the Poisson and Binomial distributions use discrete data.

Normal

A widely used probability distribution which uses continuous data.

8.2 What Are Probability Distributions

We have so far looked at ways of presenting data. A useful diagram we know is the Histogram which shows relative frequencies. In the last lecture our concept of probability was 'frequentist'. Probability distributions are about relative

D. Özdemir, *Applied Statistics for Economics and Business*,
DOI 10.1007/978-3-319-26497-4_8

frequencies. A probability distribution lists, in some form, all the possible outcomes and the probability associated with each one.

There are 3 steps for this;

- List the possible events (sample space)
- Calculate the probability of events.
- Present these probabilities in a suitable table or diagram.

Example Tossing a coin. e.g. Number of head in two tosses;

$$P(0) = 0.25$$

Number of heads	0	1	2
P	0.25	0.5	0.25

Figure 8.1. Shows these probabilities on a bar chart.

Similarly, the probable number of heads for 8 tosses (Fig. 8.2);

We have shown the distribution pictures but we do not know how to calculate these numbers yet.

For this purpose we shall explain the concept of a random variable.

8.2.1 Random Variable

It is one whose outcome or value is the result of chance and is therefore unpredictable, although the range of possible outcomes and the probability of each outcome may be known. i.e. taking the whole population is not a random selection because we are assuming population does not change.

8.3 The Binomial Distribution

In our previous chapter we discussed combination and permutations. Let us remind ourselves of these formulae;

The combinations formula is:

$$C_{n,r} = \frac{n!}{r!(n-r)!} \quad \text{(the order of 'r' is not important)}$$

Permutation;

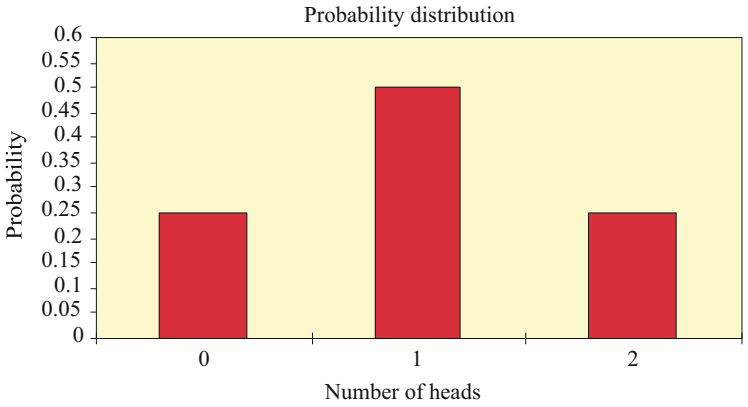


Fig. 8.1 Number of heads in two tosses

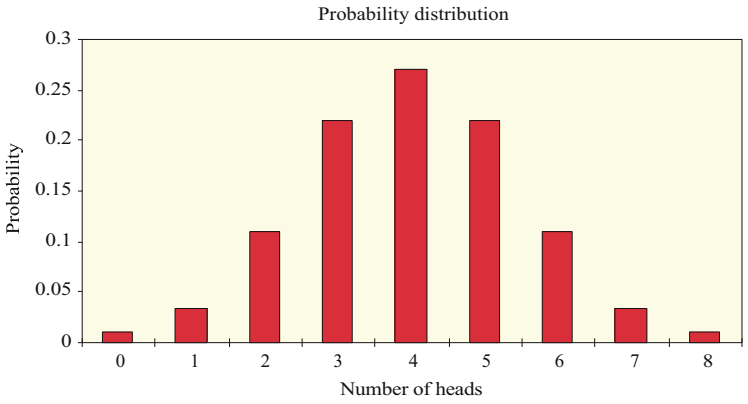


Fig. 8.2 Number of heads in eight tosses

$$P_{n,r} = \frac{n!}{(n-r)!'} \quad (\text{the order of 'r' is important})$$

One of the simplest distributions which a random variable can have is the Binomial distribution.

Say: A series of trials has the following characteristics;

1. There are a number (n) of trials.
2. Each trial has two possible outcome, 'success' (with probability P) and 'failure' (probability 1-P) = q and outcomes independent between trials.
3. The outcomes are mutually exclusive. (They cannot occur together)
4. The probability P does not change between trials.

Let us consider 8 tosses of a coin, which fulfills the above characteristics. The formula for the Binomial distribution is;

$$\mathrm{B(n,P)} = \mathrm{P(`r'\,success}\,in\,\mathrm{n\,times)}.$$
$$B(n,P) = {}^{n}C_{r}P^{r}q^{n-r}$$

Thus the formula for the binomial distribution is:

$$B(n,P) = \frac{n!}{r!(n-r)!}Prq^{n-r}$$

Examples 4. head out of 8 tosses:

$$B(n,P) = \frac{8!}{4!(4)!}0.5^{4}0.5^{4} = 0.273$$

3 head out of 8 tosses:

$$B(n,P) = \frac{8!}{3!(5)!}0.5^{3}0.5^{5} = 0.219$$

2 head out of 8 tosses:

$$B(n,P) = \frac{8!}{2!(6)!}0.5^{8} = 0.109$$

1 head out of 8 tosses:

$$B(n,P) = \frac{8!}{1!(7)!}0.5^{8} = 0.031$$

0 head out of 8 tosses:

$$B(n,P) = \frac{8!}{0!(8)!}0.5^{8} = 0.00000256$$

The shape of the binomial distribution varies with P and n (Figs. 8.3, 8.4, and 8.5).

As we examined in the diagrams above the distribution is symmetric if P = 0.5, If P = 0.1 then the distribution is skewed to the right and if P = 0.7 then the distribution is skewed to the left. If N is increasing than the distribution is flatter and broader.

Mean: μ = nP (i.e. n = 7 and P = 0.5 so the mean for 7 tosses is 3.5)

Variance is:

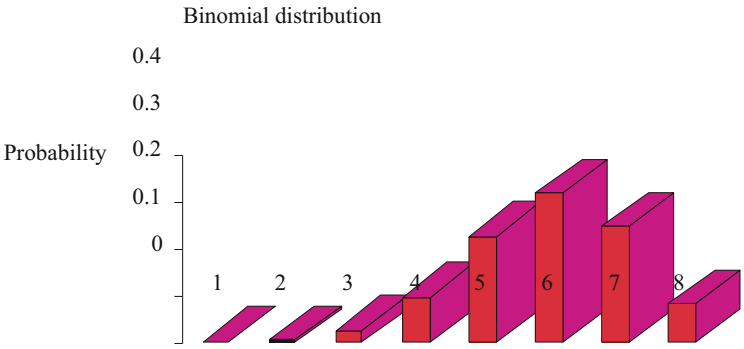


Fig. 8.3 Probability is 0.70 and skewed to left. ($\mu = P * n = 4.9$)

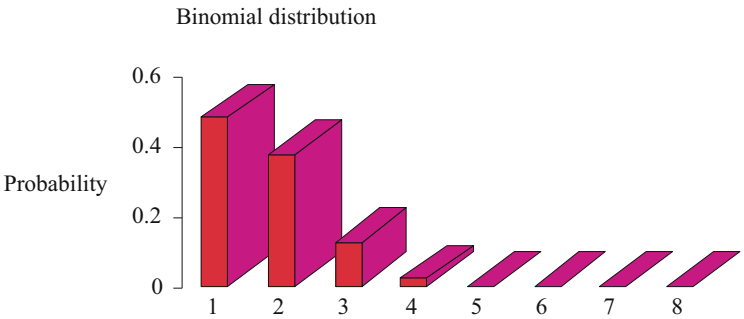


Fig. 8.4 Probability is 0.10 and skewed to right. ($\mu = P * n = 0.7$)

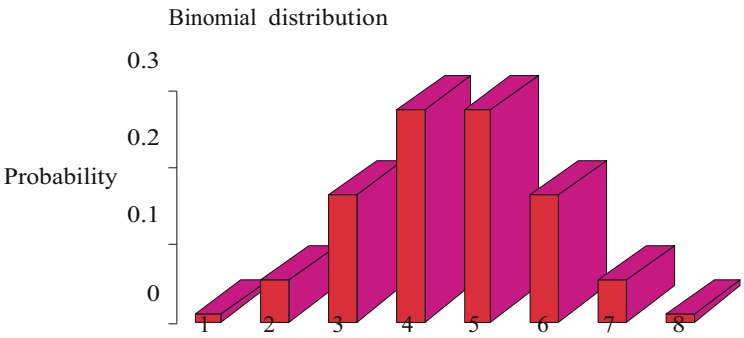


Fig. 8.5 Probability is 0.50 and the distribution is symmetric. ($\mu = P * n = 3.5$)

$$\sigma^2 = npq$$

For the standard deviation:

$$\sigma = \sqrt{npq}$$

8.4 The Poisson Distribution

Poisson distribution is a discrete probability distribution which is close to binomial distribution.

Binomial distribution can be approximated instead by a Poisson distribution. Poisson distribution would be appropriate if:

(a) Trial number (n) is large.
(b) The probability of success is very small.

Formula:

$$P\,(r\,success) = \frac{e^{-\mu}\mu^{r}}{r!}$$

where e = 2.7183 and μ = mean of the distribution.

Example The Turk Cellecom sends out 10 000 invoices. An average of 5 are returned with errors.

What is the probability that exceeds 5 units be returned in a given month?

Answer:

$$\mathrm{r} = 5 \Rightarrow \mu = 5$$
$$P(5) = \frac{e^{-5}5^{5}}{5!} = \frac{(0.0067)(3127)}{120} = 0.1755$$

For example it is very useful to calculate probability of accidents:

Say that in Istanbul there are 5000 people who commute to work by car. On average, they have 0.5 no of accidents a day. What is the probability of 3 accidents?

$$P(3) = \frac{e^{-0.5}5^{3}}{3!} = \frac{0.0758}{6} = 0.012$$

In general we use the Poisson distribution when

(a) Events are independent
(b) The probability of an event happening in an interval is proportional to that interval.
(c) An infinite number of events should be possible in an interval.

Mean: μ = np

Variance is:

$$\sigma^2 = np$$

For the standard deviation:

$$\sigma = \sqrt{np}$$

Slope and position of probability distribution defined by p. As μ increases distribution becomes more symmetrical.

8.5 The Normal Distribution

Normal distribution can be used to describe continuous data. It is the most widely used probability distributions of all (Fig. 8.6).

The normal probability distribution;

- is continuous
- is symmetrical about the mean value μ
- has mean, median and mode all equal
- has the total area under the curve equal to 1
- in theory the curve extends to plus and minus infinity on the x axis.

Many natural phenomena follow this distribution e.g. the height of trees, daily temperature etc.

Mathematical form;

If x is normally distributed and a random variable which is normally distributed its probability distribution is given by this formula:

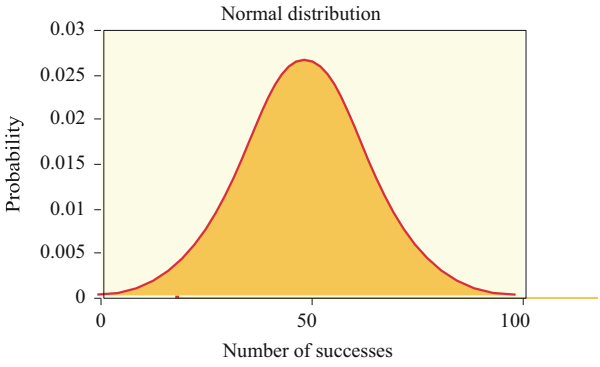


Fig. 8.6 Normal probability distribution

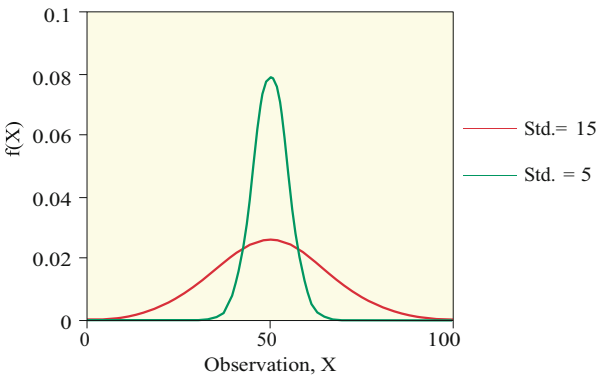


Fig. 8.7 Normal distribution with the same mean but two different standard deviations

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

Where;

π and e are constants (3.14) and (2.718)
$\sigma =$ Standard deviation $= \sqrt{\sigma^2}$
$\mu =$ mean value and x is value of interval.
$x \sim N(\mu, \sigma^2)$ (x is normally distributed with mean μ and variance σ^2

In Fig. 8.7. the mean is the same, 50, for both distributions but the greater the variance the flatter is the distribution. The standard deviation of one of the curves is 15 and the other is 5. A smaller standard deviation or a smaller variance gives a narrower distribution.

Example A manufacturer produces Turkish delight which come in boxes. The mean weight of a box is 500 g. but there will always be small variations in the weight of each box. We want to know what percentage of boxes exceed a certain weight (inefficient), say 505 g. Area under the curve is given as p;

$$\text{P(weight} > 505) + P(\text{weight} \leq 505) = 1$$
$$(\text{The total area is equal to one})$$

There are three ways of finding the area under the tail of the distribution. These are

- Integrating as a summation; Evaluate the definite integral of the curve between 505.0 and infinity (which would be very difficult)
- Use a computer (There could still be some problems)
- We could look up values in standard tables. (the best option)

Normal distribution tables (based on a Z value) are calculated from mean of a distribution and its standard deviation. Z is a number of standard deviations by

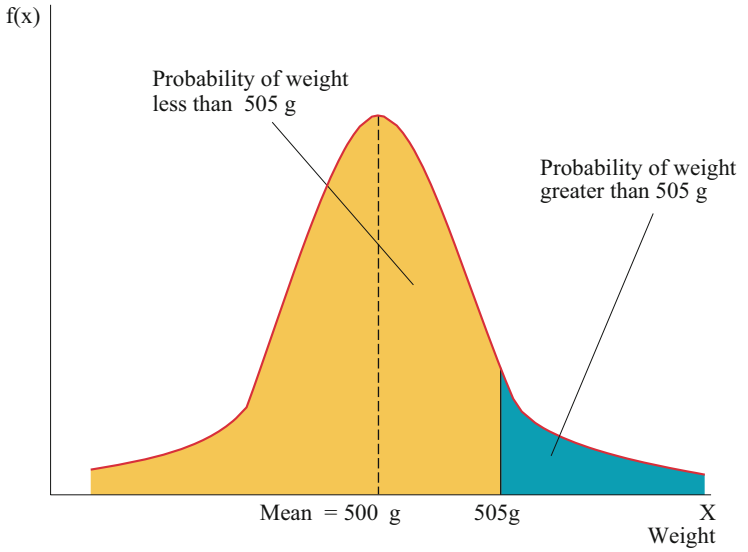


Fig. 8.8 Distribution of weights of Turkish delight boxes

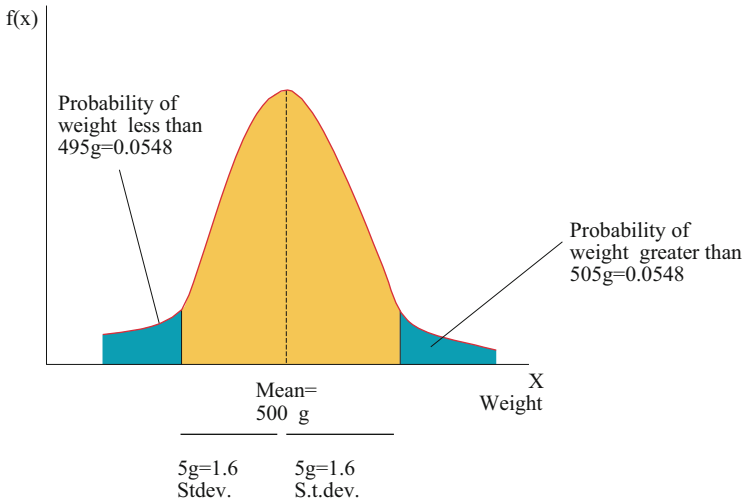


Fig. 8.9 Symmetrical distribution of weights of Turkish delight boxes

which a point is away from the mean, and tables show the probability that a value greater than this will occur (Figs. 8.8 and 8.9).

For the Turkish Delight example: the mean weight is 500 g. Assume that the standard deviation is 3 g. To find the probability that a box's weight is greater than 505 g, we need the area in the tail of the probability distribution. To find this we

calculate the number of standard deviations that the point of interest (505 g) is away from the mean, the tables will give the corresponding probability.

Z: Number of standard deviations.

$$\mathrm{Z} = \frac{Value - Mean}{S.t.dev.} = \frac{x - \mu}{\sigma}$$
$$X \sim N(\mu, \sigma^2), Z \sim N(0, 1)$$

Hence for the example:

$$\mathrm{Z} = \frac{505 - 500}{\sigma} = \frac{5}{3} = 1.6$$

Look that up in the Normal distribution table (Z table) 1.6: 0.0548 $\mathbf{P} = \mathbf{0.0548.}$

8.6 The Sample Mean of a Normally Distributed Variable

An important normally distributed variable is the sample mean, $(\overline{x})$, because we often use the sample mean to tell us something about an associated population. Firstly the distribution should be known. From our previous argument the sample mean was:

$$\overline{x} = \frac{1}{n}(x_1 + x_2 + \ldots\ldots\ldots\ldots x_n)$$

In this formula each observation is a normally distributed random variable and it has mean μ and the standard deviation is σ.

A linear combination of the random variables, as long as normally distributed and independent are itself normally distributed.

We also need to define the parameters of the mean and the variance:

The mean:

$$E(\overline{x}) = \frac{1}{n}(E(x_1) + E(x_2)\ldots\ldots\ldots\ldots)$$
$$= \frac{1}{n}(\mu + \mu + \mu\ldots\ldots\ldots\ldots) = \frac{1}{n}n\mu = \mu$$

The variance:

$$V(\overline{X}) = \frac{1}{n^2}(V(x_1) + V(x_2) + \ldots\ldots\ldots\ldots)$$
$$= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Putting the argument together we reach:

$$\overline{X} - N\left(\mu, \frac{\sigma^2}{n}\right)$$

So, the same mean as the population. The sample mean is taken from a Normally distributed population with mean μ and variance has a Normal sampling distribution with mean μ and variance σ^2/n.

This theorem has important consequences. It means that any sample taken from a normally distributed population would also have same characteristics as that population.

Samples are taken from a population. When samples are Normally distributed with the sample mean centered around the population mean, μ then the difference between population distribution and the sample mean distributions are shown in the diagram below (Fig. 8.10).

The larger the sample size the closer the sample mean is to the population mean. However, if the sample size is very small then this may fail. If the sample size is reasonably large then the sample mean would be closer to the population mean, μ. This is the reason why the variance of the distribution of the sample mean is (σ^2/n). This can solve a number of problems, an example of which is given below;

Example What is the probability that a random sample of 10 female students will have a mean height greater than 172 cm. The height of all female students is known to be normally distributed with mean $\mu = 167$ cm and variance, $\sigma^2 = 80$.

For population: $x \sim N\ (\mu,\ \sigma^2)$: If the given numbers are inserted in to this statement we have $x \sim N\ (167, 80)$.

For the sample mean: $\overline{x} \sim N(\mu, \sigma^2/n)$ Inserting the given numbers:

$\overline{x} \sim N\left(167, \frac{80}{10}\right)$ The Fig. 8.11 below shows this situation.

The answer of the question can be found after finding the shaded area on the figure. Firstly, the Z score needed to be calculated. For the Z score σ^2/n should be taken (not σ^2) because the distribution of the sample mean is used (not the population).

$$= \frac{\overline{x} - \mu}{\sigma^2/n} = \frac{172 - 167}{80/10} = 1.77$$

The next step is to look up the Normal distribution table for the Z value, which is $Z = 0.0384$. This value shows us the shaded area in the diagram above and it means 3.84 % of the sample exceeds or is equal to the 172 cm height.

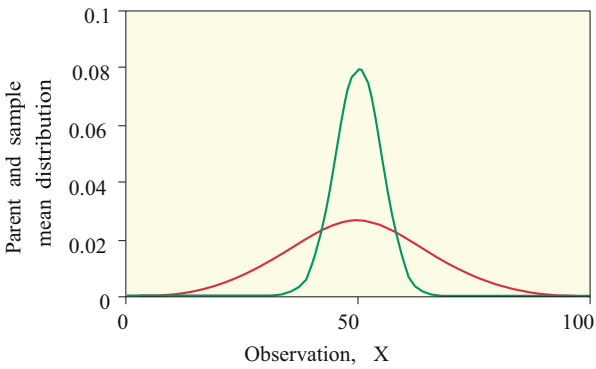


Fig. 8.10 The parent distribution and the distribution of sample mean

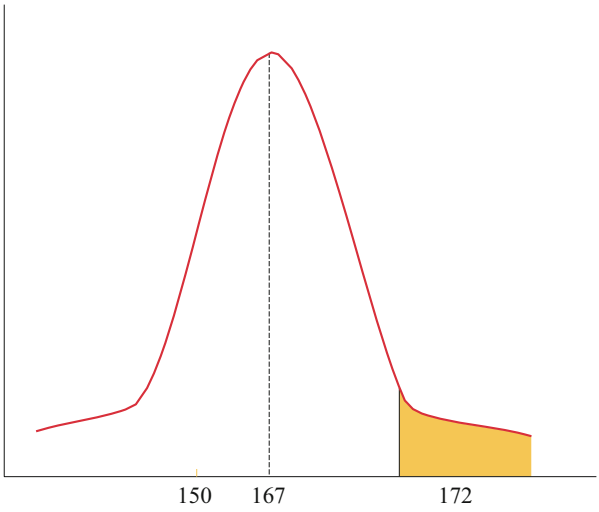


Fig. 8.11 Proportion of sample means being greater than 172

8.7 Sampling from Non-Normal Populations

The previous theorem and examples were examined under the assumption of a Normal Distribution. However, there are a number of populations which are not normally distributed. To be able to accommodate this condition we have to look at a different theorem. This is called Central Limit Theorem.

8.7.1 Central Limit Theorem

In any selected sample from any normally distributed population, the distribution of the sample means is approximately a normal distribution. If the sample size increases then this approximation improves. Returning to our notation, $\overline{x}$ is the

sample mean selected from a normally distributed population which has mean μ and variance σ^2. This selected sample mean's distribution approaches a normal distribution as the sample size increases. Then the sample size has mean μ and variance σ^2/n.

Thus the distribution of $\bar{x}$ is:

$$\bar{x} \sim N\left(\mu, \sigma^2/n\right)$$

This theorem basically states that in any infinite sample size the distribution is normal and any finite sample size the distribution is approximately normal.

For practical purposes, if the population distribution is nearly normal than a smaller sample size may be enough but if the population distribution is not normal (skewed) then more than 25-30 observations is recommended.

Example Here we can return back to the Turkish income distribution data in chapter 4. The average income level was 314.623 (Million TL) and the variance 198841. We are assuming that the parent data is the population and we draw a 100 household sample from the original data. We all know that it is highly skewed (not normally distributed). What is the probability that the sample mean is greater than 350 Million TL?

For this question we can apply the Central Limit Theorem because the original data is not distributed normally and the sample size is higher than 30 (it is 100).

The distribution of the sample mean is: $\bar{x} \sim N(\mu, \sigma^2/n)$ and with the given values this would be: $\bar{x} \sim N(314.623, 198841/100)$. We need to find the area beyond 350 Million TL. For this purpose, first the Z score should be calculated.

$$Z = \frac{\bar{x} - \mu}{\sigma^2/n} = \frac{350 - 314.623}{198841/100} = 0.793$$

Looking at the standard normal table 0.793 corresponds to 0.2148 (the area of the tail). This means that there is a probability of 21.48 % finding a sample size 100 with a mean of 350 Million TL or greater. It is a quite high probability for a value which is so different from the mean 314. This is because the income distribution data has a very high dispersion.

This examination can be extended with the following question: What is the probability of the sample mean lying between 214.623 and 414.623 Million TL. This situation is explained in the figure below (Fig. 8.12).

In the figure above, the area I is equal to the area II Hence if we only find one of the areas, say II, then we would know the total shaded area.

For II calculation of Z score is as follows:

$$Z = \frac{414.623 - 314.623}{198841/100} = 2.24$$

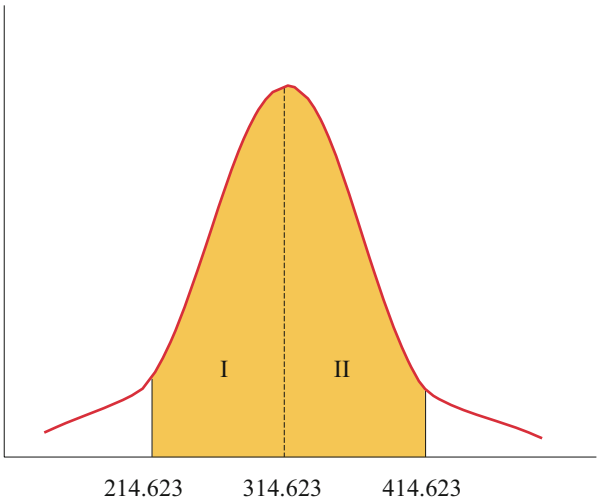


Fig. 8.12 Probability of the mean lying between 214.623 and 414.623

The number 2.24 corresponds to 0.0125 in the normal distribution table. That is the area on the upper tail. 1.25 % from the upper tail and 1.25 % from the lower tail 2.5 % is the total unshaded area. Hence the total shaded area I + II = 1 − 0.025 = 0.975 or the total shaded is 97.5 % of the distribution. Hence we can conclude that there is a 97.5 % probability of sample mean falling between the range of 214.623 and 414.623. This can also be shown as in the following form:

$$\Pr(214.623 \leq \bar{x} \leq 414.623) = 0.97.5$$

This could also be represented in terms of the formula:

$$\Pr\left(\mu - 2.24\sqrt{\sigma^2/n} \leq \bar{x} \leq \mu + 2.24\sqrt{\sigma^2/n} = 0.97.5\right.$$

8.8 Approximation of Binomial and Normal Distribution

Let us examine this approximation with an example: 80 students (n) take the course in Macroeconomic Theory. The result is pass or fail. The probability of passing is 70 % (P = 0.7 and (1-P) = 0.3). What is the probability of at least 50 students passing the course. For this question it is possible to use both Binomial and Normal distributions. We will first consider the binomial solution and then the Normal distribution solution.

8.8.1 The Use of Bionominal Distribution

$$r \sim B(n, P), \quad \text{Mean} = \text{nP}, \quad \text{Variance} = \text{nP}(1 - \text{P})$$

The important point in this approximation is that the probability should not be too close to zero, because the mean should be greater than 5. For binomial distribution we need to find the probability of 50 or above students passing.

Probability of 50 Students Passing

$$\Pr(50) = {}^{n}C_r P^r q^{n-r} = {}^{n}C_r 0.7^{50} 0.3^{30} = \mathbf{0.03285}$$

For the probability of 50 students or above passing we need to calculate 51, 52,.. up to 80 students. Calculating these and summing them gives the result of **0.941252**. This result tells us the probability of at least 50 students passing is 94 %.

8.8.2 The Use of Normal Distribution

The mean is nP and variance nP(1-P). The condition of the approximation of the binomial distribution to the normal distribution is that both the mean and the variance have to be greater than 5. In our above example both are greater than five hence there is no problem for approximation. Therefore,

$$r \sim N(nP, nP(1 - P)) = r \sim N(56,\ 16.8)$$

Another important adjustment is the fact that the binomial distribution uses discrete data and normal distribution uses continuous data. Hence it needs continuity correction. That is, in our example 50 in the binomial distribution is represented by the area under the normal distribution between 49.5 and 50.5 or similarly 80 is represented as being between 79.5 and 81.5. The area under the normal distribution to the right of 49.5 (not 50) needs to be calculated.

The Z score is:

$$Z = \frac{49.5 - 56}{16.8} = 1.59$$

After comparing with the normal distribution table this gives the area of 0.0559 or 5.59 % (The area of the tail). The result is different to the binomial calculation result but not too far away.

8.9 A Review of this Chapter

In this chapter we have examined some fundamentally important probability distributions such as the Binomial, Poisson and Normal distributions. We also considered sampling from non-normal populations.

8.9.1 Review Problems for Probability Distributions

8.1. Define what is meant by and give an example of

(a) a random variable,
(b) a discrete random variable, and
(c) a discrete probability distribution.
(d) What is the distinction between a probability distribution and a relative-frequency distribution?

8.2. (a) State the conditions required to apply the binomial distribution.
(b) What is the probability of 3 heads out of 7 tosses of a balanced coin?
(c) What is the probability of less than 3 heads out of 7 tosses of a balanced coin?
(d) Calculate and plot the probability distribution for a sample of 4 items taken at random from a production process known to produce 25 % defective items.

8.3. (a) What is the difference between the Binomial and Poisson distributions? Give some examples of when we can apply the Poisson distribution.
(b) Under what conditions can the Poisson distribution be used as an approximation to the binomial distribution? Why can this be useful?
(c) There are 5000 people who commute to work by car in Istanbul. On average, they have 5 number of accidents a day. What is the probability of having 2 accidents?

8.4. Past experience indicates that an average number of 10 customers per hour stop for petrol at the Bornova TP petrol station. What is the probability of:

(a) Four customers stopping in any hour?
(b) Four customers or less in any hour?
(c) What is the expected value, or mean, and standard deviation for this distribution?

8.5. (a) Define what is meant by a continuous variable and give some examples.
(b) Define what is meant by a continuous probability distribution.
(c) What are the standard normal and the normal distribution ?

8.6. Assume that family incomes are normally distributed $\mu = \$1000$ and $\sigma = \$200$.

What is the probability that a family picked at random will have an income: Between \$900 and \$1200? b) Below \$900? c) Above \$1200? d) Above \$2000?

8.9.2 Computing Practical for Probability Distributions

C.8.1. Obtain and Chart the Probability Distribution

The AkGaranty Bank offers a discount on credit card accounts paid by the 29th of the month. In the past, 70 % of accounts have been paid by this date. Accounts are sent out to 7 customers on a particular morning. Calculate the probability that of these customers the number who paid by the 29th of the month is: (a) 3 (b) 3 or less (c) 7 (d) 6 or more.

Use Excel and obtain and chart the probability distribution of the number of accounts paid by the 29th of the month for the AkGaranty and see how it alters if different values are used for p, the probability of an account being paid by the 29th.

Begin by entering the X values you require in the spreadsheet, as follows: To enter the list of X values, or any series, just enter the first two, select them and use the Fill handle to fill in the reminder of the series. If you fill more values than intended, select, those you don't want, press Delete and click OK to remove them.

Form a column of values of X!. To do this, click on the cell where the top X! value is to appear. Choose Formulas then Math&Trig. Scroll through those offered to you until you find FACT under Math&Trig. Click it so it appears in the formula bar, and click the number whose factorial you want so its cell reference appears in the function. Use the Fill handle to replicate the formula down the column.

Check the factorials are correctly calculated by entering a formula in a blank cell of the spreadsheet to calculate one of them using multiplication signs.

Form a column of (n-X)!. Use the factorials you have calculated to form a column of values of ${}^{n}C_{x}$.

Now discover the use of the COMBIN formula, which would calculate these values directly, and use it as a check on them. COMBIN is obtained in a similar way to that in which you obtained FACT. Its first argument should be the cell reference for n. Notice you will have to make this absolute if you want to replicate the formula.

When you have the first part of the formula correct, click on the second argument. You need to make this the X value by clicking on the cell containing X.

Enter p for the Binomial distribution in a cell, and use it in a formula to create values of p^x. To do this give a name to the probability cell.

Names

If you define a name for a cell, you can use that name in a formula and it is assumed to be an absolute cell address. Thus it may be convenient to name the cell which contains probability p, and you may call it 'prob'.

To do this, make active the cell you wish to name, then choose Formulas, Name Manager and type in the name you wish.

Create Names is an alternative method of naming cells, whereby labels you have entered in the spreadsheet are used as names for adjacent cells.

Find also a column of values of $(1 - p)^{n-X}$. Now form a column of binomial probabilities (Hint: Use the formula; Bprob $= {}^{n}C_{x} * p^{x} * (1 - p)^{n-X}$) and obtain a chart of the probability distribution.

To select two non-adjacent ranges to be used in your chart, select the first then press and hold CTRL while you select the next range.

Try altering the value of p and see how the shape of your binomial distribution changes.

Now discover how you can obtain values of the Binomial distribution directly. This requires the use of array function.

Begin by selecting a column of cells to contain the Binomial probabilities.

These should be empty cells in the same rows as each of the X values, to which they are to correspond. Paste the function BINOMDIST. You will see under more functions in the formulas bar that it requires four arguments. The first is the set of values which X can take, so select them, then click on the next argument. The second argument, called trials, is the value of n for the binomial distribution, so point to the cell where this is entered. Click on the third argument and insert the probability. The fourth argument tells Excel whether you want the separate probabilities that X is the specified number or less. For this argument you must type either true or false. We do not want a cumulative distribution, so type false.

DON'T HIT OK YET – keep reading!

To enter an array function: An array function is a function entered simultaneously into all the selected cells. Once an array function is in the formula bar and ready to enter in the spreadsheet, you hold CTRL and SHIFT and simultaneously press Enter. If you press Enter on its own, the formula will be entered in only one cell. What you should then do is to click on the formula in the formula bar. This makes it ready to edit or re- enter and you can try again to use CTRL + SHIFT + ENTER.

C.8.2. Normal Distribution

Open a new Excel Sheet. Type X in Cell A2 and Enter the number of successes from the zero through to 100 in cells A3:A103. Select the adjacent column (B) from B3:B103. Choose Functions, More Functions, Statistical then NORMDIST in Paste Function. Obtain two normal distribution curves on a same diagram with mean is 50. (Hint: Take two different standard deviations such as 5 and 15 for the same mean).

For Excel calculations use the similar steps as you did in C.8.1 BINOMDIST array function entry.

What is the effect of the variance on your distribution?

Confidence interval

Enter your data in the spreadsheet and paste functions from the statistical list to find the Average and Stdev of the data. (Note the STDEV function uses (n-1) as divisor, the STDEVP function uses n.)

The first argument of this function is the probability in two tails of the distribution, so using 0.05 as that argument will give you the t value for a 95 % confidence interval. (NB: The Confidence function is for finding confidence intervals for a Normal distribution).

Chapter 9
Estimation and Confidence Intervals

9.1 Introduction

In our first chapter we stated that statistics is the study of how to collect, organize, analyze, and interpret numerical data. In statistical inference we draw a sample of observations from a larger population. Estimation is the use of sample data in order to derive conclusions about the population. If we denote the Sample parameters as $\overline{x}$, S^2 then the Population Parameters are μ, σ^2 for the mean and variance.

There are two ways of presenting an estimate of a parameter: A point estimate and an interval estimate. An estimate of a population parameter given by a single number is called a point estimate, i.e. the average Turkish person drinks 4 glasses of tea in a day. Instead if we say that the average Turkish person drinks between 2 and 6 glasses of tea per day this is an interval estimate.

For estimation we need some estimators such as the sample mean, smallest sample observation, the first sample observation, median etc. But the question is: which one is the best estimator? In order to have some idea about which one is the best estimator we have two important criteria, bias and precision. If an estimator gives the correct answer on average then the estimator is unbiased, in another words an unbiased estimator does not systematically mislead the researcher from the correct value of the parameter (say μ). Technically, an unbiased estimators' expected value is equal to the estimated parameter; $E(\overline{x}) = \mu$

For example, if we use the smallest sample observation, which is smaller than the sample mean then its expected value must be less than the population mean (μ). So we can say that this estimator is biased downwards;

$$\text{Biased} = E(x_s) - \mu$$

Thus the sample mean and the first sample observation estimators are unbiased. However the smallest sample observation estimator is biased. We also need precision in our estimator. Unlike bias, precision is a relative concept and it involves

D. Özdemir, *Applied Statistics for Economics and Business*,
DOI 10.1007/978-3-319-26497-4_9

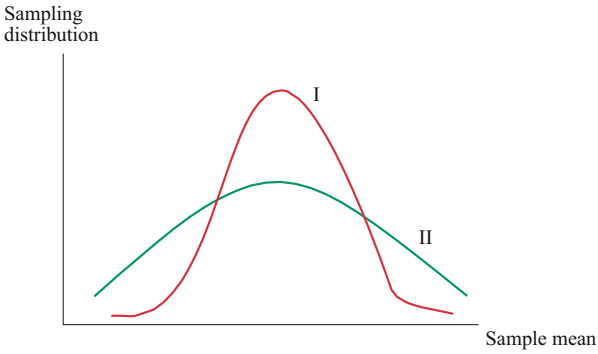


Fig. 9.1 Sampling distribution of two estimators

comparing one estimator to another. For example the sample mean is more precise then other estimators because taking single sample observation means that it is likely to be unrepresentative of the population as a whole, and thus leads to a poor estimate of population mean. The sample mean therefore is a more precise estimator of population mean. Precision therefore can be defined in terms of variance. If an estimator has a smaller variance compared to another estimator then the estimator with the smaller variance is more precise.

The probability distribution of the sample mean is; $\overline{x} \sim N(\mu, \sigma^2/n)$
The variance of the sample mean is; $V(\overline{X}) = \sigma^2/n$

The larger the sample size, the smaller the variance of the sample mean will be, thus the estimator becomes more precise. We can argue that large samples give better estimates than small samples. Figure 9.1 shows the difference on the two estimators' distribution, a sample mean (I) and a single observation (II).

What we have argued is that the method of estimation differs according to the size of the sample. Thus we will deal with these separately. If sample size is less then 25 then it is considered to be a small sample size, if it is greater than 25 it is large sample size. We will now examine the case of large sample size.

In this chapter we will divide the estimation topic into the two parts. The first part will examine the estimation with large samples. The second part will examine the estimation with small samples: the *t* distribution. Both parts will concern the estimating mean and the difference between two means. In the first part we will also focus on the confidence interval, estimating a proportion and the difference between two proportions.

9.2 Estimation with Large Samples

Using a sample mean to estimate a population mean can be difficult even if there is a large sample size, because it is not exactly possible to say how close the sample mean $(\overline{x})$ is to the population mean (μ) when μ is unknown. Therefore the error of estimate is $(\overline{x} - \mu)$ is unknown.

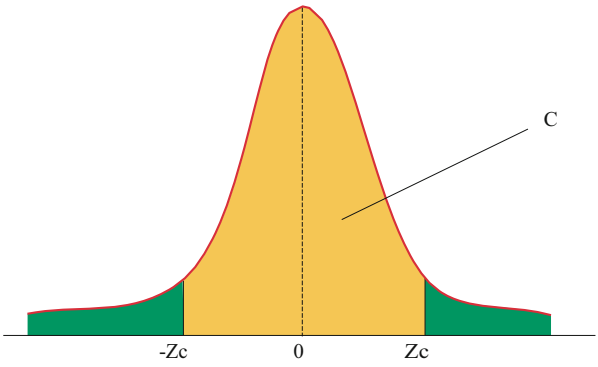


Fig. 9.2 A standard normal distribution curve with the confidence level (the area c) and corresponding critical value Z

9.2.1 Estimating a Mean

Confidence Interval

When we use as $\overline{x}$ a point estimate of μ we need a knowledge of probability to give us an idea of the size of the error of the estimate. For this purpose we need to know about confidence levels.

The area under the curve is equal to one for a normal distribution. For confidence levels therefore we could choose any value between 0 and 1. The confidence level is usually taken as 0.90, 0.95 or 0.99. Each level is a point such as Zc in the Fig. 9.2. Thus the value Zc is called critical value for a confidence level c. The area under the curve between -Zc to Zc is the probability that the standardized normal variable Z lies in that interval;

$$P(-Zc < Z < Zc) = C \tag{1}$$

An estimate is not very valuable unless we have some kind of measure of how 'good' it is. With the aid of probability we can now have an idea about the size of the error of an estimation of the population mean from sample mean. In our last lecture we outlined the central limit theorem which says if the sample size is large, then the sample mean has a distribution approximately normal with $\mu_{\overline{x}} = \mu$.

The standard deviation is $\sigma_{\overline{x}} = \sigma/n$. With the information we have we can now have a new probability statement.

Equation (1) clarifies the size of Z but we need to know the link between Z and $\overline{x} - \mu$. The standard Z formula is given by;

$$Z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

Substitution of this value into Eq. (1) gives;

$$P\left(-Zc < \frac{\overline{x} - \mu}{\sigma/n} < Zc\right) = C$$

Multiply all parts by σ/ n gives;

$$P\big(-Zc\frac{\sigma}{n} < \overline{x} - \mu < Zc\frac{\sigma}{n} = C$$

$$E = Zc\sigma/n \Rightarrow \text{maximal error tolerance.}$$

and could be read as; The probability is C that our point estimate $\overline{x}$ is within a distance $+Zc(\sigma / n)$ of the population mean.

$$\begin{gathered}\text{Thus } P(-E < \overline{x} - \mu < E) = C \\ \text{Or } -E < \overline{x} - \mu < E \\ \text{Hence } \overline{x} - E < \mu < \overline{x} + E \\ \text{Or } P(\overline{x} - E < \mu < \overline{x} + E) = C\end{gathered} \tag{3}$$

Equation (3) can be read as: 'there is a chance of C that the population mean μ lies in the interval from $\overline{x} - E$ to $x + E$' which is called the **confidence interval** for μ. For large samples (n > 25) the confidence interval for population mean μ is;

$$\overline{x} - E < \mu < \overline{x} + E$$
$$\text{where } \; E \approx Zc\frac{s}{n} \Rightarrow \sigma = 8$$

S = Sample standard deviation.
C = Confidence level (0 < C < 1)
Zc = Critical value for confidence level C
n = Sample size. n > 25

Figure 9.3 represents this relationship: The probability is C that $\overline{x}$ is within $\pm Z c\frac{S}{\sqrt{n}}$ of the true population mean μ.

Example Özlem likes jogging 3 days of a week. She prefers to jog 3 miles. For her 95 times, the mean was $\overline{x} = 24$ minutes and the standard deviation was S = 2.30 minutes. Let μ be the mean jogging time for the entire distribution of Özlem's 3 miles running times over the past several years. How can we find a 0.99 confidence interval for μ?

(a) What is the table value of Z for 0.99? ($Z_{0.99}$)?
(b) What can we use for σ ? (sample size is large)
(c) What is the value of? $Zc\frac{\sigma}{\sqrt{n}}$
(d) Determine the confidence interval level for μ.

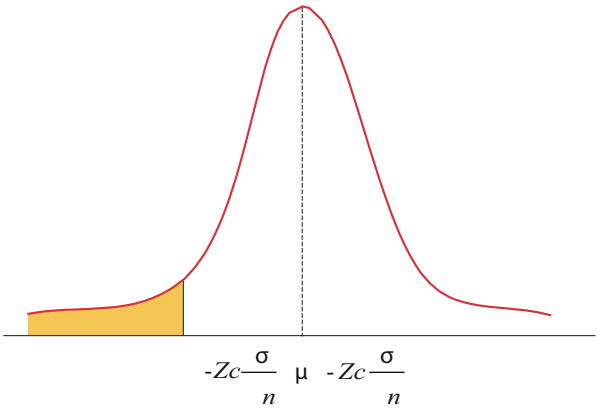


Fig. 9.3 Distribution of sample means

Answer

(a) From the Z table (Normal distribution) we can see that both tails are 0.01. Hence $0.01/2 = 0.005$. The nearest corresponding (lower 0.0049) number to that (in the table) is 2.58. Different textbooks provides the different areas of the parts of the distribution curve. But since the whole area under the curve is equal to 1 then it is not difficult to obtain the necessary area value. For example some statistics texts gives the area C instead of taking the area outside C in Fig. 9.3. An extract of a Z table is provided below:

An extract of the normal distribution table.

Z	**0.00**	**.....**	**.....**	**0.08**
. . .	. . .	. . .	.....	....
. . ..	.....	.....	.....	....
. . ..	. . ..	. . ..	. . ..	....
2.5	. . ..	.....	.....	0.0049

(b) This is given in the question: $\sigma \approx S = 2.30$.

(c) The maximal error tolerance level is:

$$E = Zc\frac{\sigma}{n} \approx 2.58\left(\frac{2.30}{95}\right) = 0.61$$

(d) The confidence interval; $\overline{x} - E \approx 24 - 0.61 = 23.39$ and $\overline{x} + E \approx 24 + 0.61 = 24.61$

$\overline{x} - E < \mu < \overline{x} + E = 23.39 < \mu < 24.61$

We are 99% confident that Özlem runs 3 miles between 23.39 minutes and 24.61 minutes.

9.2.2 Estimating a Proportion

Rather than the average, we are sometimes interested in the proportion of a population with a particular characteristics. i.e. car ownership.

Example A sample of 100 Yasar University Students is observed and 60% of them own their own cars.

n = 100, p = 0.60.

We would like to estimate the population proportion π. The sampling distribution of p is;

$$P \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right.$$

i.e. normal approximation of binomial distribution.

For a 95% confidence interval z : 1.96. (For 95% it is 1.96 but for 99% it is 2.58, different confidence levels would have different values).

Since the value of π is unknown the confidence interval cannot yet be calculated so the sample value (p) has to be used instead;

$$\left[p - 1.96\sqrt{\frac{0.6(1-0.6)}{100}}, p + 1.96\sqrt{\frac{0.6(1-0.6)}{100}}\right] = [0.504, 0.696]$$

So we say that we are 95% confident that the true proportion of Yaşar students' car ownership lies between 50.4% and 69.6%.

9.2.3 Difference Between Two Means

This is about an estimate for $\mu_1 - \mu_2$ in the form of both points and interval estimates.

Example Take as an example the servicing costs of two cars Meugeot and Siat. It is an estimate that obtains the true difference between the two cars average servicing costs.

$$A = n_1 = 100 \quad \bar{x}_1 = 105 \quad S_1 = 20$$
$$A = n_2 = 64 \quad \bar{x}_2 = 99 \quad S_2 = 14.5$$

$$\text{If } \bar{x}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \quad \textit{and} \quad \bar{x}_2 \sim N\left(\mu_2 \frac{\sigma_2^2}{n_2}\right)$$

The formula for the difference between the two cars' average servicing cost:

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Substitute the existing numbers: $\bar{x}_1 + \bar{x}_2 = 105 - 99 = 6$

$$\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} = \frac{20^2}{100} + \frac{14.5^2}{64} = 7.285$$

Construction of a 95% confidence interval:

$$[6 - 1.96 \quad 7.285, \quad 6 + 1.96 \quad 7.285,] = [0.71, 11.29]$$

The estimate is that Siats' average service cost is between 11.29 and 0.71 percentage points below Meugeots.

9.2.4 Difference Between Two Proportions

Example A survey of 90 People in Turkey showed that 26 owned mobile phones. A similar survey of 60 Britons showed 12 with mobile phones. Are mobile phones more widespread in Turkey than Britain?

In this example the aim is to estimate the difference between the two population proportions. We need to find the difference of the sample proportion. Similarly to the derivation of the sample mean, the probability distribution is;

$$p_1 - p_2 \sim N\big(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$$

$$p_1 = \frac{26}{90} = 0.29, \qquad p_2 = \frac{12}{60} = 0.2$$

Thus p_1 - $p_2 = 0.0875$ or 8.75%

For 95% confidence interval:

$$\Big[0.29 - 0.2 - 1.96\sqrt{\frac{(0.29)(0.71)}{80} + \frac{(0.2)(0.8)}{50}}, 0.29$$

$$-0.2 + 1.96\sqrt{\frac{(0.29)(0.71)}{80} + \frac{(0.2)(0.8)}{50}} = [-0.06, 0, 24]$$

This result mean that we can be 95% sure that the percentage difference of mobile phone users for Turkey and Britain is between – 6% and 24%. Since the interval −0.06 to 0.24 is not all negative (nor all positive), we cannot say that

p_1 - $p_2 < 0$ or p_1 - $p_2 > 0$ This means that at the 95% confidence level we cannot conclude that $p_1 > p_2$ or $p_1 < p_2$. Some people use mobile phones in both countries but we cannot be 95% confident that there is any true difference at all between Turkey and Britain.

9.3 Estimation with Small Samples: t Distribution

This is based on normal distribution and considers a sample size of less than 25. If we have large sample size then we can approximate the population standard deviation σ by S, the sample standard deviation. Then we can use the central limit theorem to find the boundaries or the error of estimate and confidence interval for μ.

However if we have a sample size smaller than 25 then in order to avoid the error of replacing σ by S we need the error variable called Student's *t* variable and its distribution. The formula of the t variable is as follows:

$$t = \frac{\overline{x} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}$$

The above equation describes a random variable that has a t distribution with n-1 degrees of freedom. This t distribution approaches the standard normal distribution as the sample size gets larger. The t distribution is similar to the normal distribution - it is unimodal, symmetric, centered on zero, bell shaped and extends from minus infinity to plus infinity.

Unlike the normal distribution, it has larger variance and has one parameter rather than two; the degrees of freedom. d.f. = v (pronounced nu) = n − 1 (Fig. 9.4).

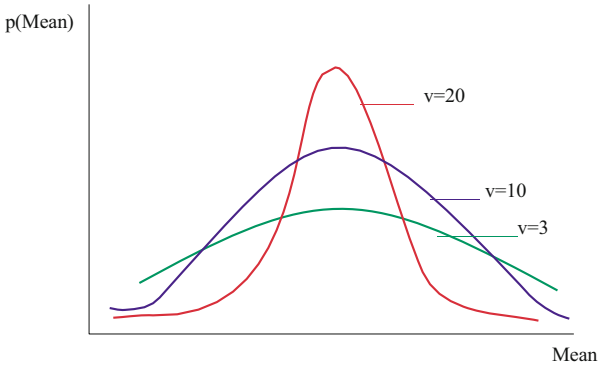


Fig. 9.4 The fewer the degrees of freedom, the more dispersed is the t distribution

9.3.1 Estimating Mean

One of the best way to demonstrate the estimation of the population mean is to use an example. As we mentioned earlier the weights are normally distributed.

Example TamSüt is a milk producing company. A sample of 20 boxes of milk showed an average weight of 1000g, with standard deviation 2.5. Estimate the true specific weight of TamSüt milk by constructing 95% confidence interval.

In this question the sample mean is an unbiased estimator of the population mean and is equal to 1000g. The sample size is small (i.e. less than 25) and the population variance is unknown hence we can use the t distribution.

Thus the information given in the question:

$\overline{x} = 1000g$
S = 2.5
n = 20.

$$\frac{\overline{x} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}$$

The 95% confidence interval;

$$\left[\overline{x} - t_{n-1}\sqrt{S^2/n}, \quad \overline{x} + t_{n-1}\sqrt{S^2/n}\right]$$

where t_{n-1} is the value of t which shows the tails of the distribution. We are concerned with 95% confidence which means that the tails are 5% (2.5 % in each tail) of the t distribution with 19 degrees of freedom. (Degrees of freedom = v, pronounced nu). The degrees of freedom is calculated as: d.f. = 20-1 = 19.

Looking at the t distribution table with 0.025, which indicates the area cut off in each tail and 19 degrees of freedom gives us $t_{19} = 2.093$

Hence the confidence interval is;

$$\left[1000 - 2.093\sqrt{2.5^2/20}, \quad 1000 + 2.093\sqrt{2.5^2/20}\right] = [999.346 \quad 1000.654]$$

We are 95% confident that the true specific weight of these milk boxes lies within the range of 999.346 and 1000.654.

9.3.2 Difference Between Two Means

When the population variances are unknown and the sample size is small then the *t* distribution can be used for the estimation of the difference between two means. Again we will use an example to demonstrate this estimation.

Example Let us say that we would like to examine the expenditure efficiency of the big city municipalities in Turkey. A sample of 18 Mother Path Party (MPP) controlled municipalities and 14 Republican Left Party (RLP) municipalities are examined in a Survey. It is observed that the RLP spends average 200 Million Turkish Liras (MTL) per taxpayer on administration and the Standard Deviation is 30MTL. The MPP, on average, spends 165 MTL per tax payer with standard deviation of 35MTL. We would like to find out if there is any true difference in administrative expenditures between these two party controlled municipalities.

Let us summarize the given information:

$\bar{x}_1 = 200$	$\bar{x}_1 = 185$
$S_1 = 30$	$S_1 = 35$
$n_1 = 18$	$n_1 = 14$

Here we would like to estimate the difference of the population means (μ_1 - μ_2). By using the sample means the unbiased point estimate is $\bar{x}_1 - \bar{x}_2 = -200 - 185 = 15$.

Since we have small sample and the population variances are unknown, we use t distribution confidence interval;

$$\left[\bar{x}_1 - \bar{x}_2 - t_v\sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_2}}, \qquad \bar{x}_1 - \bar{x}_2 + t_v\sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_2}}\right]$$

The pooled variance is;

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

The degrees of freedom is: $v = n_1 + n_2 - 2$.

Substitution of these values in to the pooled variance formula gives;

$$S^2 = \frac{(18 - 1)30^2 + (14 - 1)35^2}{18 + 14 - 2} = 100.8$$

The 95% co confidence interval;

$$\left[15 - 2.042\frac{1040.8}{18} + \frac{1040.8}{14}, \quad \left[15 - 2.042\frac{1040.8}{18} + \frac{1040.8}{14},\right] = [-8.48, 38.48]\right.$$

Since the interval ranges from negative to positive, the true difference is uncertain and the evidence is consistent with MPP municipalities spending more than the RLP municipalities but we cannot generalize this evidence for whole population. This is because we have a small sample and the variances are large for both samples.

9.4 A Summary of the Chapter

In this chapter we have examined the estimation topic for both large and small sample sizes. Estimation is divided into two different forms: the point estimate and the interval estimate. For both large and small samples we provided examples for estimating mean, the difference between two means, estimating proportions, the difference between two proportions and the confidence intervals.

9.4.1 Review Problems for Estimation and Confidence Intervals

9.1. (a) What is meant by statistical inference and estimation?
(b) What is meant by a point estimate, unbiased estimator and interval estimate?
(c) Under what conditions can we not use the normal distribution but can use the t distribution to find confidence intervals for the unknown population mean? What is meant by degrees of freedom?

9.2. The university entry exam is taken by 500 000 students. A random sample of 90 students is taken out of this population. The average mark of the sample students is 180 and the standard deviation for the entire population is 25. Construct the 95% confidence interval for the unknown population mean score.

9.3. A factory employs 1500 workers. As a factory employment manager you need to retire some workers. Instead of asking everybody about whether they would like to be retired, you selected a random sample of 120 workers and 80 of them prefer to be retired. Construct a 95% confidence interval for the proportion of all workers in the plant who prefer their own retirement.

9.4. (a) How can you find the t value for 10% of the area in each tail for 12 df?
(b) In what way are t values interpreted differently from z values?
(c) Find the t value for 5, 2.5, and 0.5% of the area within each tail for 12 df.
(d) Find the t value for 5, 2.5, and 0.5% of the area within each tail for a sample size, n, that is very large or infinite. How do these t values compare with their corresponding z values?

9.5. A car battery importer received a large shipment of batteries. These batteries known to have a normally distributed operating life. A random sample of n=15 car batteries with a mean operating life of 600days and a standard deviation of 50 days is picked from this shipment.
(a) Construct the 90% confidence interval for the unknown mean operating life of the entire shipment.
(b) Show your results of part a) on a distribution diagram.

Chapter 10
Hypothesis Testing

10.1 Introduction

In our last chapter we emphasized the importance of inferences about populations based on samples. The issues discussed in this chapter are very similar to estimation, but have some important subtle differences. There are two ways of drawing inferences about the parameters of a population. The first one is to make decisions concerning the value of the parameter and the second one is to actually estimate the value of the parameter. Estimating the value of parameters was examined in the last chapter. In this lecture we will examine the decisions concerning the value of a parameter: hypothesis testing.

10.2 The Nature of Hypotheses and Scientific Method

We have to be very careful about believing hypotheses to be 'true' or not, since if we are careless in our methods we will not make progress.

It is better if we take a cautious approach, and be skeptical about any so-called truths. For centuries it was believed that the sun revolved around the Earth rather than vice versa, until Galileo and Copernicus proved otherwise (at much personal cost!).

There are many different philosophical approaches to proving or disproving hypotheses. Karl Popper argued that we can only ever disprove hypotheses, never prove them. E.g. you cannot prove that all swans are white, but you could potentially disprove this statement if you found a black swan.

Thus the 'truth' consists only of those hypotheses which have not yet been proved false. In Poppers' view, the key issue was falsification. In fact, there is a view that says that one particular distinction between scientific and non-scientific statements is that scientific statements can in principle be falsified.

D. Özdemir, *Applied Statistics for Economics and Business*,
DOI 10.1007/978-3-319-26497-4_10

Language should reflect this caution. When hypothesis testing, we 'reject' or 'do not reject', rather than 'prove' or 'disprove'. We have the 'maintained' or 'null' hypothesis. All scientific truths should be subject to the most rigorous testing, and independently checked. Unfortunately economics often does not live up to these ideals.

10.3 The Null and Alternative Hypotheses

The **null hypotheses** is the one that is assumed to be true until it is proved otherwise, when the alternative hypothesis replaces it. Thus when carrying out hypothesis tests, these are done on the basis of the truth of the null hypothesis (Ho).

Thus the null hypothesis has to be precise enough for statistical testing to take place. For example, the hypothesis:

Average Turkish income is $3500 pa. is satisfactory, but not: Average income in Turkey is above $3500 since the latter is not precise enough to permit proper testing.

The alternative hypothesis has the advantage of being less precise. In the above case, the alternative hypothesis would be: Average income in Turkey is not $3500 pa. Sometimes the null is so precise and the alternative so vague, that the null is bound to be rejected. The above case is an example: it is very unlikely that average income in Turkey is **exactly** $3500.

This means that how we specify the test can influence the outcome of the test.

10.4 Two Types of Error

There could be two types of error for the testing procedure of a hypothesis.

Type I error: rejecting Ho when it is true.

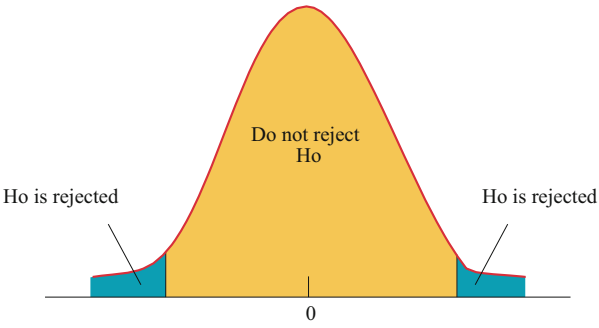
Type II error: do not reject Ho when it is false (i.e. rejecting H1 when it is true). We always have to attach a significance level to a test. This is very similar to a confidence interval. A five percent significance level corresponds to a 95 % confidence interval. In other words, if our test statistics (i.e. the number on which the test is based) falls within a 95 % confidence interval, we do not reject the null hypothesis.

However, if it falls into the remaining 5 % region, we reject it. (See Fig. 10.1)

Similarities to Estimation Hypothesis testing has some similarities to the estimation which we discussed in the previous chapter. These are:

1. The significance level plus the confidence level is equal to one.
2. We also distinguish between large samples and small samples. → Z versus t-distributions.

Fig. 10.1 Critical region



3. Very often the test statistics are based on z, $t = (x - \mu)/S$. where x and μ can be sample means, proportions, or differences between them.

10.5 Large Samples: Testing a Sample Mean Using the Normal Distribution

The best way to clarify the issue is to set an example.

Example The Beyazjean company knows that the average size of order received is 600 with the standard deviation of 250. The last 35 orders have averaged 550 units. Does this indicate a fall in their demand?

Ho: $\mu = 600 \rightarrow$ demand has not fallen

H_1: $\mu < 600 \rightarrow$ demand has fallen.

If H_1 is true, the firm might want to take action, e.g. reduce production levels. But we don't know which is true. Hence we could take the wrong action, e.g. lay off workers and reduce investment when demand has not fallen, or continue as before when demand has fallen. There is a cost associated with each type of error. Let us consider each type of errors within our example.

Type I Error Reject Ho when it's true. The demand actually was fallen but due to the error we have taken a wrong decision. Hence the firm continued to produce a high level of output which creates excess supply and losses for the firm.

Type II Error Don't reject Ho when it's false. The demand has actually not fallen but your test result says that it is. A misleading concept and will cause a different cost. The firm will reduce output and may have some layoffs but it shouldn't have taken any action.

There is a tradeoff between these two costs. Assume Ho is true until proven otherwise.

Draw diagram of distribution $\bar{x}$ of under Ho. $\bar{x} = 550$, $S = 250$, $n = 35$.

Hence, $Z = \frac{\bar{x}-\mu}{\sqrt{S^2/n}} = \frac{550-600}{\sqrt{250^2/35}} = -1.18$, the corresponding value in the Z Table.

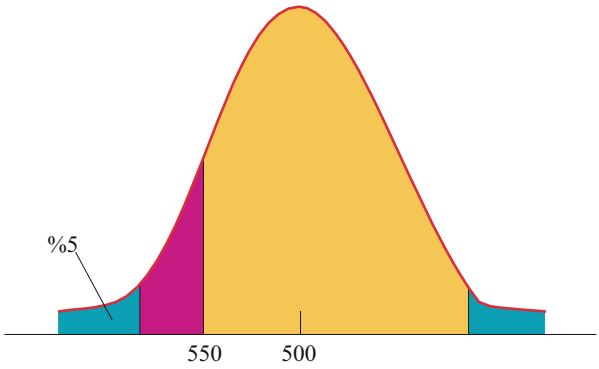


Fig. 10.2 Ho is not rejected

gives: 0.1190 or 11.90 % in tail. (See Fig. 10.2 for a graphical explanation). Is this small enough to reject Ho? It's a question of judgment.

Choose (arbitrarily) a cut-off point of 5 %, called the significance level of the test. That is, if under the assumption that Ho is true, the probability of getting the sample evidence is less than 5 %, we reject Ho. So in this case, do not reject Ho.

Alternatively find the critical value of the test, which cuts off 5 %, in this case $Z^* = 1.64$. Since $1.18 < 1.64$, we do not reject Ho.

The critical value of the test divides the distribution into a rejection region and an acceptance region.

The significance level of the test is the probability of a Type I error.

10.5.1 Choice of Significance Level

The choice of a significance level depends on the relative costs of Type I and Type II errors (and on the sample size). The greater the cost of a Type I error relative to a Type II error, the lower the significance level should be set. Note that the lower the probability of a Type I error the higher the probability of a Type II error. Take the example given before. If the firm reduces the capacity wrongly after a test result, it will lose some of its market share which will be hard to recover (e.g. newspapers) then a low significance level needs to be set. If fixed costs are high relative to variable costs (so capacity is costly) you should set a high significance level to avoid a type II error.

The convention when it is hard to decide on the relative costs is to set the significance level at 4 %, or sometimes 1 %.

10.5.2 The Power of a Test

The power of a test is another important concept and it is;

$$\text{Power of a test} = 1 - \text{P(Type II error)}$$

Note that it is usually impossible to know the probability of a Type II error because we don't know the precise value of μ under Ho. However, we can evaluate the power of a test against a specific alternative.

The closer the two alternatives, the less powerful the test is going to be. (e.g. how do you decide between 0.99 and 0.98?).

10.5.3 One and Two Tail Tests

We could explain one and two tail tests with how the H1 is set. For example if the hypothesis is set in the following way then a one tail test is taken.

$$H_0 : \mu = k$$
$$H_1 : \mu < k$$

where k is any number. The Ho population mean is equal to k and the H1 says population mean is less than k. hence this test is a left tail test. (See Fig. 10.3).

If the alternative test is set as a reverse case of the first explanation then a right tail test needs to be set (Fig. 10.4).

$$H_0 : \mu = k$$
$$H_1 : \mu > k$$

If the test requires both side of the population mean mu then the rejection region is divided into the two tails of the distribution and it is called a two tail test. (Fig. 10.5)

$$H_0 : \mu = k$$
$$H_1 : \mu \neq k$$

It may be not be a good idea to use sample evidence to decide on one or two tail tests. If in doubt, use two tail tests.

10.5.4 Testing a Sample Proportion

The best way to explain testing a sample proportion is to use an example.

Example Otofas car manufacturer argue that their cars are so good that no more than 5 % of its cars need repair in the first three years. As a customer we are

Fig. 10.3 A left tail test

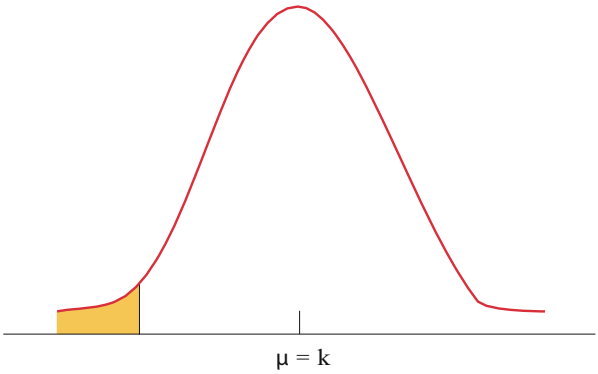


Fig. 10.4 A right tail test

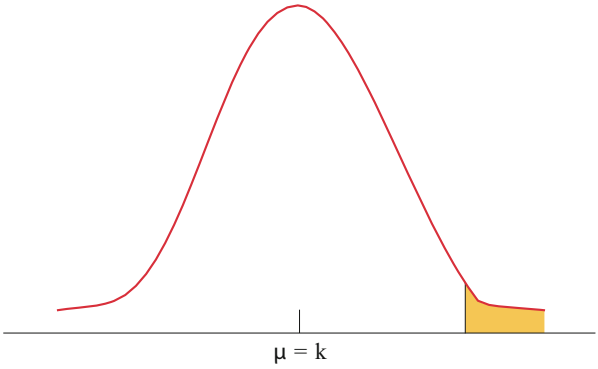
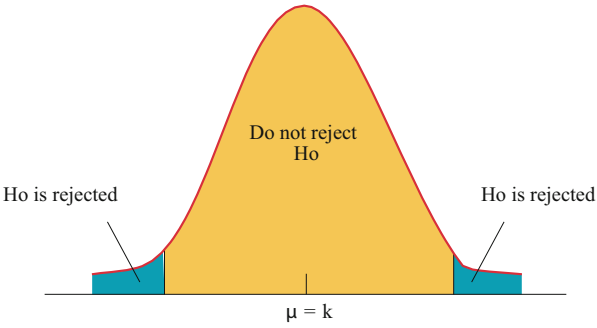


Fig. 10.5 A two tail test



skeptical about this claim. We took a sample of 100 three year old cars and find that 7 of them needed attention. Do you agree with the Otofas?

The summary of the question is given: $p = 0.07$, $n = 100$.

Ho: $\pi = 0.05$, H: $\pi \neq 0.05$ and the significant level is 5 % hence $Z^* = 1.96$.

$$Z = \frac{(p-\pi)}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{(0.07-0.05)}{\sqrt{\frac{0.05(1-0.05)}{100}}} = \frac{0.02}{0.0218} = 0.92$$

$Z < Z^*$ so do not reject Ho with 95 % confidence. The Otofas is likely to be right about its claim.

The Effect of Sample Size

Sometimes a large sample size with a small significance level may itself cause a rejection of Ho. Suppose the data were $p = 0.11$, $n = 3500$, Then $Z = 1.972$, and Ho is rejected. However, the extra 1 % is not very large, and Ho is rejected simply because of large sample size. Hence ideally we should alter significance level with sample size.

10.5.5 Testing Two Sample Means for Equality

Following Ho and H1 hypotheses are set.

$$H_0 : \mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

The appropriate test statistics is:

$$Z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Again the best way to clarify this test is to use an example.

Example You are a stock control manager of a company, which is a large buyer of car batteries. There are two possible types of batteries to purchase and you need to make a decision between them. You want to decide which one of two equally priced brands to purchase at the 5 % level of significance. To do this, you take a random sample of 100 batteries of each brand and find that brand 1 lasts 1060 hours on the average, $\overline{x}^-$ with a sample standard deviation, S_1 of 90 hours. For brand 2, $\overline{x} = 1100$ hours and $S_2 = 140$ hours. Which brand should you purchase if you want to reach a decision at the significance level of: **a)** 5 %? **b)** 1 %?

a) For a 5 % significance:

$$\begin{array}{llll} H_0: & \mu_1 = \mu_2 & Or & H_0: \quad \mu_1 - \mu_2 = 0 \\ H_1: & \mu_1 \neq \mu_2 & Or & H_1: \quad \mu_1 - \mu_2 \neq 0 \end{array}$$

$$\begin{array}{lll} \overline{x}_1 = 1060h & S_1 = 90h & n_1 = 100 \\ \overline{x}_2 = 1100h & S_2 = 140h & n_2 = 100 \end{array}$$

This is a two-tail test with an acceptance region within ± 1.96 under the standard normal curve. Therefore,

$$Z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$\frac{(1060 - 1100) - 0}{\sqrt{\frac{90^2}{100} + \frac{140^2}{100}}} = \frac{1060 - 1100}{\sqrt{81 + 196}} = -2.40$$

Since the calculated value of Z falls within the rejection region for Ho ($2.40 > 1.96$), the buyer should accept H1, that $\mu_1 \neq \mu_2$ at the 5 % level of significance (and presumably decide to purchase brand 2)

b) For a 1 % significance:

At the 1 % level of significance, the calculated Z value would fall within the acceptance region for Ho ($2.40 < 2.58$). This would indicate that there is no significant difference between μ_1 and μ_2 at the 1 % level, so the buyer could buy either brand. Note that even though brand 2 lasts longer than brand 1, brand 2 also has a greater standard deviation than brand 1.

10.5.6 Testing Two Sample Proportions for Equality

Here we test whether two sample proportions are different or not.

$$\begin{array}{llll} H_0: \pi_1 - \pi_2 = 0 & Or & \pi_1 = \pi_2 \\ H_0: \pi_1 - \pi_2 \neq 0 & Or & \pi_1 \neq \pi_2 \end{array}$$

$$Z = \frac{(p_1 - p_2)}{\sqrt{\frac{p'(1-p')}{n_1} + \frac{p'(1-p')}{n_2}}} \text{ where p' is } \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

To understand this concept we consider two sample proportions in an example.

Example Let us consider two industrial areas,Cukurova and Marmara, and assume that there are 70 factories in Marmara and 50 factories in Cukurova which can pollute nature. The factories in both districts follow some of governments recently introduced antipollution standards but with different degrees. (They try to get away as much as they can from polluting). Further assume that 55 % of Marmara region's 70 factories are applying the antipollution standards and 45 % of Cukurova regions 50 plants abide by the antipollution standards. Is the percentage of plants abiding by the antipollution standards significantly greater in Marmara as opposed to Cukurova at:

a) the 5 % level of significance? **b)** the 10 % level of significance?

a) The 5 % level of significance:

$$H_0 : \pi_1 = \pi_2 \quad \text{and} \quad H_1 : \pi_1 > \pi_2$$

$$p_1 = 0.55 \quad \text{and} \quad n_1 = 70$$
$$p_2 = 0.45 \quad \text{and} \quad n_2 = 50$$

This is a right-tail test and the acceptance region for Ho with a = 0.05 lies to the left of 1.64 (for 5 % level of significance) under the standard normal curve:

$$Z = \frac{(p_1 - p_2)}{\sqrt{\frac{p'(1-p')}{n_1} + \frac{p'(1-p')}{n_2}}}$$

Where p' is $\frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{70(0.55)+50(0.45)}{70+50} = 0.51$

Thus

$$Z = \frac{(0.55 - 0.445)}{\sqrt{\frac{0.51(1-0.51)}{70} + \frac{0.51(1-0.51)}{50}}} = 0.093$$

Since $Z^* = 1.64 > 0.093$ we accept Ho that $\pi_1 = \pi_2$, with the 5 % level of significance. ($\alpha = 0.05$) There is no statistical difference in applying the anti-pollution standards between the two regions at 5 % level of significance.

b) The 10 % level of significance:

With $\alpha = 0.10$, the acceptance region for Ho lies to the left of 1.28 under the standard normal curve. Since the calculated Z falls within the acceptance region, we accept Ho at $\alpha = 0.10$ as well.

10.6 Small Samples: Using the t Distribution

10.6.1 Testing the Sample Mean

You would like to examine corruption and bribery in municipalities in Turkey. You have looked at a number of Justice and Development Party (JDP) controlled municipalities and they take an average of 300 Million Turkish Lira (MTL) per job, the so-called donation-bribe. This amount seems normal to local citizens since the municipalities have a lack of funding. These donations of course are not spent for community purposes but are extra salary for the municipality officers. A sample of 10 Republican Left Party (RLP) controlled municipalities reveals the average donation-bribe to be 320MTL, with standard deviation 40. Do the RLP controlled municipality authorities charge more than JDP ones (i.e. which one is more corrupt)?

$$\begin{aligned} H_0 &: \mu = 300 \\ H_1 &: \mu > 300 \end{aligned}$$

The significance level is 5 % thus t* = 1.833 (v = 9 in t table)

$$\frac{\overline{x} - \mu}{S^2/n} \sim t_v = \frac{320 - 300}{40^2/10} = 1.58$$

Comparing with the table value; $t < t^* = 1.58 < 1.833$ we cannot reject Ho. There is no statistically significant difference between JDP and RLP authorities at 5 % level of significance.

10.6.2 Testing Two Sample Means' Difference

Here we have again small sample sizes. The hypothesis is about the difference of two means.

$$\frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_v \text{ where } S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

and $v = n_1 + n_2 - 2$

This again is best explained with an example.

Example The Turkish Consumer Association observed a number of toothpastes which all seem to claim they are the best for fighting tooth decay. They wanted to test which one is the best for this aim. After consulting the experts in the area they reduced the number to two brand names, Mapana and Disgate. They now want to test which of two toothpastes is better for fighting tooth decay. They observed 16 randomly selected people using each of the toothpaste over an 8 year period. The

average number of cavities for the Mapana users was 28 with a standard deviation of 6. The Disgate users' average number of cavities was 25 with a standard deviation of 5. Assuming that the distribution of cavities is normal for all the users of Mapana and Disgate and that, $\sigma_2^1 = \sigma_2^2$ determine if $\mu_1 = \mu_2$ at the 5 % level of significance.

$$\begin{array}{llll} Ho: & \mu_1 = \mu_2 & and & H_1: \quad \mu_1 \neq \mu_2 \\ \overline{x}_1 = 28 & S_1 = 6 & & n_1 = 16 \\ \overline{x}_2 = 25 & S_2 = 5 & & n_2 = 16 \end{array}$$

Since the two populations are normally distributed but n_1 and $n_2 < 30$ and and it is assumed that $\sigma_2^1 = \sigma_2^2$ (but unknown), the sampling distribution of the difference between the means has at distribution with $n_1 + n_2 - 2$ degrees of freedom. Since it is assumed that $\sigma 2 = \sigma 2$ (and we can use S2 as an estimate of $\sigma 2$ and S2 as an estimate of estimate of σ_2^2), we get

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

S^2 is a weighted average of S_1^2 and S_2^2. The weights are $n_1 - 1$ and $n_2 - 1$ for S_1^2 and S_2^2, in order to get 'unbiased' estimates for σ_1^2 and σ_2^2. This is a two-tail test acceptance region for Ho within ± 2.042 under the t distribution with $\alpha = 5$ % and

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{(16 - 1)6^2 + (16 - 1)5^2}{16 + 16 - 2} = 30.5$$

$$t = \frac{(28 - 25) - 0}{\sqrt{\frac{30.5}{16} + \frac{30.5}{16}}} = \frac{28 - 25}{1.95} \cong 1.54$$

Since the calculated value of t falls within the acceptance region, $(-2.042 < t^* < 2.021)$ we cannot reject Ho, that $\mu_1 = \mu_2$. Hence there is no true statistical difference between Mapana and Disgate at 5 % significance level.

10.7 Conclusion

A standard method for the hypothesis testing.

1. Formulate null and alternative hypotheses.
2. Decide on significance level (weigh up costs and benefits of the two types of errors)
3. Use relevant test statistics (sample size? means or proportions ? standard error?)

Problems

Suppose something is just significant at 5 % level?

Only the null ever gets tested, and one's pet theory is often the alternative. Hence you set up a weak null and find it rejected! There is asymmetry between null and alternative hypotheses.

10.8 A Review of This Chapter

This chapter examined hypothesis testing. Firstly the nature of hypotheses and scientific method with different types of errors were discussed. For a large or small samples, different test statistics were used. Testing sample proportions, choice of significance level, power of test, one and two tail tests were discussed.

10.9 Review Problems for Hypothesis Testing

10.1. a) What is meant by testing a hypothesis? What is the general procedure?
b) What is meant by type I and type II errors?
c) What is meant by the level of significance? The level of confidence?

10.2. The Mavijean company knows that the average size of order received is 500 with the standard deviation of 200. The last 32 orders have averaged 480 units. Does this indicate a fall in their demand?

10.3. Landford car manufacturer argues that their cars are so good that no more than 6 % of its cars need repair in first three years. As a customer we are skeptical about this claim. We took a sample of 120 three year old cars and find that 8 of them needed attention. Do you agree with the Landford?

10.4. You are a stock control manager of a company, which is a large buyer of car batteries. There are two types of batteries possible to purchase and you need to make a decision between them. You want to decide which one of two equally priced brands to purchase at the 5 % level of significance. To do this, you take a random sample of 120 batteries of each brand and find that brand 1 lasts 1050 hours on the average $\bar{x}_1$, with a sample standard deviation, S_1 of 95 hours. For brand 2, $\bar{x} = 1200$ hours and $S_2 = 145$ hours. Which brand should you purchase if you want to reach a decision at the significance level of: **a)** 5 %? **b)** 1 %?

10.5 You would like to examine corruption and bribery in municipalities in Turkey. You have looked at a number of Mother Path Party (MPP) controlled municipalities and they take an average of 250 Million Turkish Lira (MTL) per job, the so-called donation-bribe. This amount seems normal to local citizens since the municipalities have a lack of funding. These donations of course are not spent for community purposes but are extra salary for the municipality officers. A sample of 12 Republican Left

Party (RLP) controlled municipalities reveals the average donation-bribe to be 220MTL, with standard deviation 35. Do the RLP controlled municipality authorities charge more than MPP ones (i.e. which one is more corrupt)?

10.6 The Turkish Medical Association observed a number of tooth pastes which all seem to claim they are the best for fighting tooth decay. They wanted to test which one is the best for this aim. After consulting the experts in the area they reduced the number into to two brand names, Mapana and Disgate. They now want to test which of two tooth pastes is better for fighting tooth decay. They observed 18 randomly selected people using each of the toothpaste over a 12 year period. The average number of cavities for the Mapana users was 30 with a standard deviation of 5. The Disgate users' average number of cavities was 28 with a standard deviation of 4. Assuming that the distribution of cavities is normal for all the users of Mapana and Disgate and that $\sigma_1^2 = \sigma_2^2$, determine if $\mu_1 = \mu_2$ at the 5 % level of significance. Which toothpaste is the best for fighting the tooth decay.

Chapter 11
The Chi-Squared, F-Distribution and the Analysis of Variance

11.1 Introduction

In this chapter we will consider two main, widely used distributions; the chi-squared and F distributions. Firstly we will focus on to the two similar distributions; the F and the chi-squared distributions. They share a number of common features. The main common features are that they are

- Always non-negative.
- Skewed to the right.

We will examine three applications of χ^2 distribution and two applications of the F distribution.

11.2 The Chi-Squared (χ^2) Distribution

The diagram below clarifies the point that as the degrees of freedom increase, the chi- squared distribution becomes more bell-like and begins to look more like the normal distribution. Since it is positive we always use a positive test.

We will focus on three applications of chi-squared tests in this part. The first one is to calculate a confidence interval estimate of the population variance. The second one is to compare actual observations on a variable with the expected values and the third one is to test for association between two variables in a contingency table.

The online version of this chapter (doi:10.1007/978-3-319-26497-4_11) contains supplementary material, which is available to authorized users.

D. Özdemir, *Applied Statistics for Economics and Business*,
DOI 10.1007/978-3-319-26497-4_11

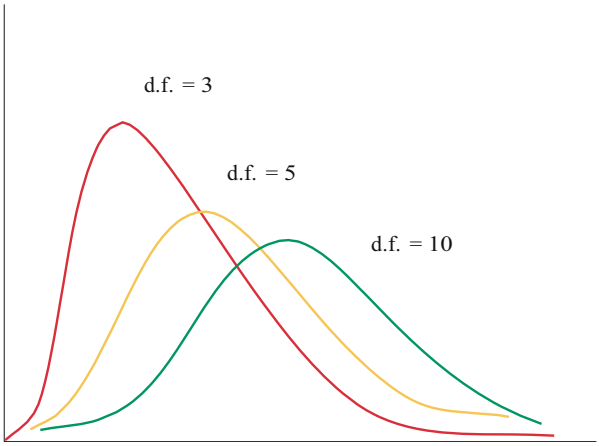


Fig. 11.1 The chi-squared distribution

11.2.1 To Calculate the Confidence Interval Estimates of a Population Variance

In the previous chapter we have assumed that the population variance is known but we did not explain how we can estimate the population variance. In this section we will go into the details of this. For this purpose we need to know how the distribution of variance is;

$$\text{Remember we had}: \quad S^2 = \frac{\Sigma(x_i - \overline{x})^2}{n-1}$$

$\rightarrow S^2$ is an unbiased estimate of σ^2, so consider it as a point estimate. To obtain the distribution of variance we will look at their formulae:

$$S^2 = \frac{(X_1 - \overline{X})^2}{n-1} + \frac{(X_2 - \overline{X})^2}{n-1} + \ldots\ldots\ldots\ldots + \frac{(X_n - \overline{X})^2}{n-1}$$

$$\text{So } (n-1)S^2 = (X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots\ldots\ldots\ldots + (X_n - \overline{X})^2$$

$$\text{Or } \frac{(n-1)S^2}{\sigma^2} = \frac{(X_1 - \overline{X})^2}{\sigma^2} + \frac{(X_2 - \overline{X})^2}{\sigma^2} + \ldots\ldots\ldots\ldots + \frac{(X_n - \overline{X})^2}{\sigma^2}$$

In fact the last equation looks very similar to squared Z distribution.

$$\frac{(X_n - \overline{X})^2}{\sigma^2} \rightarrow \frac{(X_i - \overline{X})^2}{\sigma^2}$$

So $\dfrac{(n-1)S^2}{\sigma^2}$ is a chi-squared distribution with $v = n-1$ degrees of freedom.

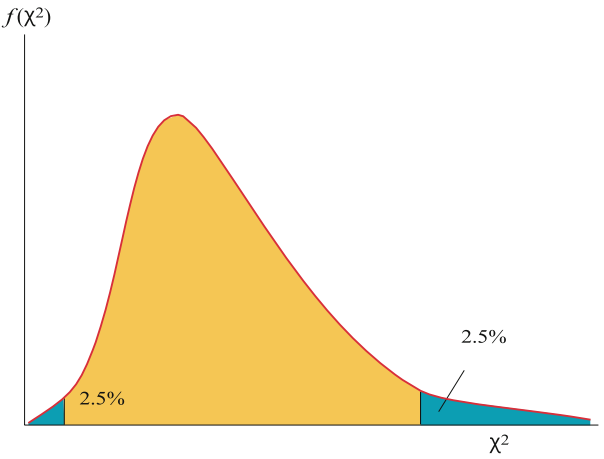


Fig. 11.2 Critical values of chi- squared distribution for the 95 % confidence interval

In general; $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

We would like to have the estimate of σ^2. In order to do this we need to obtain a critical value for σ^2. The rejection region cuts off an area a in either or both regions. Figure 11.2 shows rejection region cuts in both areas. For a 95 % confidence interval, the critical values of χ^2 distribution are cut off at 2.5 % in each tail. Since the χ^2 distribution is not symmetric it is different to the Normal and t distributions. Figure 11.2 shows the critical values for the 95 % confidence interval.

Example Let us examine this case in an example. Assume that the sample size n is 8 and the sample variance is 3.6. What is the population variance? Construct the 95 % confidence interval around the point estimate.

The summary of the given information:

$$\text{n} = 8 \quad \text{S}^2 = 3.6$$
$$\text{Thus} \;\; \text{v} = \text{n} - 1 = 8 - 1 = 7$$

We now look at the χ^2 distribution table for 7 degrees of freedom and 95 % confidence interval. For the left tail, it is 0.975 on the top of the table and corresponds to 1.68987. For the right tail it is 0.025 on the top of the table and corresponds to 16.013.

$$\text{So}: \frac{\frac{(n-1)S^2}{\sigma^2} < K_u}{\frac{(n-1)S^2}{\sigma^2} > K_L} \qquad \text{Or} \;\; K_L \leq \frac{(n-1)S^2}{\sigma^2} \leq K_u$$

$$\frac{(n-1)S^2}{K_L} \geq \sigma^2 \geq \frac{(n-1)S^2}{K_U} \qquad \frac{(8-1)3.6}{1.690} \geq \sigma^2 \geq \frac{(8-1)3.6}{16.013}$$

$$\rightarrow \quad 14.91 \geq \sigma^2 \geq 1.57$$

Thus the point estimate 3.6 is no longer at the center of the interval but is closer to the limit. (Because the chi-squared distribution is skewed to the left).

11.2.2 Comparing Actual Observations with Expected Values

In general

- We have a random sample of n observations.
- They can be classified according to K categories.
- The number of observations falls into each category of;

$O_1, O_2, \ldots\ldots\ldots O_K$

In this test if the observed and expected values differ significantly after the use of the chi-squared test then the null hypothesis is rejected.

Example To take a simple example, say we have observed the outcome of three different events (Table 11.1):

So $O_1 = 8$, $O_2 = 10$, $O_3 = 15$, $K = 3$, $n = 33$,

We would like to test the fact that the observations are no different to the population.

Ho: The population fits the given distribution.
H1: The population has a different distribution.

We need to use a chi-squared test to clarify these points, hence the formula of our general test's statistic is:

$$\sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{K-1}$$

Firstly we need to obtain the expected values (E_i) for three observations.
These would be $33/3 = 11$ for all.
Hence applying the formula above gives;

$$\frac{(8-11)^2}{11} + \frac{(10-11)^2}{11} + \frac{(15-11)^2}{11} = \frac{9+1+16}{11} = \frac{26}{11} = 2.36$$

In the next step we need to check the critical values:
If the critical value is 2.5 %; $v = K - 1 = 2 \rightarrow$ tabular value is 7.378. (one tail)
If the critical values 5 % $v = 2 \rightarrow$ tabular value is 5.991 (one tail).
If the tabular values are greater than the calculated value we do not reject Ho.

Table 11.1 Simple events

Category	1	2	3	Total
Number of Observations	8	10	15	33

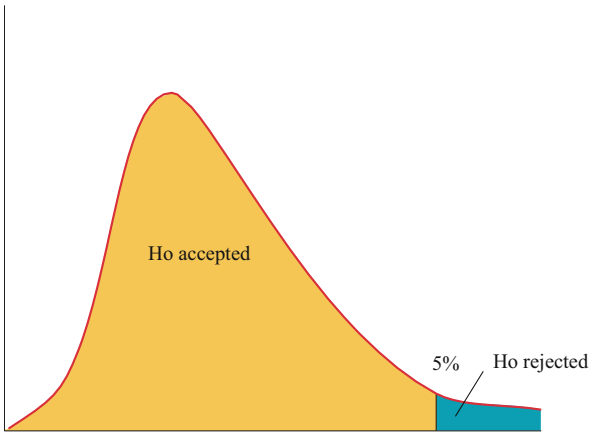


Fig. 11.3 Chi-squared test

Since the calculated chi-squared is less than the tabular values the null hypothesis is not rejected. (Fig. 11.3)

Let us **outline each step**:

- State the null and alternative hypothesis.
- Find the value of the chi-squared statistics for the sample.
- Find the degrees of freedom and the appropriate critical chi-squared value.
- Sketch the critical region and locate your sample chi-squared value and critical chi-squared value on the sketch.
- Decide whether you should reject or fail to reject the null hypothesis.

11.2.3 To Test the Association Between Two Variables in a Contingency Table

Let us examine this case with an example.

Example The number of heart attacks suffered by males and females of various age groups in Istanbul is given in the contingency Table 11.2 below. Test at the 1 % level of significance the hypothesis that age and sex are independent in the occurrence of heart attacks.

To test this hypothesis, expected frequencies (E_i) must be estimated.

Expected Frequencies

For the cell in the first row, r, and first column, c:

Table 11.2 Heart attacks suffered by males and females of various age groups in Istanbul

Age group	Male	Female	Total
Below 30	10	10	20
From 30–60	50	30	80
Above 60	30	20	50
Total	90	60	150

Table 11.3 Expected frequencies

Age group	Exp. male	Exp. female	Total
Below 30	12	8	20
From 30–60	48	32	80
Above 60	30	20	50
Total	90	60	150

$$E_{1m} = \frac{row\,total \times column\,total}{grand\,total} = \frac{\sum r \sum c}{n} = \frac{(20)(90)}{150} = 12$$

For the cell in the second row and first column, c:

$$E_{2m} = \frac{\sum r \sum c}{n} = \frac{(80)(90)}{150} = 48$$

All other expected frequencies can be obtained by subtracting from the appropriate row or column totals and are presented in the following Table 11.3:

Therefore;

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(10-12)^2}{12} + \frac{(10-8)^2}{8} + \frac{(50-48)^2}{48} + \frac{(30-32)^2}{32}$$

$$+ \frac{(30-30)^2}{30} + \frac{(20-20)^2}{20} = \mathbf{1.04}$$

$$\text{d.f.} = 3 - 1 = 2$$

Since the calculated chi-squared is smaller than the tabular chi-squared, we accept the null hypothesis, Ho, that age is independent of sex in the occurrence of heart attacks. To be sure, males seem more likely to suffer heart attacks, but this tendency does not differ significantly with age at the 1 % level of significance.

Table 11.4 Monthly income data

Range (TL) Monthly Income	Numbers of Households f
0-	950
80-	1800
130-	1300
250-	610
450-	400
800-	200
1000-	180
1500-	80
Total	5520
TL = Turkish Liras	

11.2.4 Test for the Normal Distribution

In our earlier discussions we determined some of the distributions as normal. A sample data's distribution sometimes is not easy to determine. Let us take the income data, which we have examined earlier in chapter 6. The data is re-presented in Table 11.4. Is there any evidence that the income distribution in this data distributed normally. Test for 5 % significance level.

Thus:

H_0: The distribution is normal.
H_1: The distribution is not normal

The degrees of freedom is (v): Income groups minus the number of columns minus one: v = 8 - 2 - 1 = 5. Hence the critical value from tables = 11.1

ch11data1.xls

The mean of the sample is 281.301 and the standard deviation is 343.3 as was calculated before. (See the relevant formulas for the mean and the standard deviation in chapter 3 and 4 for). The first step is to calculate the expected values. For this purpose we need to convert the original upper limit to Z scores. Subtract the mean from the upper range and then divide by the standard deviation. These Z scores are shown in column 3 of Table 11.5. After obtaining the Z scores, we find the corresponding values from the normal distribution table at the end of the text. These values are presented under the heading called probability on the table below. The next step is to derive the expected frequencies. These are obtained by probabilities multiplied by the sum of the frequencies and shown on the last column of the Table 11.5.

The chi-squared formula is: $\chi^2 = \sum \left[\frac{(O-E)^2}{E} \right]$ where O denotes the observation-sand E is the expected frequency values. The observed values are the

Table 11.5 Expected frequency calculations

Mid point	Numbers of x households f	Upper limit- mean/ std Z	Probability	E prob*5520 exp. freq.
40	950	−0.58	0.281	1551.12
105	1800	−0.44	0.33	1821.6
190	1300	−0.09	0.4641	2561.832
350	610	0.49	0.3121	1722.792
800	400	1.51	0.0655	361.56
900	200	2.09	0.01831	101.0712
1250	180	3.55	0.00023	1.2696
1750	80	5	0.00002	0.1104
	5520			

Table 11.6 Chi-square calculations

O	(O - E)	(O - E)^2	(O-E)^2)/E
950	−601.12	361345.3	232.9576
1800	−21.6	466.56	0.256126
1300	−1261.83	1,592,220	621.5162
610	−1112.79	1,238,306	718.7786
400	38.44	1477.634	4.086828
200	98.9288	9786.907	96.83181
180	178.7304	31944.56	25161.12
80	79.8896	6382.348	57811.12
		chi-sqr.=	84646.67

frequencies in Table 11.6. Subtracting the expected frequencies from the observed values gives the following relevant calculations:

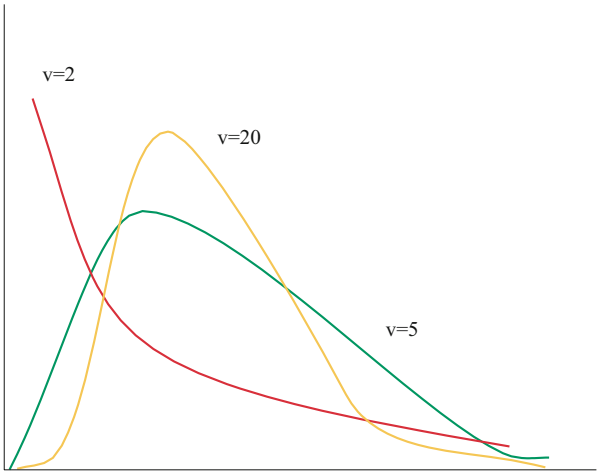
The obtained chi-squared value (84646.67) is higher than the tabular value (11.1) so we therefore reject the null hypothesis. We can conclude that there is no evidence that the income distribution is normal.

11.3 The F- Distribution

This is similar to the chi-squared distribution, i.e. always negative and skewed to the right but it has two sets of degrees of freedom. There are a number of different uses of the F distribution in statistics but we will look at two of these in this chapter. These are for testing for the equality of two variances and the analysis of variance (ANOVA) test. We will also use the F distribution in the later topic of regression analysis. The Fig. 11.4 below shows a typical F distributions.

Followed by the ratio of two independent chi-squared and each is divided by its associated degree of freedom.

Fig. 11.4 F distribution



$$\frac{\chi_1^2/_{v1}}{\chi_2^2/_{v2}} \sim F_{v_1, v_2}$$

Looking at the critical value is a bit more complicated than in the chi-squared distribution.

At the end of the text there are 4 different tables, each gives;

$v_1 \rightarrow$ associated with the numerator.
$v_2 \rightarrow$ associated with the denominator.

Specially defined for the upper 5 %, 2.5 %, 1 %, 0.5 %.

Let us say we have to just accounted a variable which has chi-squared data;

$$\rightarrow \sigma^2$$
$$\frac{(n-1)S^2}{\sigma^2}$$

So the F distribution is based on a test for the equality of two variances.

The hypothesis test can be based on a two or single tail rejection region for the F distribution.

11.3.1 Testing for the Equality of Two Variances

Earlier we conducted a hypothesis test on a mean. Now we will conduct a test on variance. If there are two samples, the most usual test is then the testing for the equality of two variance. S^2, v_1, S^2, v_2, First and second sample variances and their relevant degrees of freedoms are denoted with the following:

The first samples chi-squared distribution:

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}$$

The second samples chi-square distribution:

$$\frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$$

Setting the hypothesis:
Under Ho: $\sigma_1^2 = \sigma_2^2$, (Population variances are equal)
H1: $\sigma_1^2 \neq \sigma_2^2$ (Population variances are not equal)
Or $H_0 : \sigma_1^2/\sigma_2^2 = 1. H_1 : \sigma_1^2/\sigma_2^2 \neq 1$
From the previous definition

Also $$\frac{\chi^2_{n_1-1} \Big/ n_1 - 1}{\chi^2_{n_2-1} \Big/ n_2 - 1} \sim F_{n_1-1,\, n_2-1}$$

Replace the chi-squared values;

$$\frac{(n_1-1)S_1^2/\sigma^2(n_1-1)}{(n_2-1)S_2^2/\sigma^2(n_2-1)} = \frac{S_1^2/\sigma^2}{S_2^2/\sigma^2} = \frac{S_1^2}{S_2^2} \sim F_{n_1-1,n_2-1}$$

This last equation tells us that under Ho, the ratio of two sample variances has an F distribution with degrees of freedom n_1-1, n_2-1.

This can again be best understood with an example.

Example The following information is obtained from two samples which provide two sample variance values and the size of these two samples:

	Sample 1	Sample 2
s^2	123.35	8.02
n	16	11

The ratio of the two sample variance is:

$$\frac{S_1^2}{S_2^2} = \frac{123.35}{8.02} = 15.38$$

The two samples' critical values are;
$v_1 = 16 - 1 = 15$, $v_2 = 10$
From the table 5 % (v1 = 15, v2 = 10) corresponding to 2.845.
1 % (v1 = 15, v2 = 10) corresponding to 4.5581.

So the calculated value is greater than the tabular value for both 5 % and 1 %. Thus: reject Ho even at 1 % significance level.

The population variances are not equal.

11.4 Analysis of Variance (ANOVA)

Our earlier discussions thought us how to test the means of two samples. In fact it is possible to generalize this to more than two samples. Earlier we used the t test or the Z test depending on the sample size. Analysis of variance is a technique to generalize these types of tests by using the F distribution. In fact the analysis of variance is a test which tests whether there is a statistical difference between more than two means. It allow us to examine whether all the sample means are equal or not as opposed to the one or more than one mean not being equal.

11.4.1 One-Way Analysis of Variance

There are six steps for the analysis of variance. These are:

(a) **Step 1**: determine the Null and the alternative hypothesis. Testing the equality of several means of a random sample.

H_0: $\mu_1 = \mu_2 = \ldots\ldots\ldots\ldots\ldots = \mu_K$
H_1: At least one is different.

This is known as **one-way analysis of variance** and the simplest type of ANOVA. How can we test this?

Firstly we want to take an overall sample variance.

(b) **Step 2**: Find Total Sum of Squares (TSS).
(c) **Step 3**: Find the Sum of Squares Group (SSG). This measures the variability of group means.
(d) **Step 4**: Find the Sum of Square variations within group (SSW).
(e) **Step 5**: Find the variance estimates. (Mean Squares)
(f) **Step 6**: Find the F ratio and complete the ANOVA test.

Put all the results in a table, which is called the analysis of variance Table 11.7.

$$\text{We have } n = \sum_{i=1}^{K} n_i \quad \text{Overall mean}: \overline{X} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i} X_{ij}}{n} = \frac{\sum_{i=1}^{K} n_i \overline{X}_i}{n}$$

Table 11.7 An example table for the ANOVA results

Source of variation	Sum of squares	Degrees of freedom	Mean square (variance)	F Ratio	F Critical value	Test decision
Between Groups	SSG	d. f. group	MSG	MSG/ MSW	From Table	Reject Ho or Fail to
Within Groups	SSW	d. f. w.	MSW			Reject Ho
Total	TSS	N−1				

Table 11.8 Servicing cost in Turkish Liras (TL)

	Meugeot	Biat	Zord
	220	192	203
	215	193	200
	210	185	197
	205	180	190
	200	175	185
Total	1050	925	975
Mean	210	185	195

Example This can again be best examined with an example. Let us take three different hypothetical car companies, Zord, Biat and Meugeot in Turkey. A survey of servicing cost in 5 different service stations for each car is presented in the Table 11.8.

H_0: $\mu_1 = \mu_2 = \ldots\ldots\ldots\ldots\ldots = \mu_K$

H_1: At least one is different.

Let us re-write the 3 samples' sample means and the sample sizes:

$$\begin{aligned} \overline{X}_1 &= 210 \quad n_1 = 5 \\ \overline{X}_2 &= 185 \quad n_2 = 5 \\ \overline{X}_3 &= 195 \quad n_3 = 5 \end{aligned}$$

Thus $n = 5 + 5 + 5 = 15$

The overall mean is:

$$\overline{X} = \frac{5(210) + 5(185) + 5(195)}{15} = 196.67$$

The next step is to distinguish between two types of variability.

→ the variance due to difference between groups.

→ the variance due to difference within groups (also known as the error variance)

So for each group:

$$SS_1 = \sum_{j=1}^{n_1} \left(X_{1j} - \overline{X}_1\right)^2$$
$$SS_2 = \sum_{j=1}^{n_2} \left(X_{2j} - \overline{X}_2\right)^2$$
$$\cdots\cdots\cdots$$
$$\cdots\cdots\cdots$$
$$\cdots\cdots\cdots$$
$$SSW = SS_1 + SS_2 + \cdots\cdots\cdots SS_K$$
$$= \sum_{i=1}^{K}\sum_{j=1}^{K}\left(X_{ij} - \overline{X}_i^2\right)$$

In our example:
$SS_1 = (220-210)^2 + (215-10)^2 + \ldots\ldots\ldots + (200-10)^2 = 250$
$SS_2 = (192-185)^2 + (193-185)^2 + \ldots\ldots\ldots + (175-185)^2 = 238$
$SS_3 = (203-195)^2 + (200-195)^2 + \ldots\ldots\ldots + (185-195)^2 = 218$
Thus $\mathrm{SSW} = \mathrm{SS}_1 + \mathrm{SS}_2 + \mathrm{SS}_3 = 250 + 238 + 218 = 706$
Variance between groups:
Take the general mean as an overall area

$$\left(\overline{X}_1 - \overline{X}\right)^2, \left(\overline{X}_2 - \overline{X}\right)^2, \ldots\ldots\ldots, \left(\overline{X}_K - \overline{X}\right)^2$$
$$SSG = \sum_{i=1}^{K} n_i \left(\overline{X}_i - \overline{X}\right)^2$$

And weigh for distribution of the groups.
In our example:

$$\mathrm{SSG} = 5(210 - 196.67)^2 + 5(185 - 196.67)^2 + 5(195 - 196.67)^2$$
$$= 1583.334$$

Then we have the overall SS

$$SST = \sum_{i=1}^{K}\sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_i\right)^2$$

$\mathrm{TSS} = \mathrm{SSW} + \mathrm{SSG} = 706 + 1583.334 = 2289.334$
Then the test statistics are as follows:

$$F = \frac{SSG'_{K-1}}{SSW'_{n-K}} \sim F_{K-1,n-K}$$

In our example (where $k = 3$ groups):

Table 11.9 Summary of ANOVA results

Source of variation	Sum of squares	Degrees of freedom	Mean square (variance)	F Ratio	F Critical value	Test decision
Between Groups	1583.334	3−1 = 2	1583.334/ 2 = 791.667	791.667/ 58.83 = 13.46	For 99 % 6.92	
Within Groups	706	15−3 = 12	706/ 12 = 58.83		For 95 % 3.88	Reject Ho
Total	2289.334	14				

$$F = \frac{1583.334/_{3-1}}{706/_{15-3}} = \frac{791.667}{58.83} = 13.46$$

Is there a rationale for this test?
Critical values:

$$F_{2,12} \rightarrow Fv_1, v_2 \quad \begin{matrix} 99\% \rightarrow 6.92 \\ 95\% \rightarrow 3.88 \end{matrix} \quad \text{Ho is rejected}$$

The calculated value is greater than the tabular values. Hence the population mean servicing cost is not likely to be same for both 1 % and 5 % level of significance.

A summary of these results in a table called the analysis of variance table and is presented in Table 11.9 below.

11.4.2 Two-Way Analysis of Variance

In the example of one way analysis, our primary concern was a genuine difference between service cost and a random variation. However each service might have some training differences in their staff levels and we can classify these into five different classes. The first class holds the best trained service staff, the second class is not as good as the first and so on to the fifth class. In this question we are trying to obtain information about two factors simultaneously. The first point is: are such differences due to the different model cars in the average servicing cost; second one is whether the training level of staff effect the average servicing cost (More training-higher salary- higher cost). Due to the fact that we are examining two factors, this analysis is called two-way analysis of variance. The additional variable (in this case the training levels of staff) is called a blocking variable. The sample data is presented in the Table 11.10 below.

To test whether the population means are the same or not for all K groups, we need to calculate the sample means for each group.

Table 11.10 Sample observations on servicing cost recorded for three types of cars serviced by six different trained mechanics (MTL)

Service Class	Meugeot	Biat	Zord	Sum
1	220	192	203	615
2	215	193	200	608
3	210	185	197	592
4	205	180	190	575
5	200	175	185	560
Total	1050	925	975	2950

$$\overset{M}{x_i} = \frac{\sum_{j=1}^{H} x_{ij}}{\mathrm{H}} (i = 1, 2, \ldots\ldots\ldots K)$$

Where K denotes groups and H blocks. The $\overline{x}_i$ is the mean of the ith group. From the Table 11.10:

$$\overline{x}_1 = \frac{1050}{5} = 210, \overline{x}_2 = \frac{925}{5} = 185, \overline{x}_3 = \frac{975}{5} = 195$$

The second way was the population block way. Hence we also need to obtain sample means for H blocks:

$$\overset{M}{x_j} = \frac{\sum_{j=1}^{H} x_{ij}}{\mathrm{K}} (i = 1, 2, \ldots\ldots\ldots H)$$

By using the Table 11.10 we can obtain the population block means.

$$\overline{x}_1 = \frac{615}{3} = 205, \overline{x}_2 = \frac{608}{3} = 202.67, \overline{x}_3 = \frac{592}{3} = 197.3$$

$$\overline{x}_4 = \frac{575}{3} = 191.67, \overline{x}_5 = \frac{560}{3} = 186.67$$

The total number of observations:

$$n = HK$$

The sample mean of all observations is given by:

$$\overline{x} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{H} x_{ij}}{n} = \frac{\sum_{i=1}^{K} \overline{x}_i}{K} = \frac{\sum_{j=1}^{H} \overline{x}_j}{H}$$

Thus from the Table 11.10;

$$\overline{x} = \frac{2950}{15} = 196.67$$

$$\text{SST} = (220 - 196.67)$$

$$2 + (215 - 196.67)^2 + \ldots\ldots\ldots\ldots + (185 - 196.67)^2 = 2289.334$$

$$\text{SSG} = 5(210 - 196.67)^2 + 5(185 - 196.67)^2 + 5(195 - 196.67)^2 = 1583.334$$

Here we also need to define between blocks sum of squares (SSB)

$$SSB = K\sum_{j=1}^{K}\left(\overline{X}_j - \overline{X}\right)^2$$

$$\text{SSB} = 3\left[(205 - 196.67)^2 + (202.67 - 196.67)^2 + (197.3 - 196.67)^2 + (191.67 - 196.67)^2 + (186.67 - 196.67)^2 = 692.3574\right.$$

The error sum of squares (SSE):

$$\text{SSE} = \text{SST} - \text{SSG} - \text{SSB} = 2289.34 - 1583.334 - 692.3574 = \mathbf{13.65}$$

The rest of the two-way ANOVA is very similar to the one-way analysis of variance. We obtain the mean squares by dividing each sum of squares by the relevant degrees of freedom:

$$\text{Between-Groups} : MSG = \frac{SSG}{K-1} = \frac{1583.334}{3-1} = 791.667$$

$$\text{Between-Blocks} : MSB = \frac{SSB}{H-1} = \frac{692.3574}{5-1} = 173.089$$

$$\text{Error Mean Square} : \text{MSE} = \frac{\text{SSE}}{(K-1)(H-1)} = \frac{13.65}{8} = 1.70625$$

Here since we have two way ANOVA then we have two null and the alternative hypothesis.

The first one is

For testing the K population group means:

H_0: K population group means are all the same.
H_1: K population group means are not all same.

$$\text{Reject Ho if} \qquad \frac{\text{MSG}}{MSE} > \text{F}_{K-1,\,(K-1)(H-1),\,\alpha}$$

$$\frac{MSG}{MSE} = \frac{791.667}{1.70625} = 463.98$$

Let us test this for a 1 % level of significance; The table value of F is:

$$F_{K-1,\,(K-1)(H-1),\,\alpha} = F_{2,8,0.01} = 8.6491$$

On the evidence of our data we can say that the hypothesis of equal average population servicing cost is rejected for the three brand named cars at 1 % level of significance.

The second one is

Testing the H population block means:

The null hypothesis that, population values of mean servicing cost are the same for each different level of trained service mechanics.

H_0: H population block means are all the same.
H_1: H population block means are not all the same.

$$\text{Reject Ho if} \quad \frac{MSB}{MSE} > F_{H-1,\,(K-1)(H-1),\,\alpha}$$

$$\frac{MSB}{MSE} = \frac{173.089}{1.70625} = 101.44$$

For the 1 % level of significance the table value of F is:

$$F_{H-1,\,(K-1)(H-1),\,\alpha} = F_{4,8,0.01} = 7.0061$$

We can say that the null hypothesis is rejected at 1 % level of significance. That means the alternative hypothesis is accepted, namely the population values of mean servicing cost are **not** the same for each different level trained service mechanics.

Again the best way to present all these results is to use an ANOVA table as we did in one way analysis of variance discussion (Table 11.11).

Now we can construct our examples' table with the obtained results. These results are presented in Table 11.12 below.

Table 11.11 An example table of the two-way ANOVA results

Source of variation	Sum of squares	Degrees of freedom	Mean square	F Ratio	F Critical value	Test decision
Between Groups	SSG	K−1	$\text{MSG} = \frac{SSG}{K-1}$	SG	From Table	Reject Ho or Fail to
				MSE		
Between Blocks	SSB	H−1	$\text{MSB} = \frac{SSB}{H-1}$	MSB		Reject Ho
				MSE		
Error	SSE	(K−1) (H−1)	$\text{MSE} = \frac{SSE}{(K-1)(H-1)}$			
Total	SST	n−1				

Table 11.12 Two-way analysis of variance table for car servicing cost data

Source of variation	Sum of squares	Degrees of freedom	Mean square	F Ratio	F Critical value	Test decision
Cars	1583.334	2	791.667	463	8.6491	Reject Ho
Mechanic's Training levels	692.3575	4	173.089	101.44	7.0061	Reject Ho
Error	13.65	8	1.70625			
Total	2289.34	14				

11.5 A Summary of This Chapter

In this chapter, we firstly examined two similar distributions: the chi-squared and the F distributions. Both distributions were presented with examples. In the second part we looked into the Analysis of Variance (ANOVA) as it is relevant to the F distribution. The ANOVA examined within an example both with one-way and two-way analysis of variance.

11.5.1 Review Problems for the Chi-Square and F-Distributions, ANOVA

11.1 You are an employment manager of a factory and concerned with the number of health reports submitted by the workers. In the past year a number of sick days claimed caused a serious decline in the production. You look at a random sample of 100 sick days. Your findings are as follows; a 30 % of the factory labour force in the 20–29 age group took 26 of the 100 sick days. The 40 % of the labour force in the 30–39 age group took 37 sick days. The 20 % in the 40–49 age group took 26 sick days, and that 11 % in the 50-and-over age group took 24 sick days. How can you test at the 5 % level of significance the hypothesis that age is not a factor in taking sick days?

11.2 Traffic accidents are a serious problem in Istanbul and you would like to find out more about the types of drivers causing these accidents. The number of car accidents caused by males and females drivers of various age groups in Istanbul is given by a contingency table. Test at the 1 % level of significance the hypothesis that age and sex are independent in the occurrence of traffic accidents.

ch11data2.xls

Contingency table			
Age Group	Male	Female	Total
Below 35	50	30	80
From 35–60	10	10	20
Above 60	30	20	50
Total	90	60	150

11.3 The table below gives the Sample observations on servicing cost recorded for three types of cars serviced by six different class service stations. Assume that the class difference criteria is determined by the training of the car mechanics.

ch11data3.xls

Service class	Talkswagen	Mofas	Sayota	Sum
1	225	192	203	620
2	220	193	200	613
3	215	185	197	597
4	210	180	190	580
5	205	175	185	565
Total	1075	925	975	2975

(a) Find the sample means for each group and obtain sample means for blocks. What is the sample grand mean of all observations?
(b) Estimate the population variance from the variance between the means or columns.
(c) Estimate the population variance from the variance within the samples or columns.
(d) Test the hypothesis that the population means are the same at the 5 % level of significance.

11.4 (a) From the results obtained in problem 11.3, find the value of SSA, SSE, and SST; the degrees of freedom for SSA, SSE, and SST; and MSA, MSE, and the F ratio.
(b) From the results in part a) construct an ANOVA table.
(c) Conduct the analysis of variance and draw a figure showing the acceptance and rejection regions for Ho.

11.5 The table below gives the first- year earnings (in 000 Turkish Liras per annum) of students with master's degrees from 5 Departments and for

3 class rankings at graduation. Test at the 5 % level of significance that the means are identical. a) for department populations and b) for class-ranking populations

ch11data4.xls

Class Ranks	*EC*	*IR*	*LAW*	*MED*	*LIT*	*Sample Mean*
Top 1/3	10	9	8	7	6	**8**
Middle 1/3	9.5	8	6.5	6	5	**7**
Bottom 1/3	9	7	5	5	4	**6**
Sample mean	**9.5**	**8**	**6.5**	**6**	**5**	**7**

First-Year Earnings of MA Graduates of 5 Departments and 3 Class Ranks.

11.5.2 Computing Practical for Chi-Square Test Statistics and ANOVA Using Excel

Chi-Square Using Excel

Use a spreadsheet to solve the following problems.

C.11.1. We are interested in determining whether voters' attitudes regarding a new tax proposition are independent of the income group to which they belong. We have arbitrarily defined 4 mutually exclusive and exhaustive income intervals and 3 attitudinal levels. Every voter will fall into 1 of the 12 categories displayed in the table below. The results from a random sample of 320 voters are reported in the table.

ch11data5.xls

	Income	Groups		
Attitude	1	2	3	4
For	33	38	35	25
Against	37	36	20	15
Indifferent	32	14	21	14

Show the significance level to be used in the spreadsheet, perform a chi-square test and explain the meaning of your conclusions. Do your conclusions depend on the significance level stated?

Hints

Type in a table with data. Calculate row and column totals.

Create another table with the expected numbers in each cell assuming the null hypothesis is true.

Set up another table containing in each cell (observed - expected)^2/expected value. Calculate the value of the test statistics.

To get the critical value of the χ^2 distribution, use Formulas, More Functions then CHISQ.INV. This function needs two parameters. The first parameter is the significance level, the second one is the degrees of freedom.

a. Printout your spreadsheet
b. Show the solution
c. Show the printout giving your calculations (Use options, show formulas and print a few columns, in particular the one with the calculations for the income group nr.

4. Include also the area where you used the CHISQ.INV function)

C.11.2. Data for income group 3 were typed incorrectly. There should be twenty-five people

'For', two 'Against' and thirty-one 'Indifferent'.' Perform a chi-square test again for the revised data.

Describe how you modified your previous spreadsheet and give your solution. No extra printout is needed.

ANOVA in Excel

C.11.3. One-way ANOVA

The table below gives the Sample observations on servicing cost (TL) recorded for three types of cars.

ch11data6.xls

Talkswagen	Mofas	Sayota	Sum
225	192	203	620
220	193	200	613
215	185	197	597
210	180	190	580
05	175	185	565

(a) Enter the labels and data shown in the table in an Excel sheet.
(b) From the Tools menu, choose data Analysis (If you do not have the data analysis option in your Tools menu go to add-ins then chose Analysis Tool pack first). In the Data Analysis dialog box, double-click Anova: Single-Factor in the Analysis Tools list box. For input range select the data with labels. Tick

the labels in the first row. Then choose a cell for output range. Make sure that 16 below and 7 on the right columns of the output range cell are free.

(c) What are your test results for the null hypothesis, which claims that all population means are the same?

C.11.4. Re-questions 11.3 and 11.2: **Two-way ANOVA:**

The table below gives the Sample observations on servicing cost recorded for three types of cars serviced by six different class service stations.

(a) Construct an ANOVA table by using Excel and draw a figure showing the acceptance and rejection regions for Ho.

Service class	Talkswagen	Mofas	Sayota	Sum
1	225	192	203	620
2	220	193	200	613
3	215	185	197	597
4	210	180	190	580
5	205	175	185	565
Total	1075	925	975	2975

ch11data7.xls

Enter the data in a sheet. From the Tools menu, choose Data Analysis dialog box, double-click Anova: Two-Factor Without replication in the Analysis Tools list box. Select sall data including the labels for the input range, click Labels and again choose a cell for the output range. Finally click OK. Comment on the results of the Two-way Anova output.

Conduct the analysis of variance for 5 % level of significance. Print all your work.

Chapter 12
Correlation

12.1 Introduction

So far we have discussed the statistical properties of single variables. In this chapter we will discuss bivariate relationships (relationships between two variables). For example inflation and unemployment, quantity and price, consumption and income etc. In the general sense correlation is about two or more sets of things being related. It is quite often defined as how things (in our case two things) 'move' together. In statistics we need to define and establish a measurement for this. In the first chapter, we examined the relationships between two variables (XY charts) which can give an immediate impression of a set of data. In the case of two variables the appropriate method of illustration is by scatter diagrams.

12.2 Scatter Diagrams

When we deal with two variables then the appropriate method of illustration is by the scatter diagram. It allows us to show two variables together, one on each axis.

Let us take some data to examine two variables. Table 12.1 shows the data on Unemployment and inflation between 1988 and 2000. Data is obtained from the SIS sources as a Semiannual Inflation and Unemployment. For example 1998-1 shows the first quarter of the unemployment and Inflation rates and 1998-2 for the third quarter of the same year. The Inflation rates are the GNP Implicit Price Index where $1987 = 100$.

The online version of this chapter (doi:10.1007/978-3-319-26497-4_12) contains supplementary material, which is available to authorized users.

D. Özdemir, *Applied Statistics for Economics and Business*,
DOI 10.1007/978-3-319-26497-4_12

Table 12.1 Short term relationship between inflation and unemployment in Turkey

Semiannual Data		
Year	UnempR.	GNP PrIn.
1998-1	6.9	87.7
1998-2	6.7	73.2
1999-1	7.9	55.3
1999-2	7.4	63.8
2000-1	8.3	66.1

Source: The TUIK www.tuık.gov.tr

ch12data1.xls

The variable 'Unemployment is on one axis and 'Inflation' on another. Figure 12.1 shows this relationship. From Fig. 12.1 we can infer that, in general, the higher the Inflation the lower the Unemployment rate. This is not a long term deterministic relationship and does not imply a cause and effect between the two variables: it merely means that times of high unemployment were also times of low inflation rates.

From Fig. 12.1 we see a reasonably tidy relationship between the unemployment and the inflation rate. It has a negative relationship: if unemployment was high then inflation was low. If we spread the same data for a longer period, say ten years in annual periods, then from theory we know is that there should not be a relationship between these two variables.

Table 12.2 below shows the data for unemployment rates and inflation rates in Turkey for a longer period and is taken from the same data source. The unemployment figures show percentage annual figures from 1990 till 1999. Inflation rates are taken from the GNP implicit price index, 1987 = 100.

ch12data2.xls

A scatter diagram of the data presented on Table 12.2 is shown in Fig. 12.2 below.

There appears to be little correlation between the unemployment rate and the inflation rate for this ten year period in Turkey.

Let us have a look at another example, which is more substantial: The data is taken from the OECD web site for 1997. The first column in the table below shows the OECD countries. The second column shows the Gross Domestic Products for 29 OECD countries. The total fertility rate is the average number of children per women aged 15–49. The data is mainly from 1997. However, since data was not available for 1997 for some countries and they have been replaced by the most up-to-date data. For Total fertility rate Australia, Canada, Korea, Mexico and the New Zealand has 1996 data Turkey has 1994 data. For Infant Mortality which

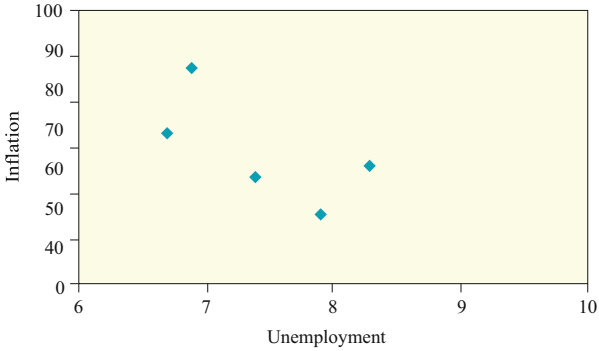


Fig. 12.1 Observations on a pair of linearly related variables: Unemployment and inflation

Table 12.2 Inflation and unemployment for a longer period

Year	UnempR.	GNP PrIn.
1990	8.2	57.6
1991	7.9	59.2
1992	8.1	63.5
1993	7.8	67.4
1994	8.1	107.3
1995	6.9	87.2
1996	6	78
1997	6.7	81.2
1998	6.8	75.3
1999	7.7	56.2

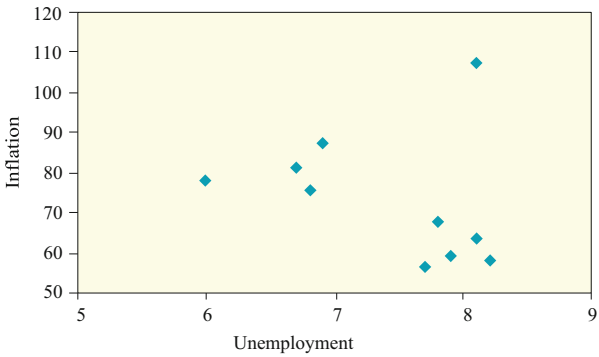


Fig. 12.2 Inflation and Unemployment in the long run

shows per 1000 live births, Belgium, Canada, Denmark, Korea, Luxembourg, Netherlands, Norway, Poland, Spain, Sweden, Switzerland and the USA has 1996 data but Italy has 1995 data (the rest have 1997 data). The GDP growth rates are the average annual real percentage change between 1995 and 1999. The GDP per capita is in US dollars, using purchasing power parities in 1999.

Table 12.3 Fertility rate, GDP per capita, growth and health expenditures

	GDP	GDP pc	Growth	Fertility R.	InfantMort.	HealthExp.
Australia	47,030	24,400	4.2	1.8	5.3	8.4
Austria	207,014	24,600	2.1	1.4	4.7	8.3
Belgium	753,270	24,300	2.3	1.6	6	7.6
Canada	77,956	25,900	3.1	1.6	6	9.1
Czech Rep.	118,815	13,100	0.3	1.2	5.9	7.2
Denmark	91,685	26,300	2.5	1.8	5.6	8
Finland	46,276	22,800	4.8	1.8	3.9	7.4
France	788,682	21,900	2.2	1.7	4.8	9.6
Germany	383,431	23,600	1.6	1.4	4.8	10.7
Greece	2,816,689	14,800	3.2	1.3	6.4	8.6
Hungary	584,000	10,900	3.6	1.4	9.9	6.5
Iceland	41,651	27,300	5.5	2	5.5	7.9
Ireland	3564	25,200	9	1.9	6.2	6.3
Italy	1.64E+08	21,800	1.2	1.2	6.2	7.6
Japan	37,658,200	24,500	1.2	1.4	3.7	7.2
Korea	22,571,360	15,900	3.6	1.6	8	6
Luxembourg	37,247	39,300	5.1	1.7	4.9	7
Mexico	150,000	8100	5	2.7	16.4	4.7
Netherlands	63,495	25,100	3.4	1.5	5.7	8.5
New Zealand	7449	18,000	1.6	2	6.9	7.6
Norway	88,450	27,500	3	1.9	4	7.5
Poland	28,888	8100	5.3	1.5	12.3	5.2
Portugal	1,397,000	16,500	3.4	1.5	6.4	7.9
Spain	5,762,000	18,100	3.4	1.2	5.5	7.4
Sweden	148,240	23,000	2.5	1.5	4	8.6
Switzerland	38,044	27,500	1.4	1.5	4.7	10.3
Turkey	1.15E+09	6300	3.7	2.7	38.2	4
U.K.	53,917	22,300	2.5	1.7	5.9	6.9
USA	1,070,366	33,900	4.1	2.1	7.8	13.9

Source: OECD www.oecd.org/els

ch12data3.xls

From the data in Table 12.3 let us say we would like to see the fertility rate against GDP. The data used in the table indicates the fertility rates and the GDP's of the 29 OECD countries. Our investigation would therefore only explain these countries' fertility and GDP relationship. These countries are randomly chosen, and even the availability of data may affect the choice. They may not represent the link between these two variables for the other countries.

The variables are defined as follows:

Total fertility rate: Average number of children per women aged 15–49 in 1997.
GDP per capita: 1999 gross domestic product p.c., in US $ using PPPs.
Growth rate: The growth rate of GDP per capita per annum, 1995 - 99
Health Expenditures: the percentage of GDP in 1997. Infant mortality: per 1000 live births in 1997.

We will firstly have a look at XY graphs of the data which may provide some useful information. The scatter diagram in Fig. 12.2 plots the relationship between fertility rate and GDP p.c.

There appears to be little correlation between the fertility rate and per capita GDP. They seems to have a zero correlation.

However Fig. 12.4 indicates that high values of health expenditures tend to be associated with low values of infant mortality rate and vice-versa. This is termed **negative** correlation.

Figure 12.5 (rather weakly) indicates that high values of GDP growth rate are associated with high values of fertility. This is a **positive correlation**.

The relationship graphed in these figures can be summarized numerically. This numerical measure is called the correlation coefficient. The correlation coefficient can measure the relationship between any pair of variables.

12.3 Cause and Effect Relationships

It is easy if we know which variable causes the change. For example if we have a scatter diagram which has a wage rise on the one axis and price on the other and say that the diagram shows a strong linear relationship, i.e. high wages corresponds to high prices. Can we say wage rises cause price rises or price rises cause wage rises.

Are they both correct?

Only one of the variables can be causal at a particular point in time, but they could both be effects of a common cause.

For example take a relationship (a scatter diagram for two variables) between the number of ducks in Britain and the number of births in Turkey. Say both are increasing. In a scatter diagram you have a strong upward relationship. There appears to be a correlation. This is called **spurious correlation** as both variables are changing over time. Time itself cannot be a cause, except perhaps in terms of the aging process.

The cause and effect is not always an easy question to answer. The variables can be divided into two types; stable and differential variables. The differential variables are most likely to be identified as a cause. We may have some idea about the cause and effect if we were to lag one of the variables. In our previous example we then relate the wage rises in one month to price rises in the previous month, which will give some evidence of prices as a cause This could also be repeated with wage rises lagged one month behind price increases giving some evidence of wages as a cause.

Table 12.4 Rankings of exam and assignment performance

Exam %	Rank	Assignments %	Rank
81	5	74	7
62	10	71	8
74	7	69	9
78	6	76	5
93	1	87	2
69	9	62	10
72	8	80	3
83	4	75	6
90	2	92	1
84	3	79	4

Correlation should be used as a suggestive and descriptive tool rather than a technique which gives definite answers.

12.4 Spearman's: Rank Correlation Coefficient

If the data is given in the form of ranks rather than the actual values then the rank correlation can be used. Rankings are often used where people or companies are asked to express preferences or to put a series of items in to order.

Table 12.4 provides data on exam and tutorial assignment performances in a tutorial group. The rankings are organized as follows; The highest exam mark (93 %) is the first and the second highest mark (90 %) second, etc. The question we need to answer is: Is there an association between the rankings of the exam and the assignment marks?

ch12data4.xls

We can calculate a rank correlation coefficient between these two sets of data. We are trying to find out whether there are similarities in the rankings of the assignments and exam marks. For this we need to use a formula which is called **Spearman's coefficient of rank correlation**.

$$r = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

where d is the difference in the ranks and n is the sample size. Table 12.5 shows the Excel calculations of the rank correlation.

If we substitute the numbers we have obtained from the table above the following correlation coefficient is derived:

Table 12.5 The rank correlation calculations

Exam %	Rank %	Assignment %	Rank	Difference	
				d	d^2
81	5	74	7	2	4
62	10	71	8	−2	4
74	7	69	9	2	4
78	6	76	5	−1	1
93	1	87	2	1	1
69	9	62	10	1	1
72	8	80	3	−5	25
83	4	75	6	2	4
90	2	92	1	−1	1
84	3	79	4	1	1
					46

$$r = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6\Sigma(46)}{10\left(10^2-1\right)} = 0.721$$

There is a positive rank correlation between the marks obtained from assignments and the exam. This implies that high assignment marks are associated with high exam performance.

Let us take another example by using the previous data on fertility and growth rate in the table below. In the table, the country with the highest growth rate (Ireland) is ranked 1 for variable RankGro ; Iceland, the next fastest growth nation, is ranked 2, etc. For the fertility rate, Turkey and Mexico are ranked one, having the highest fertility rate, 2.7. Spain, Italy and Czech Republic have the lowest birth rate and ranked 11 for variable RankFer. The Excel calculations for the rank differences (d) and the sum of the d^2 are presented in the same Table 12.5 below.

Substitution of the derived values into the formula provides:

$$r = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6\Sigma(3596)}{29\left(29^2-1\right)} = 0.11$$

This indicates that there is a very small positive rank correlation between the two variables. This means that the countries which have higher economic growth rates also tend to have higher birth rates but this rank correlation is small (Table 12.6).

12.5 Covariance

Assume again we are intending to look into a relationship between two variables. X and Y. This gives us the measure of the association.

If we have a positive association then X↑ and Y↑

Table 12.6 Fertility rate and growth rates as a rank

Countries	Growth	Fertility R.	RankGro.	RankFert.	d	d^2
Australia	4.2	1.8	7	5	2	4
Austria	2.1	1.4	23	9	14	196
Belgium	2.3	1.6	21	7	14	196
Canada	3.1	1.6	16	7	9	81
Czech Rep.	0.3	1.2	29	11	18	324
Denmark	2.5	1.8	18	5	13	169
Finland	4.8	1.8	6	5	1	1
France	2.2	1.7	22	6	16	256
Germany	1.6	1.4	25	9	16	256
Greece	3.2	1.3	15	10	5	25
Hungary	3.6	1.4	10	9	1	1
Iceland	5.5	2	2	3	−1	1
Ireland	9	1.9	1	4	−3	9
Italy	1.2	1.2	28	11	17	289
Japan	1.2	1.4	27	9	18	324
Korea	3.6	1.6	11	7	4	16
Luxembourg	5.1	1.7	4	6	−2	4
Mexico	5	2.7	5	1	4	16
Netherlands	3.4	1.5	14	8	6	36
New Zealand	1.6	2	24	3	21	441
Norway	3	1.9	17	4	13	169
Poland	5.3	1.5	3	8	−5	25
Portugal	3.4	1.5	12	8	4	16
Spain	3.4	1.2	13	11	2	4
Sweden	2.5	1.5	20	8	12	144
Switzerland	1.4	1.5	26	8	18	324
Turkey	3.7	2.7	9	1	8	64
U.K.	2.5	1.7	19	6	13	169
USA	4.1	2.1	8	2	6	36
						3596

A negative association means X↑ while Y↑
No linear association = 0

The population covariance between two variables, X and Y is denoted as Cov(X, Y) (or sometimes σ_{XY}) and defined as:

$$Cov\ \ (X,Y) = \frac{\sum (X_i - \mu_X)(Y_i - \mu_Y)}{N}$$

For the sample covariance $\mu_X = \overline{X}$ and $\mu_{Y=\overline{Y}}$

$$\text{Sample} \quad Cov(X,Y) = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

There are two alternative way to express these:

$$Cov(X,Y) = \frac{\sum X_i Y_i - n(\overline{XY})}{n-1}$$

$$\text{Or} \quad Cov(X,Y) = \frac{\sum X_i Y_i - \frac{1}{n}\sum X_i \sum Y_i}{n-1}$$

Sometimes these two equations expressed as

$$\begin{array}{l} \text{Population covariance} \quad Cov(X,Y) = \sigma_{XY} \\ \text{Sample covariance} \quad Cov(X,Y) = S_{XY} \end{array}$$

If small values of X (that is smaller than the average of X) tend to be associated with small values of Y (smaller than average of Y) and large values of X with large values of Y then the covariance will be positive. If on the other hand, there is negative association, that is if small values of X tend to be associated with large values of Y and large values of X with small values of Y, then the covariance will be negative. If there is a positive association than the covariance will be positive. A negative association will give a negative covariance.

12.6 Correlation Coefficient (Pearson)

If the data is provided in the form of actual values rather than ranks then we need to use a different coefficient of correlation. This is called **Pearson's correlation coefficient** (or just 'correlation coefficient' as it is in most text books).

The earlier Figs. from 12.2, 12.3, 12.4, 12.5 can also be summarized numerically by measuring the correlation coefficient. The correlation coefficient can be negative, positive or zero. It has to be between zero and one or minus one. The correlation coefficient only represents a linear relationship between two variables. The Correlation coefficient formula is (for a population) equal to:

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

For a sample

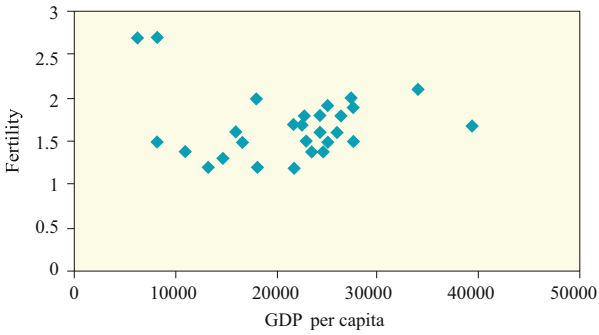


Fig. 12.3 Fertility and GDP per capita

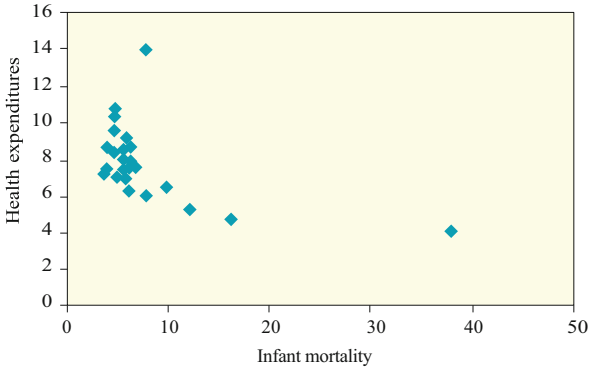


Fig. 12.4 Health expenditures and the infant mortality rate

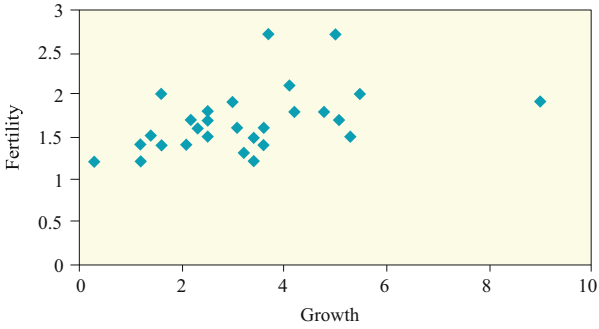


Fig. 12.5 GDP growth and the fertility

$$r = \frac{est.Cov(X,Y)}{\sqrt{est.Var(X)est.Var(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

The correlation coefficient formula can be calculated by substituting the values of the variance and the covariance as follows:

$$r = \frac{\sum X_i Y_i - \frac{1}{n}\sum X_i \sum Y_i}{n-1}$$
$$\sqrt{\left(\frac{\sum X_i{}^2 - \frac{1}{n}(\sum X_i)^2}{n-1}\right).\left(\frac{\sum Y_i{}^2 - \frac{1}{n}(\sum Y_i)^2}{n-1}\right)}\, r$$

$$= \frac{\sum X_i Y_i - \frac{1}{n}\sum X_i \sum Y_i}{\sqrt{\left(\sum X_i{}^2 - \frac{1}{n}(\sum X_i)^2\right).\left(\sum Y_i{}^2 - \frac{1}{n}(\sum Y_i)^2\right)}}\, r$$

$$= \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{\left(n\sum X_i{}^2 - (\sum X_i)^2\right).\left(n\sum Y_i{}^2 - (\sum Y_i)^2\right)}}$$

This last equation is called the Pearson correlation coefficient formula.

Let us now have a look at a calculation of the correlation coefficient using sample data.

Example Table 12.7 below indicates data on sales and prices of a particular good.

ch12data5.xls

The summary of the data calculations are:

$$\sum X = 3240 \quad \sum X^2 = 1320200 \quad \sum Y = 45$$
$$\sum Y^2 = 257 \quad \sum XY = 18060 \quad n = 8$$
$$(\sum X)^2 = 3240^2 = 10497600 \quad \left(\sum Y^2\right) = 45^2 = 2025$$

Substitution of these values into the formula for the correlation coefficient gives:

$$r = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{\left(n\sum X_i{}^2 - (\sum X_i)^2\right).\left(n\sum Y_i{}^2 - (\sum Y_i)^2\right)}}$$
$$r = \frac{8*(18060) - (3240*45)}{\sqrt{(8*1320200) - (3240)^2)*(8*257) - (45)^2}} = -0.937$$

This result indicates a strong negative correlation between the sales and prices. If the prices are higher than the sales tend to be lower. Higher prices are associated with lower sales or lower sales are associated with higher prices.

Table 12.7 Excel calculations for correlation coefficient

Sample data on Sales (S) and Prices (P)					
	Sales	Prices	S × P	S^2	P^2
	1	2	3	4	5
	420	5.5	2310	176,400	30.25
	380	6.0	2280	144,400	36.00
	350	6.5	2275	122,500	42.25
	400	6.0	2400	160,000	36.00
	440	5.0	2200	193,600	25.00
	380	6.5	2470	144,400	42.25
	450	4.5	2025	202,500	20.25
	420	5.0	2100	176,400	25.00
Total =	3240	45.0	18,060	1,320,200	257.00
Mean =	405	5.625			

Let us take another example. Table 12.8 below presents data concerning a country import data and Real GDP in Million Euro (M €). We would like to see the correlation coefficient of these two variables, i.e. Imports and Real GDP.

ch12data6.xls

Taking the calculated values from Table 12.9 and putting them into the correlation formula is as follows:

$$r = \frac{9*(3775.576) - (1214.435*39.211)}{\sqrt{(9*115905.5) - (1214.435)^2)*(9*124.8762) - (39.211)^2}} = 0.88$$

From the correlation coefficient we can obtain the r square value which is the coefficient of determination. (The use of the r square will be discussed later on in the topic about regression.)

$$r^2 = 0.774$$

The correlation coefficient has the following properties:

- It is always between zero and plus or minus one.
- A positive value of the coefficient indicates positive correlation and a higher value means a stronger correlation between these two variables. If the coefficient is equal to 1, this mean that all the observations lie on the straight line with positive slope.
- Negative value means a negative correlation. Larger negative value mean a higher negative correlation. If the coefficient is equal to −1 then this means there is a perfect negative correlation.

Table 12.8 Correlation data of a country Imports and the real GDP

	Imports	Real GDP
Year	(Million €)	(Million €)
1987	2.106	74.722
1988	2.409	76.306
1989	2.42	76.498
1990	2.483	83.579
1991	2.594	84.353
1992	2.588	89.4
1993	2.836	96.592
1994	3.154	91.321
1995	3.097	97.888
1996	3.191	104.745
1997	3.526	112.631
1998	4.079	116.114
1999	4.728	110.286
Totals	39.211	1214.435

Table 12.9 Excel correlation calculations of imports and GDP

	Imports	Real GDP			
Year	(Million €)	(Million €)	Im^2	GDP^2	Im * GDP
1987	2.106	74.722	4.435236	5583.377	157.3645
1988	2.409	76.306	5.803281	5822.606	183.8212
1989	2.42	76.498	5.8564	5851.944	185.1252
1990	2.483	83.579	6.165289	6985.449	207.5267
1991	2.594	84.353	6.728836	7115.429	218.8117
1992	2.588	89.4	6.697744	7992.36	231.3672
1993	2.836	96.592	8.042896	9330.014	273.9349
1994	3.154	91.321	9.947716	8339.525	288.0264
1995	3.097	97.888	9.591409	9582.061	303.1591
1996	3.191	104.745	10.18248	10,971.52	334.2413
1997	3.526	112.631	12.43268	12,685.74	397.1369
1998	4.079	116.114	16.63824	13,482.46	473.629
1999	4.728	110.286	22.35398	12,163	521.4322
Totals	39.211	1214.435	124.8762	115,905.5	3775.576

- $r = 0$ (or very close to zero) means that there is a lack of correlation between these two variables.

The relationship is symmetric. For example the correlation between X and Y is the same as the correlation between Y and X.

12.7 A Review of this Chapter

In this chapter we have examined the correlation. Two correlations were considered; Spearman's rank correlation and Pearson's correlation coefficients. We used some Turkish data to explore these coefficients.

12.8 Review Problems For Correlation

12.1. The following data on the rankings of exam and assignment performance is given. Calculate the rank correlation coefficient and comment.

Exam %	Rank	Assignments %	Rank
83	5	77	7
64	10	74	8
76	7	73	9
80	6	79	5
95	1	90	2
71	9	65	10
74	8	83	3
85	4	78	6
92	2	95	1
86	3	82	4

12.2. Draw a scatter diagram of the following data, treating nominal GNP as the dependent variable. Calculate r and r^2, and comment.

ch12data7.xls

	Nominal GNP(Million $)	Money Supply, M1
Year	At current purchasing price	(Million $)
1991	150,168	8,559,576
1922	158,121	7906.05
1993	178,717	8,399,699
1944	132,298	5,556,516
1995	167,182	6,520,098
1996	178,133	6,412,788
1997	192,226	7,112,362

(continued)

Year	Nominal GNP(Million $) At current purchasing price	Money Supply, M1 (Million $)
1998	206,279	7,219,765
1999	185,136	7,590,576

12.3. The data below show real GDP and Imports data for 13 years:

ch12data6.xls

Year	Imports (Million €)	Real GDP (Million €)
1987	2.106	74.722
1988	2.409	76.306
1989	2.42	76.498
1990	2.483	83.579
1991	2.594	84.353
1992	2.588	89.4
1993	2.836	96.592
1994	3.154	91.321
1995	3.097	97.888
1996	3.191	104.745
1997	3.526	112.631
1998	4.079	116.114
1999	4.728	110.286
Totals	39.211	1214.435

(a) Draw an *XY* scatter diagram of the data and comment.
(b) From the chart, would you expect the line of best fit to slope up or down? *In theory*, which way should it slope?
(c) What would you expect the correlation coefficient to be, approximately? Which variable is more likely to depend on the other, and why?
(d) Calculate the correlation coefficient between Import and real GDP.
(e) Calculate r and r^2. What information do they give?

Chapter 13
Simple Regression

13.1 Introduction

In Chap. 12 we discussed bivariate relationships, (relationships between two variables) and examined correlation. We have emphasized that correlation is a measure of the strength of any linear association between a pair of random variables. In this chapter we will continue to discuss linear relationships between pairs of variables but in terms of one variable depending on the other variable. We no longer have a symmetric relationship.

Thus regression analysis is different from correlation because it asserts the direction of causality. This is not something the data teaches us but the theory. We have two types of variables: dependent and explanatory or independent and exogenous variable.

We can have more than one exogenous variables but in this chapter we will focus on one independent variable and one endogenous variable.

13.2 The Linear Regression Model

Let us call our variables X and Y with Y depending on X, i.e. any change in X will cause a change in Y.

$X \Rightarrow Y$

'Explanatory' 'Explained'

'Independent' 'Dependent'

'Exogenous' 'Endogenous'

'Regressor' 'Regressand'

The online version of this chapter (doi:10.1007/978-3-319-26497-4_13) contains supplementary material, which is available to authorized users.

D. Özdemir, *Applied Statistics for Economics and Business*,
DOI 10.1007/978-3-319-26497-4_13

We can have more than one independent variable (multiple regression), in this case we would look into the influence of each independent variable on the dependent variable. The objective is to find the line of best fit.

Example

The following table provides data on Infant mortality (per 1000 live births in 1997) and Health Expenditures (the percentage of GDP in 1997) for several countries.

ch12data3.xls

Plotting the scatter diagram of the data from the Table 13.1 gives Fig. 13.1. We can see a downward slope from the data. The best fitting line to this data can be seen in Fig. 13.2.

Q How is the line of best fit determined?

It allows each observation on Y (the explained variable), to be broken down into two parts:

a) The part explained by regression line b) The residual / error (e).

Errors are the difference between the line and the observation points in the scatter diagram.

The estimated regression line is :

$$\widehat{Y_i} = a + bX_i \tag{13.1}$$

Thus the regression equation is:

$$Y_i = a + bX_i + e_i \tag{13.2}$$

The errors are :

$$Y_i - \widehat{Y_i} = e_i \,\text{Or}\, Y_i - \widehat{Y_i} = (a + bX_i + e_i) - (a + bX_i) = e_i \tag{13.3}$$

This is called the residual or error or disturbance.

Question

Why should we actually fit a regression line?

We need a rationale from the theory. The most popular regression line, for example, is the consumption function;

$C = \langle\ \alpha + \beta Y + \varepsilon$

where C is consumption, Y income, alpha the autonomous part of the consumption and beta the marginal propensity to consume. However this is as a population regression line, which is the 'true' model has some unknown variables, such as α, β, ε.

We use a sample regression line where :

a = estimate of α

b = estimate of β

Table 13.1 Health expenditures and infant mortality infant

Countries	Health Exp.	Infant Mort.
Australia	8.4	5.3
Austria	8.3	4.7
Belgium	7.6	6
Canada	9.1	6
Czech Rep.	7.2	5.9
Denmark	8	5.6
Finland	7.4	3.9
France	9.6	4.8
Germany	10.7	4.8
Greece	8.6	6.4
Hungary	6.5	9.9
Iceland	7.9	5.5
Ireland	6.3	6.2
Italy	7.6	6.2
Japan	7.2	3.7
Korea	6	8
Luxembourg	7	4.9
Mexico	4.7	16.4
Netherlands	8.5	5.7
New Zealand	7.6	6.9
Norway	7.5	4
Poland	5.2	12.3
Portugal	7.9	6.4
Spain	7.4	5.5
Sweden	8.6	4
Switzerland	10.3	4.7
Turkey	4	38.2
U.K.	6.9	5.9
USA	13.9	7.8

Source: OECD www.oecd.org/els

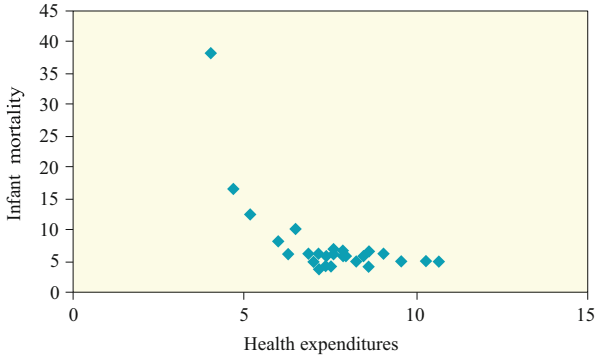


Fig. 13.1 Scatter diagram of the health expenditures and the infant mortality rates

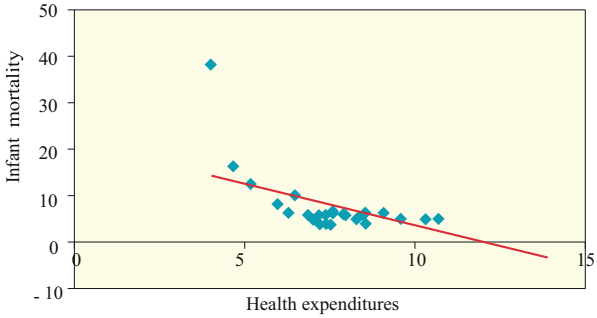


Fig. 13.2 Line of best fit

e_i = estimate of ε_i

Now we need a method to determine the regression line, i.e. find 'a' and 'b' to obtain the 'best' possible line through the data. It must be satisfied that we want the part of Y explained by the regression line is as big as possible. We want to make sure that the unexplained part (e) is as small as possible. This could be achieved by minimizing the deviations of errors from the line, i.e. obtaining the best fitting line.

The sum of the deviations from the regression line would be zero, therefore we would not be able to use the sum of the deviations (Σe_i) but we could take the square of the sum of these errors. Σe^2. Basically minimizing the sum of these squared residuals provides the minimum error-best fitting line. This is called 'the method of least squares'or OLS (Ordinary Least Squares).

13.3 The OLS (Ordinary Least Squares) method

The OLS is a method of least squares. The first concept is the Error Sum of Squares (ESS)

$$\text{ESS} = \sum_{i=1}^{n} e_i^2$$

The OLS method minimizes these errors. For this purpose we need to obtain the function for the Error Sum of Squares. In our earlier discussion we said that the regression equation was:

$Y_i = a + bX_i + e_i$

$e_i = Y_i - a - bX_i$

Thus the Error Sum of Squares (ESS):

$$\text{ESS} = \sum_{i=1}^{n} e_i{}^2 = \sum_{i=1}^{n} \left(Y_i - a - bX_i\right)^2$$

For the constant 'a' and the coefficient 'b' we need to find the partial derivatives:

For a: $\frac{\partial ESS}{\partial a} = \sum -2(Y_i - a - bX_i)$

$$= -2\sum \left(Y_i - a - bX_i\right) = 0 \tag{13.3}$$

For b : $\frac{\partial ESS}{\partial b} = \sum 2X_i\left(Y_i - a - bX_i\right)$

$$= -2\sum X_i\left(Y_i - a - bX_i\right) = 0 \tag{13.4}$$

Dividing the equations (13.3) and (13.4) by 2 (two equations—two unknowns) provides the following two equations which are derived from them: Equation (13.3) can be rewritten:

$$\begin{gathered}
\sum Y_i - \sum a - \sum bX_i = 0 \\
\sum Y_i - na - b\sum X_i = 0 \\
N.B.\ \ \overline{Y} = \frac{\sum Y}{n} \Rightarrow \sum Y = n\overline{Y} \\
n\overline{Y} - na - bn\overline{X} = 0 \\
\text{Dividing by n gives :} \quad \overline{Y} - a - b\overline{X} = 0 \\
\text{Hence :} \qquad\qquad a = \overline{Y} - b\overline{X}
\end{gathered} \tag{13.5}$$

Equation (13.4) can be rewritten:

$$\begin{gathered}
\sum X_iY_i - a\sum X_i - b\sum {X_i}^2 = 0 \\
\sum X_iY_i = a\sum X_i + b\sum {X_i}^2 \\
= (\overline{Y} - b\overline{X})\sum X_i + b\sum {X_i}^2 \\
= \overline{Y}\sum X_i - b\overline{X}\sum X_i + b\sum {X_i}^2 \\
\sum X_iY_i - \overline{Y}\sum X_i = b(\sum {X_i}^2 - \overline{X}\sum X_i) \\
\sum X_iY_i - \frac{\sum Y_i\sum X_i}{n} = b(\sum {X_i}^2 - \frac{\sum X_i\sum X_i}{n})
\end{gathered}$$

Multiply by n gives: $n\sum X_iY_i - \sum Y_i\sum X_i = b(n\sum {X_i}^2 - \sum X_i\sum X_i)$

Dividing both side by (n − 1) gives: $\frac{n\sum X_iY_i - \sum Y_i\sum X_i}{n-1} = b\left(\frac{n\sum {X_i}^2 - (\sum X_i)^2}{n-1}\right)$

Or Cov (X, Y) = b Var (X)

Hence b $\Rightarrow$ $b = \frac{Cov(X,Y)}{Var(X)}$

$$b = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum {X_i}^2 - \left(\sum X_i\right)^2} \tag{13.6}$$

Can also be written as:

$$b = \frac{\sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(X_i - \overline{X}\right)^2} \tag{13.7}$$

Closely related to the correlation coefficient:

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \qquad b = \frac{Cov(X,Y)}{Var(X)}$$

$$\text{Or}\, b = r\frac{\sqrt{Var(Y)}}{\sqrt{Var(X)}}$$

A Numerical Example

Let us take the relationship between health expenditures and the infant mortality. (The data presented in Table 13.1). The Infant mortality, the dependent variable (Y), is per 1000 live births in 1997. Health expenditures is in terms of the percentage of GDP in 1997 is the explanatory variable.

Let us remind ourselves of the formulae for the constant 'a' and the coefficient 'b' again:

$$a = Y - bX, b = \frac{n\Sigma X_i Y_i - \Sigma X_i \Sigma Y i}{n\Sigma X_i - \left(\Sigma X_i\right)2}$$

The Excel calculations of the data for these formulae are presented in Table 13.2. The summary of these calculations is as follows:

n=29	$\Sigma XY = 1503{,}2$	$\Sigma X = 225{,}9$	$\Sigma Y = 215{,}6$
	$\Sigma X^2 = 1859{,}61$	$(\Sigma X)^2 = 51030{,}81$	$\overline{X} = 7{,}7897$
	$\overline{Y} = 7{,}4345$		

Substitution of these calculations into the formulae gives:

$$b = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum {X_i}^2 - \left(\sum X_i\right)^2}$$

$$b = \frac{29{*}1503,2 - (2.15,6{*}225,9)}{29{*}1859,61 - 51030,81} = -1.764$$

a= $\overline{Y} - b\,\overline{X}$, a = 7,4345 – (−1,764) * (7,7897) = 21,1755

Table 13.2 Regression calculations in Excel

Countries	HealthExp.	InfantMort.			Fitted Line
	X	Y	X * Y	X^2	Y=a+b*X
Australia	8.4	5.3	44.52	70.56	6.3562
Austria	8.3	4.7	39.01	68.89	6.5326
Belgium	7.6	6	45.6	57.76	7.7674
Canada	9.1	6	54.6	82.81	5.1214
Czech Rep.	7.2	5.9	42.48	51.84	8.473
Denmark	8	5.6	44.8	64	7.0618
Finland	7.4	3.9	28.86	54.76	8.1202
France	9.6	4.8	46.08	92.16	4.2394
Germany	10.7	4.8	51.36	114.49	2.299
Greece	8.6	6.4	55.04	73.96	6.0034
Hungary	6.5	9.9	64.35	42.25	9.7078
Iceland	7.9	5.5	43.45	62.41	7.2382
Ireland	6.3	6.2	39.06	39.69	10.0606
Italy	7.6	6.2	47.12	57.76	7.7674
Japan	7.2	3.7	26.64	51.84	8.473
Korea	6	8	48	36	10.5898
Luxembourg	7	4.9	34.3	49	8.8258
Mexico	4.7	16.4	77.08	22.09	12.883
Netherlands	8.5	5.7	48.45	72.25	6.1798
New Zealand	7.6	6.9	52.44	57.76	7.7674
Norway	7.5	4	30	56.25	7.9438
Poland	5.2	12.3	63.96	27.04	12.001
Portugal	7.9	6.4	50.56	62.41	7.2382
Spain	7.4	5.5	40.7	54.76	8.1202
Sweden	8.6	4	34.4	73.96	6.0034
Switzerland	10.3	4.7	48.41	106.09	3.0046
Turkey	4	38.2	152.8	16	14.1178
U.K.	6.9	5.9	40.71	47.61	9.0022
USA	13.9	7.8	108.42	193.21	−3.3458
Totals	225.9	215.6	1503.2	1859.61	
Mean	7.7896552	7.43448276			

The estimated regression line is therefore: $\hat{Y} = 21,1755 - 1.764X$

The interpretation :

- An increase in health expenditures by one percent of the countries' GDP, on average lowers the infant mortality per 1000 live births by 1.764.
- 21,1755 is a constant and represents the rate of infant mortality not affected by health expenditure.

The fitted line is represented by the estimated regression equation, this line's data is presented in the last column of the Table 13.2. The figure of these data is presented as a straight line in Fig. 13.2.

We have so far assumed that there is a linear relationship between these two variables. But why? We could equally assume that there is a nonlinear relationship between Y and X.

13.4 Nonlinear Regression

Linearity of a functional form does not necessarily provide a best fit to some of these relationships.

For example the linear relationship ($Y_i = a + bX_i$) could easily be in the following forms:

$$y_i = AX_i^b \text{ or } y_i = a + bX + cX^2$$

(Quadratic function)

In order to use OLS the function has got to be linear. Simply we linearise the function by taking the logarithmic forms of these variables. The linearisation of the first function above is as follows:

Take the logs: $1nY = 1n\ (AX^b)$

$1nY = 1n\ A + 1nX^b$

$1nY = 1n\ A + b*1nX$

We can simply rewrite this as:

$y = a + b\ x$

Where $y = ln\ Y$, $x = ln\ X$, $a = ln\ A$

Now we have a linear regression that we can estimate.

Table 13.3 converts these variables into logarithmic form: basically, each number's logarithmic value is considered. The scatter diagram of the logarithmic data in Fig. 13.3 provides the same diagram as in Fig. 13.1. We can also see that the best fitting line in Fig. 13.4, is also similar to diagram 13.2.

Ln Y = Ln infant mortality and Ln X = Ln health expenditures.

To calculate the regression line we can use the same OLS formulae.

To calculate the regression line, $Y_i = a + bX_i$, we calculate the constant 'a' and the coefficient 'b' and use the same formulae (Table 13.4).

The summary of the table above:

$$\Sigma y = 53,60638 \quad \Sigma x = 58,72009 \quad \Sigma xy = 106,38706 \quad \Sigma x^2 = 120,55138$$
$$\overline{x} = 2,024831 \quad \overline{y} = 1.848496$$

$$\text{a} = \overline{y} - b\,\overline{x},\ b = \frac{{}^{n}\sum X_i Y_i - \sum X_i \sum Y_i}{{}^{n}\sum X_i{}^2 - \left(\sum X_i\right)^2}$$

Table 13.3 Logarithmic form of the data

Countries	HealthExp.	InfantMort.		
	X	**Y**	Ln Y	Ln X
Australia	8.4	5.3	1.667707	2.128232
Austria	8.3	4.7	1.547563	2.116256
Belgium	7.6	6	1.791759	2.028148
Canada	9.1	6	1.791759	2.208274
Czech Rep.	7.2	5.9	1.774952	1.974081
Denmark	8	5.6	1.722767	2.079442
Finland	7.4	3.9	1.360977	2.00148
France	9.6	4.8	1.568616	2.261763
Germany	10.7	4.8	1.568616	2.370244
Greece	8.6	6.4	1.856298	2.151762
Hungary	6.5	9.9	2.292535	1.871802
Iceland	7.9	5.5	1.704748	2.066863
Ireland	6.3	6.2	1.824549	1.84055
Italy	7.6	6.2	1.824549	2.028148
Japan	7.2	3.7	1.308333	1.974081
Korea	6	8	2.079442	1.791759
Luxembourg	7	4.9	1.589235	1.94591
Mexico	4.7	16.4	2.797281	1.547563
Netherlands	8.5	5.7	1.740466	2.140066
New Zealand	7.6	6.9	1.931521	2.028148
Norway	7.5	4	1.386294	2.014903
Poland	5.2	12.3	2.509599	1.648659
Portugal	7.9	6.4	1.856298	2.066863
Spain	7.4	5.5	1.704748	2.00148
Sweden	8.6	4	1.386294	2.151762
Switzerland	10.3	4.7	1.547563	2.332144
Turkey	4	38.2	3.642836	1.386294
U.K.	6.9	5.9	1.774952	1.931521
USA	13.9	7.8	2.054124	2.631889

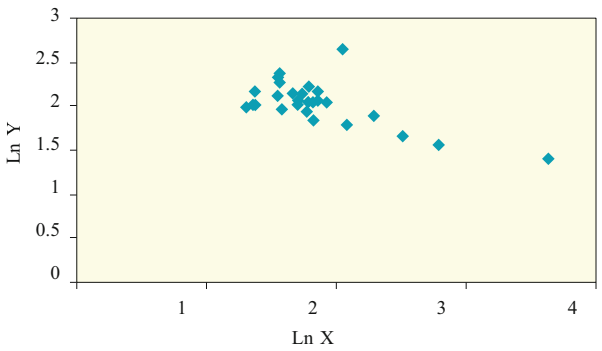


Fig. 13.3 Logarithmic scatter diagram.

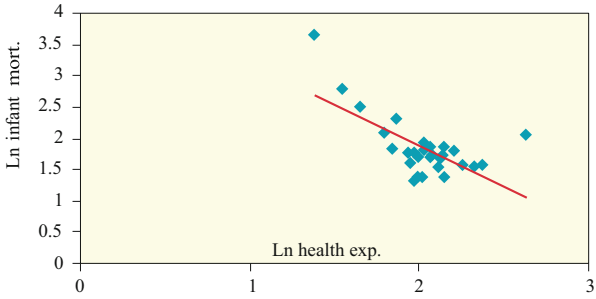


Fig. 13.4 Ln infant mortality versus Ln health expenditures and the fitted line

$$b = \frac{29*1503,2 - (2.15,6*225,9)}{29*1859,61 - 51030,81} = -1.764$$

a = 1,848496 – 1.305* 2,025 = 4,49

The estimated regression line is therefore $\hat{Y} = 4{,}49–1{,}30465x$

INTERPRETATION

In a 'double-log' regression, the estimated coefficient is the elasticity of Y with respect to X. This implies that a 1% increase in health expenditures is associated with a 1,305% decrease in the infant mortality.

In general there are many functional forms that the underlying relationship might take.

13.5 Goodness of Fit

This question is: how well does the regression equation fit the data? (Figs. 13.5 and 13.6

Total = Explained + Error $\Rightarrow (Y - \overline{Y}) = (\hat{Y} - \overline{Y}) + e$

The part explained by the regression line, $\hat{Y} - \overline{Y}$ (i.e. explained by the value of Xi) This is the length $\hat{Y} - Y$ bar distance in the figure.

The error term $e_i = Y_i - \hat{Y}_i$.

Thus in algebraic terms: $Y_i - \overline{Y} = \left(Y_i - \hat{Y}\right) + \left(\hat{Y}_i - \overline{Y}\right)$ A good regression model should explain a large part of the differences between the Y_i values and $\overline{Y}$. For example the length $\hat{Y}_i - \overline{Y}$ should be large relative to $Y_i - \hat{Y}_i$.

Table 13.4 Logarithmic calculations in Excel for non-linear regression.

Countries	HealthExp.	InfantMort.					y = 4.49-
	X	**Y**	Ln Y	Ln X	LnX*LnY	Ln X^2	1.30466*x
Australia	8.4	5.3	1.667707	2.128232	3.5492665	4.5293702	1.713589
Austria	8.3	4.7	1.547563	2.116256	3.2750377	4.4785374	1.729213
Belgium	7.6	6	1.791759	2.028148	3.6339538	4.1133853	1.844162
Canada	9.1	6	1.791759	2.208274	3.9566966	4.8764759	1.609161
Czech Rep.	7.2	5.9	1.774952	1.974081	3.5038998	3.8969959	1.914701
Denmark	8	5.6	1.722767	2.079442	3.5823924	4.3240771	1.777243
Finland	7.4	3.9	1.360977	2.00148	2.7239674	4.0059222	1.878955
France	9.6	4.8	1.568616	2.261763	3.5478376	5.1155723	1.539377
Germany	10.7	4.8	1.568616	2.370244	3.7180021	5.6180554	1.397848
Greece	8.6	6.4	1.856298	2.151762	3.9943119	4.6300806	1.682889
Hungary	6.5	9.9	2.292535	1.871802	4.2911715	3.5036434	2.048139
Iceland	7.9	5.5	1.704748	2.066863	3.5234803	4.2719217	1.793654
Ireland	6.3	6.2	1.824549	1.84055	3.3581735	3.387623	2.088913
Italy	7.6	6.2	1.824549	2.028148	3.7004564	4.1133853	1.844162
Japan	7.2	3.7	1.308333	1.974081	2.582755	3.8969959	1.914701
Korea	6	8	2.079442	1.791759	3.7258591	3.210402	2.152567
Luxembourg	7	4.9	1.589235	1.94591	3.0925089	3.7865663	1.951454
Mexico	4.7	16.4	2.797281	1.547563	4.3289677	2.3949497	2.471159
Netherlands	8.5	5.7	1.740466	2.140066	3.7247128	4.5798832	1.698149
New Zealand	7.6	6.9	1.931521	2.028148	3.9174118	4.1133853	1.844162
Norway	7.5	4	1.386294	2.014903	2.7932487	4.0598342	1.861443
Poland	5.2	12.3	2.509599	1.648659	4.1374725	2.7180753	2.339264
Portugal	7.9	6.4	1.856298	2.066863	3.8367132	4.2719217	1.793654
Spain	7.4	5.5	1.704748	2.00148	3.4120192	4.0059222	1.878955
Sweden	8.6	4	1.386294	2.151762	2.9829758	4.6300806	1.682889
Switzerland	10.3	4.7	1.547563	2.332144	3.6091385	5.4388951	1.447554
Turkey	4	38.2	3.642836	1.386294	5.0500423	1.9218121	2.681557
U.K.	6.9	5.9	1.774952	1.931521	3.4283585	3.730775	1.970227
USA	13.9	7.8	2.054124	2.631889	5.4062253	6.9268389	1.056492
Total			53.60638	58.72009	106.38706	120.55138	
Mean			1.848496	2.024831			

The measure of therefore: $\frac{\left(\hat{Y}-\overline{Y}\right)}{\left(Y_i-\overline{Y}\right)}$ However this measure has some problems. It only applies to one observation and often takes a negative value. In order to solve these problems we square and then sum each of the terms.

This provides:

$\sum \left(Y_i-\overline{Y}\right)^2 = \sum Y_i^2 - n\overline{Y}^2 \Rightarrow$ Total sum of squares (TSS)

$\sum \left(\hat{Y}_i-\overline{Y}\right)^2 \Rightarrow$ The regression sum of squares (RSS)

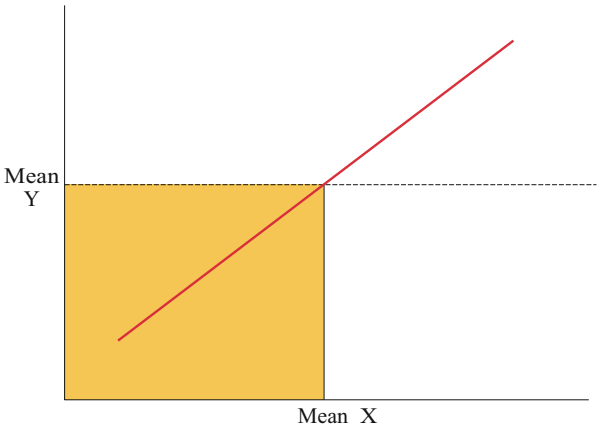


Fig. 13.5 There is no gap from the mean

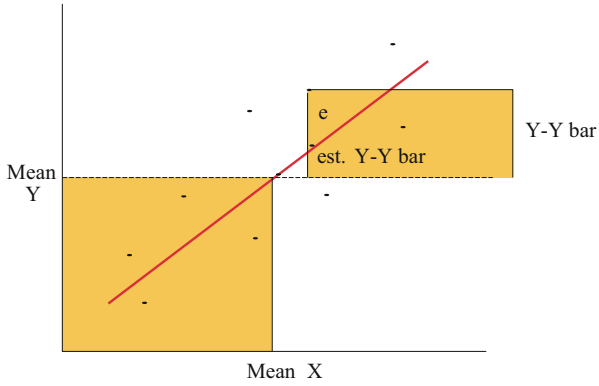


Fig. 13.6 Total gap from the mean

$$\sum \left(Y_i - \hat{Y}\right)^2 = \sum Y_i^2 - a\sum Y_i - b\sum X_i Y_i$$

⇒ The error sum of squares (ESS)

The measure of goodness of fit (R^2) is then defined as: $R^2 = \frac{RSS}{TSS}$ because TSS = RSS + ESS

For example in our data TSS = $\sum Y_i^2 - n\overline{Y}^2$ =2777,98 – 1602,8745 = 1175,1055

The ESS = $\sum Y_i^2 - a\sum Y_i - b\sum X_i Y_i$ = 2777,98– 21,1755 * 215,6 – (−1,764)*1503,2

ESS = 864,187

Since TSS = RSS + ESS then the RSS = TSS – ESS = 1175,1055 – 864,187

RSS = 310,9185

Hence the R^2 = (RSS/TSS) = 310,9185/1175,1055 = **0.2646**.

INTERPRETATION

Countries' infant mortality vary around the overall mean value of 7,435 and 26.4% of this variation is explained by variations in countries health expenditures. This means that 73,6% of the variation in Y is explained by other factors. (other than the health expenditures).

13.6 Inference in Regression

In this part we will review simple regression analysis and discuss goodness of fit, confidence interval, hypothesis testing and regression significance.

13.6.1 Review of Regression

Regression is a line which fits through data and asserts a direction of causality (i.e. dependent independent variable). A simple regression line has two characteristics which are the intercept and slope (Fig. 13.7).

$\Rightarrow Y_i = a + bX_i + e$

OLS

Method Min: $\Sigma\,(Y_i - Y)^2 = \min \Sigma e_i$

$\rightarrow$ leads to

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2} = \left(\frac{Cov(X,Y)}{Var(X)}\right), a = \overline{Y} - b\overline{X}$$

This implies that the regression line will pass exactly through the sample means (Fig. 13.8), since

$$\overline{Y} = a + b\overline{X}$$

Goodness of fit is about how well the regression line fits the data. Figures 13.9 and 13.10 provide diagrammatic examples of good and bad fit.

MEASURE: R^2

To calculate, we distinguish between fluctuations in Y around the mean due to changes in the regression line, and errors (Fig. 13.11).

$$\text{So} \quad Y_1 = \hat{Y}_1 + e_i$$

$$Y_1 - \overline{Y} = \hat{Y}_1 - \overline{Y} + e_i$$

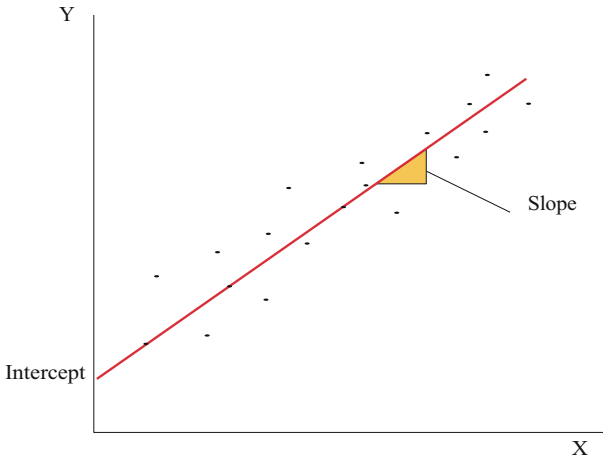


Fig. 13.7 Scatter diagram

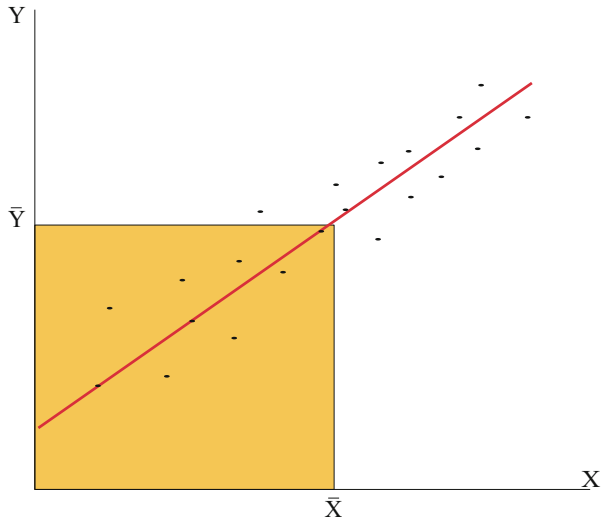


Fig. 13.8 Regression passes through the sample mean

$$R^2 = \frac{RSS}{TSS} = \frac{\sum \left(\hat{Y}_i - \overline{Y}\right)^2}{\sum \left(Y_i - \overline{Y}\right)^2} = 1 - \frac{ESS}{TSS} = 1 - \frac{\sum e_i^2}{\sum \left(Y_i - \overline{Y}\right)^2}$$

$$0 \leq R^2 \leq 1$$

To calculate R^2:

$$\text{TSS} = \sum Y^2 - n\overline{Y}^2$$

$$\text{ESS} = \sum Y^2 - a\sum Y - b\sum XY$$

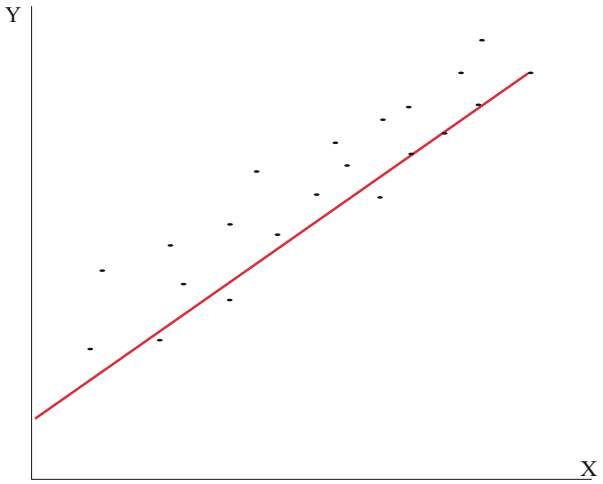


Fig. 13.9 Good fit

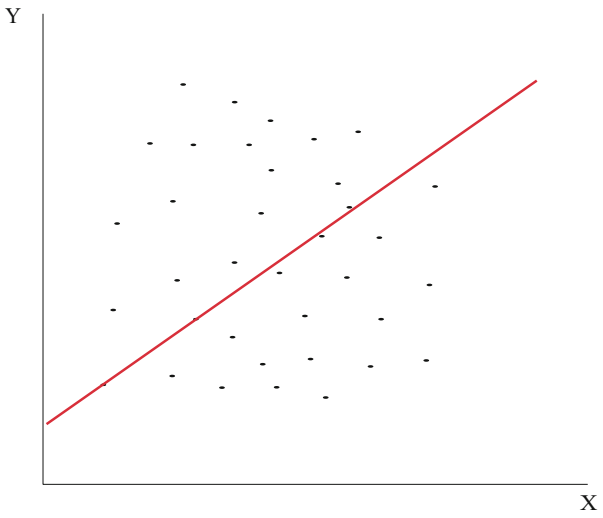


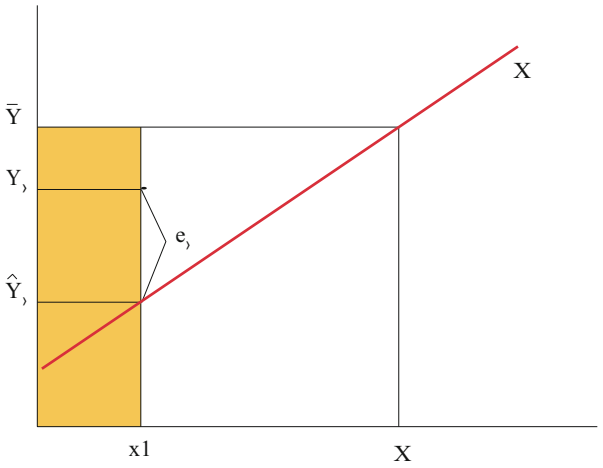
Fig. 13.10 Bad fit

$$\text{RSS} = \text{TSS} - \text{ESS}$$
$$\text{RSS} = \sum Y^2 - n\overline{Y}^2 - (\sum Y^2 - a\sum Y - b\sum XY)$$
$$\text{RSS} = a\sum Y - b\sum XY - n\overline{Y}^2$$

$$R^2 = \frac{RSS}{TSS}$$

→Where does regression come from?
→ estimate of population regression line:

Fig. 13.11



$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

estimated by $Y_i = a + bX_i + e_i$

$$a = Est(\alpha)$$
$$b = Est(\beta)$$
$$e_i = Est(\varepsilon_i)$$

It comes from theory, i.e. a hypothesized relationship. The interpretation of regression coefficient is as follows:

a = value of Y (predicted) where X = 0

b = shape of Y with respect to X, i.e. dY/dX. All results also hold for logged data.

Use of log → If underlying relationship is nonlinear.

(How do we know?)

- data.
- theory.

Interpretation of b if data in logs: elasticity.

i. e. $d1nY = \Sigma Y, X$

$d1nX$

Compared to unlogged data: to calculate $\sum = \frac{dYX}{dXY} \Rightarrow use \quad \frac{b\overline{X}}{\overline{Y}}$ calculated at the mean.

We have two further issues now, they are inference and multiple regression. We will leave multiple regression for another chapter but we willl discuss inference here.

13.6.2 Inference

We need to recognize that a and b are estimates (point estimates).

The basic point here is that if, for instance, we take a relationship between inflation and GNP, for 30 countries in 1980, 1983, 1986 and 1989, we would see that for the same two variables we would have four slightly different regression lines. Inference looks at how statistically significant these differences are.

If we assume that the underlying relationship has stayed the same then we wouldn't need to view the data for different samples.

Thus, when you estimate a regression, it is always like a drawing from a population. That is, usually, observations have one draw.

We would like our OLS estimates to have good criteria i.e. unbiasedness and efficiency. The OLS is an optimal estimator under certain conditions. It gives us a good point estimate, and a variance that is as small as possible.

Question

What are the distributions associated with a and b?

The key quantity is the estimated residual (e_i) and we need to make number of assumptions on the estimated residual:

- e_i are normally distributed.
- They are independent of each other.
- They have constant variance.
- They are not correlated with the independent variable.

Now let us denote the error variance:

$$\frac{\sum e_i^2}{n-2} = \frac{ESS}{n-2} = S_e^2$$

Thus the variance of a and b are given by; $S_b^2 = \frac{S_e^2}{\sum \left(X_i - \overline{X}\right)^2}$

Estimated standard error $= S_a^2 = S_e^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum \left(X_i - \overline{X}\right)^2} \right)$

$\Rightarrow$ Now we would like to consider what factors affect the precision of estimates a and b.

Remember that: $\Sigma(X_i - X)^2 = (n-1)Var(X_i)$

So: variance of a, b would be smaller if ;

$\Rightarrow$ n $\uparrow$

$\Rightarrow$ Var (X) $\uparrow$

$\Rightarrow s_e^2 \downarrow$ (better fit)

$\Rightarrow$ the smaller the variance, the smaller the associated confidence intervals! In regression analysis, distributions are assumed to have a t distribution.

To get a confidence interval for b;

$\left[b - t_v^*.S_b\,, \quad b + t_v^* S_b\right]$ here v = n − 2

For a;

$$\left[a - t_v^*.S_a\,, \quad a + t_v^* S_a\right]$$

HYPOTHESIS TESTING

$H_0 : \beta = 0$

$H_1 : \beta \neq 0$ (Or $\beta > 0$, $\beta < 0$)

The test statistic is;

$$\frac{b - \beta}{S_b} < t\, n - 2$$

Illustration

Take a regression of infant mortality on Income (Y). Let us say the regression is estimated for 1990 data and n = 30. The estimated regression equation is as follows:

$\hat{Y}_i = 71.738 - 0.00541 X_i$

To get the standard errors:

$$\sum \left(X - \overline{X}\right)^2 = 647953880,\ \overline{X} = 17405584$$

$$S_e^2 = \frac{ESS}{n-2} = \frac{23964.21}{28} = 255.86$$

$$S_b^2 = \frac{S_e^2}{\sum \left(X_i - \overline{X}\right)^2} = \frac{855.86}{647953880} = 0.00000132$$

$$S_b = 0.001149$$

$$S_a^2 = S_e^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum \left(X_i - \overline{X}\right)^2}\right) = 855.86\left(\frac{1}{30} + \frac{17405584}{647955880}\right) = 57.579$$

$$S_a = 7.178$$

Let us now establish confidence intervals for both a and b; Firstly we need to find the t*, v = 28 ® 95% ± 2.0480.

So for a: 95% confidence interval:

= [71.738 − (2.048 7.178), 71.738 + (2.048 7.178).

= [57.035, 86.441]

For b: 95% confidence interval:

= [−0.00541 − (2.048 0.001149), − 0.00541 + (2.048 0.001149)]

$= [-0.0078 - 0.0031]$

HYPOTHESIS TEST

The most common test is;

$H_0 : \beta = 0$ $H_0 : \alpha = 0$

$H_1 : \beta \neq 0$ $H_1 : \alpha \neq 0$

i.e. Are the estimated coefficients significantly different from zero. (particularly b)

To test:

$$t = \frac{a - 0}{s_a} = \frac{71.738}{7.178} = 9.99$$

$t^*_{28} = \pm 2.048 \Rightarrow$ 'highly significant'

$$t = \frac{b - 0}{s_b} = \frac{-0.00541}{0.00115} = -4.71$$

Same result, i.e. significant.

The table value is greater than the calculated values hence Ho is rejected, this implies that these parameters are statistically significant.

13.6.3 Testing the Significance of R^2; the F-Test

The significance of R^2 can be tested by using the F distribution.

H_0: $R^2 = 0$ (X does not influence Y)

$H_1 : R2 \neq 0$

The test statistics are:

$$F = \frac{R^2 / 1}{(1 - R^2) / (n-2)} \; Or \; F = \frac{RSS / 1}{ESS / (n-2)}$$

i.e. If $R^2 = 0.678$ and n = 12 then F is;

$$F = \frac{0.678 / 1}{(1 - 0.678) / 10} = 21.078$$

For 5% significance level, the critical value for F distribution with $v_1 = 1$ and $v_2 = 10$ is: $F^*_{1,10} = 4.96 \Rightarrow$ The test statistics exceeds this. So the regression as a whole is significant or the relationship between dependent and independent variables (or variables in multiple regression case) is significant.

The F test statistics are the square of the t statistics in a single regression, i.e.

$F_{1,n-1} = t^2_{n-2}$

F test in a way becomes more crucial in multiple regression (discussed in the next chapter).

IN PRACTICE

We estimate the regression on the Excel, Eviews, STATA etc. They provide;

- Coefficient, Standard error
- t statistics, Confidence interval
- ESS and sample size n.

R^2, and $S_e \rightarrow$ Standard error.

13.7 A Review of This Chapter

In this chapter, simple regression was examined and one dependent and one independent variable model with the method of OLS was used. Firstly we have examined the calculation of the coefficients and then we looked at inference in regression. For the inference section we particularly focused on the t test and the F test of the estimated regression equation.

13.7.1 Review Problems For Simple Linear Regression

13.1 Use the data below and calculate the linear regression line of GNP on M1. Add the regression line to your scatter diagram which you have drawn in question 12.2. Comment.

ch12data7.xls

Year	Nominal GNP(Million $) At current purchasing price	Money Supply, M1 (Million $)
1991	150168	8559.576
1922	158121	7906.05
1993	178717	8399.699
1944	132298	5556.516
1995	167182	6520.098
1996	178133	6412.788
1997	192226	7112.362
1998 1999	206279 185136	7219.765 7590.576

Source: TUIK general indicators.

13.2 The data below shows the number of tourist arrivals to Turkey and the reel exchange rate (1982 January = 100). The reel exchange rate based on 0.75.

USD + 0.25 DM basket, in the relative price calculations, producers prices for USA, industrial products producer prices for Germany and wholesale prices for Germany and wholesale prices for Turkey are used.

Draw a scatter diagram of the data. Which variable is more likely to depend on the other, and why? Find the regression line for predicting the number of visitors to Turkey on the basis of the exchange rate. Use the regression equation to predict the number of visitors at an exchange rate of 50.50 and 45.00. Comment.

ch13data1.xls

Year	Visitors(000)	ExchangeR.
1983	1506.5	224.03
1984	1855.3	364.85
1985	2190.2	518.34
1986	2397.3	669.4
1987	2907.1	855.69
1988	4265.2	1420.76
1989	4516.1	2120.78
1990	5397.7	2607.62
1991	5552.9	4169.85
1992	7104.1	6887.51
1993	6525.2	10985.96
1994	6695.7	29704.33
1995	7747.4	45705.43
1996	8531.5	81137.16
1997	9700	146881.53

Source: TUIK Tourism statistics and IFS

13.3 Suppose that you regressed milk consumption (M shows milk in litres per capita per week) against per capita income (Yd in € per week) and got M= 1500 + 600 .06 Yd. as your regression line with $r^2 = 0.65$
What result would you have got if

a) Income had been measured in Euro per week?
b) Milk had been measured in gallons per week?
c) Milk had been measured in five kg boxes per year and income in thousands € per year?

13.4 Aggregate consumption expenditures (C) and aggregate disposable income (Yd) is given by the table below;

a) Calculate the parameters a and b and write the equation for the estimated consumption regression. Explain what is the meaning of estimator a and estimator b

b) Calculate i) S^2, ii) S_a^2 and S_a iii) $S_b^2$and S_b.
c) State the null and alternative hypothesis to test the statistical significance of the parameters of the regression equation estimated in a) and test at the 5% level of significance for parameters a and b.

ch13data2.xls

d) Construct the 95% confidence interval for parameters a and b.

Year	n	C	Y
1991	1	102	114
1992	2	106	118
1993	3	108	126
1994	4	110	130
1995	5	122	136
1996	6	124	140
1997	7	128	148
1998	8	130	156
1999	9	142	160
2000	10	148	164
2001	11	150	170
2002	12	154	178

e) Find R^2 for the estimated consumption regression.
f) The overall significance of the regression can be tested with the ratio of the explained to the unexplained variance. This follows an F distribution with $k-1$ and $n-k$ degrees of freedom, where n is the number of observations and k is the number of parameters estimated:

$$F_{k-1,n-k} = \frac{\sum \hat{y}_i/(k-1)}{\sum e_i^2/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

Test the overall significance of the regression estimated in a) at the 5% level and Comment.

13.8 Computing Practical for Simple Regression and Correlation Using Excel

C.13.1

Investigate what linear relationship may exist between investment and change in sales for the firm to which the following data relates:

Year	Investment	Change in Sales				Reg. Line
	Y	*X*	*X*Y*	*X^2*	*Y^2*	*Y=a + b*X*
1988	10	30				
1989	24	20				
1990	41	70				
1991	14	50				
1992	6	−20				
Total =						
Mean =						

a) Complete the rest of the table in Excel. Use the appropriate formulae and calculate the regression line, Y = a + b X, (Use hints 22 in your earlier computing practical sheet) and the correlation coefficient.
b) Now select two adjacent blank columns each of which has five cells. Go to Formulas, More Functions in your toolbar. Choose 'LINEST' function under the statistical functions. Click 'LINEST'. You will have 4 arguments. The first one, 'known y's', is highlighted. Select the investment column without its heading. Click on 'known x's' then. Select the change in sales column without the heading. Click on the third argument 'constant' and then, type 'true' in the space of 'constant'. Do the same typing for the fourth argument. Finally Press 'Ctrl + Shift + Enter' simultaneously. (p.s. If you can't do this simultaneously then you may not have a successful outcome, so be careful). Use the hints about the inbuilt function LINEST at the end of this practical sheet. What are the values of the correlation and regression coefficients? Have you obtained the same results as you calculated previously?
c) Estimate your regression function for each of the X values in the last column of the table above.

Now copy the following columns in the same order in a separate space on your worksheet; Change in sales, Investment and Regression Line columns. Select these columns and plot a scatter diagram. In order to obtain a best fitting straight regression line, double click on the dots of this data in your scatter diagram after you have completed it. Double click again. In the 'Format Data Series' dialogue box choose 'Patterns'. Under the 'line' chose 'custom' and under the 'Marker' choose 'None' in this box. Click back to the figure. This exercise should provide you a scatter diagram and a straight regression line in the middle. If you couldn't obtain what you supposed to then first read the hints and try again.

C.13.2

The data below is taken from a country macroeconomic variables:

ch13data3.xls

Year	Investment	GDPFC
1993	525.506	1,981,867
1994	952.322	3,868,483
1995	1,882,225	7,762,456
1996	3.757812	14,772,110
1997	7,728,372	28,835,883
1998	13,034,212	52,224,945
1999	17,261,864	77,374,802

Source: SPO and TUIK,(Values are in Current prices in million Turkish Liras)

a) Investigate whether a linear relationship exists between investment and GDP (at factor cost), by looking at a scatter diagram and calculating a linear regression, Calculate the correlation coefficient, and obtain a standard errors of the regression coefficients, You should input your own formulae to calculate the linear regression and correlation coefficient, For the standard errors you may EITHER input your own formulae OR use an inbuilt function (see hints),
b) Calculate the estimated values by inputting an appropriate formula, Chart your scatter diagram and add the regression line as an overlay,
c) Calculate the residuals for the regression equation and chart them, plotting time on the X axis,
d) Obtain the sum of squared residuals,

C,13,3

The following table gives data on weekly family consumption expenditure (Y) and weekly family income (X),

Income	80	100	120	140	160	180	200	220	240	260
Expenditure	70	65	90	95	110	115	120	140	155	150

a) Draw the scatter diagram with Y on the vertical axis and X on the horizontal axis,
b) What can you say about the relationship between Y and X?

Show all your calculations,

HINTS

Drag and drop

Put a side any variables you do not immediately require, using Drag and Drop to put them, say in columns to the right of P, Or delete them with the Edit command and save the file under a new name,

Y depends on X

Decide which variable is to be identified as Y and which as X, You may find it useful to insert a blank row and then to label the variables appropriately,

Plan the use of your spreadsheet

You will want to form columns of other variables adjacent to those you have, Plan to place single calculations and constants separately, perhaps arranging them below or to the side of the array of columns,

Regression and correlation calculation

Calculate the means of Y and X, then form columns of deviations from means and of the product XY, Remember that you can simultaneously copy down several columns, Sum appropriate columns, You can find a sum of squared deviations about the mean by using the DEVSQ function, Use the results of these computations to find estimates of ⟨, ® and r,

Square root

the square root of an expression can be found by either =SQRT(expression) or= (expression)^0,5,

Scatter diagram

Check that your columns are in the appropriate order before using Charts, Once you have created the basic chart, double click on it to open it so that you can make various alterations, Click on an axis of the graph to select it and then use Format Scale, Alter Minimum and Axis Crosses to change the scale which was automatically selected, Aim to display your points clearly,

In-built regression function, linest

The Linest array function calculates simple and multiple regression equations and automatically produces values for R^2, the standard errors and the Sum of Squares, You must select a range of appropriate size into which the array function will be entered, For a simple regression, this is 2 columns and 5 rows, You can then paste the function and define appropriate columns of data (not including either headings or totals) as the 'known y's' and 'known x's', (For a multiple regression, a larger area would be required for the function, and the various known x's should be in adjacent columns,) The third argument of the function specifies whether or not a constant term should be included, Since we want one, you should type TRUE for this, (There may be instances in which a user would want to force a regression to go through the origin, and this would happen if FALSE were typed,) The final argument offers you statistics about the regression, and to obtain them you should type TRUE, Once your formula is complete, enter it as an array function (Ctrl + Shift + Enter),

Because Linest is a function, if you alter data in the spreadsheet its result will automatically change to correspond,

Unfortunately, Linest output does not include any labelling, The outline below should help you to identify the items in each cell,

	Slope	Constant	
Coefficient			
se			
r-sq			se-yest
F			df
ESS			RSS

Predicted values

A column of predicted values can be formed by typing in a formula which, as it is copied down the column, will substitute each X value (or set of X values for a multiple regression) in turn into the regression equation, To ensure accuracy, it is important that the regression coefficients should be brought into this formula by using their cell addresses,

Residuals and squared residuals

A column of residuals and another containing the squares of these values can now be formed, Each residual is the difference between the actual Y value and its prediction from the regression line, The sum of the squared residual can then be obtained,

Chapter 14
Multiple Regression

14.1 Introduction

In this chapter, we will examine multiple regression analysis and discuss regression significance, regression specification, model building, autocorrelation, multicollinearity and some applications of multiple regression analysis.

14.2 Multiple Regression

Multiple regression is not very different from simple regression. Much of the analysis is simply an extension of the simple regression model, but there are some differences. Multiple regression includes more than one independent variable. In another words the variable that is to be explained is supposed to be a function of several variables.

Two Issues

a) When evaluating the influence of a given explanatory variable or the dependent variable we face with the probability of discriminating between its effects and the effects of other explanatory variables.
b) We have to tackle the problem of model specification.

Simple regression has only one independent factor affecting the dependent variable. This situation is quite often not true: For example, the main determinant of quantity demanded is the price, but it is not the only factor in a demand equation. For example, another factor, income affects the quantity demanded, too.

The online version of this chapter (doi:10.1007/978-3-319-26497-4_14) contains supplementary material, which is available to authorized users.

D. Özdemir, *Applied Statistics for Economics and Business*,
DOI 10.1007/978-3-319-26497-4_14

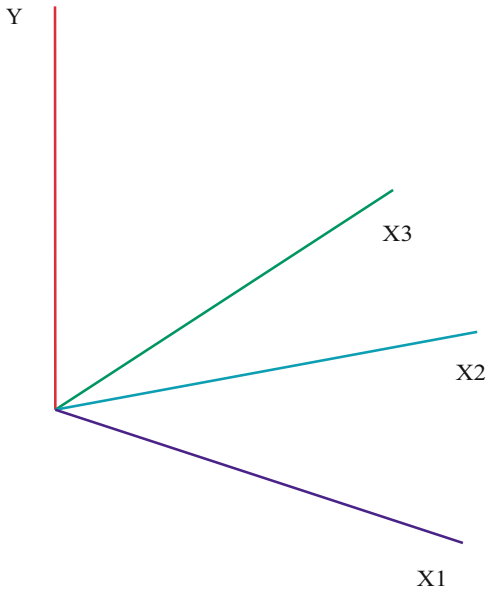


Fig. 14.1 A regression plane in four dimensions (3 independent variable and one dependent variable)

A general equation (population) for a multiple regression is:

$$Y_i = \mu + \beta_1 + x_{1i} + \beta_{2i}x_{2i} + \ldots\ldots\ldots\ldots + \beta_K x_{Ki} + \varepsilon_i$$

This equation represents the relationship between a dependent variable Y and K independent variables, X_1, X_2, ... X_K where the independent variables take the specific values x_{1i}, x_{2i}, ... x_{Ki}. The α, β_1, β_2..... β_K are constants and ε_i is a random variable with mean 0.

Firstly, we will take a simple version of multiple regression which is a regression with two independent variables.

$$Y_i = \alpha + \beta_1 + X_{1i} + \beta_{2i}X_{2i} + \varepsilon_i$$

So we have these coefficients α, β_1 and β_2) to estimate a, b_1 and b_2.

Q. How do we obtain these estimates?

The method is still the same: Use OLS. If the model estimate is;

$$Yi = a + biX_{1i} + b_2X_{2i} + ei$$

There are estimates of a, b_1 and b_2 so that ESS $= \Sigma e_i^2$ is minimized; the same estimation as in the simple regression. We will not go through the proof of OLS here again because we have already done this in simple regression analysis.

Example 1
A demand equation: A simple demand equation consists of two related variables: the quantity demanded Q_i and the price P_i. So a simple demand equation is $Q_i = \alpha + \beta P_i$ There are also other factors that affect the quantity demanded apart from the price. eg. Income (Y).

Hence the regression equation becomes or our hypothesized model becomes;

$$Q_t = \alpha + \beta_1 P_t + \beta_2 Y_t + \varepsilon$$

Let us look at the demand for food data between (1987 and 1999). The demand for food:

Q_t = Expenditure on food and beverages (with 1987 prices)
Pt = CPI (Consumer Price Index) (P_i/P, 1987:100)
Y_t = Wages and salaries.

Data source: Turkish Statıstical Instıtute (TUIK). The ultimate regression equation:

$$\hat{Q}_t = -84905154 - 158.2Pt + 7.7Yt$$
$$(46153591) \quad (144.1) \quad (1.12)$$

The numbers in brackets, which are below the each coefficient of the equation, give the standard error for each variable. What does the demand equation above tell us? What problems are there in terms of statistical point of view? Can we use the estimation for prediction? One way to explain all of these is to do all steps of a regression equation estimation.

A better way to illustrate multiple regression is to choose a case, for example what determines the volume of imports. Several factors can be included in the import demand equation. Before going into the details of the data on imports, we will outline the typical steps to estimate a regression equation.

14.2.1 The Regression Equation Estimation Steps

1. The first step concerns the theoretical considerations. For example, for an estimation of the import function, the first step is to look at the question of 'what can economic theory tell us about the determinants of the demand function? The theory says us that the import demand can be determined by several factors such as prices and income.
2. Obtaining data: After the theoretical suggestions of the parameters, we more or less know about which data we are looking for, but there might be some definitional problems in this step.

3. Transformation of the data: If the existing data is not suitable for our model, we may need to transform them into the needed form. For example, if the values are given in a nominal form, we may need to convert them into real numbers.
4. Estimation of the model: The estimation is done by computer programs. A standard spreadsheet package such as Excel can perform multiple regression analysis, but when the problem of autocorrelation present the other more specific packages are much beneficial to use and they give more detailed results. For this purpose, using soft wares such as Eviews, Stata. etc. are better. Eviews is especially good for time series and autocorrelation problems.
5. Conclusion: This last part is about the interpretation of results. Firstly, we need to comment on what the results are telling us. It may be necessary to improve the model. The most important part of this step is to look at the policy conclusions. If the result is not satisfactory the researcher may like to re-collect more data and re-estimate the regression until satisfactory conclusions are reached.

Example 2
We will now examine these five steps as methodological issues within the example data. Our example data is taken from the TUIK and the Turkish Central Bank data sources. We will now follow the five steps in the estimation of the import demand equation.

The first step was the theoretical definition of the import equation: what can economic theory tell us about the determinants of the demand function? Prices and the GDP are the most important theoretical parameters. Next, we need to look at the data sources to obtain these parameter values, which are presented on Table 14.1 in an annual form.

ch14data1.xls

The variable definitions from the Table 14.1 are as follows:

GDP
The Gross Domestic Product of Turkey at factor cost, current prices in Million Turkish
Liras.

GDP Deflator
An index of ratio of nominal GDP to real GDP 1987 = 100.

Imports
Imports of goods and services into Turkey, at current prices, in Million Turkish Liras.

Import price
The unit value index of imports, 1989 = 100.

Table 14.1 Original data for the import regression

Year	GDP (million TL) (At current prices)	GDP Deflator	Imports (Million TL) (At current prices)	CPI INDEX 1987 = 100	ImportPrice Index 1989 = 100
1982	**10,492**	19.2	**1594**	**19**	**118.4**
1983	**13,906**	24.3	**2348**	**25**	**110.5**
1984	**21,997**	36	**4410**	**37**	**105.9**
1985	**35,095**	55	**6696**	**53**	**107.5**
1986	**51,079**	74.8	**8356**	**72**	**90.1**
1987	**74,722**	100	**13,269**	**100**	**97.2**
1988	**129,225**	169.4	**22,683**	**174**	**96.6**
1989	**227,324**	297.2	**40,420**	**284**	**100**
1990	**393,060**	470.3	**69,092**	**454**	**105.3**
1991	**630,117**	747	**104,819**	**755**	**102**
1992	**1,093,368**	1223	**189,646**	**1284**	100.1
1993	**1,981,867**	2051.8	**383,358**	**2136**	93.9
1994	**3,868,429**	4236.1	**788,530**	**4406**	**94.7**
1995	**7,762,456**	7929.9	**1,890,238**	**8512**	**110.6**
1996	**14,772,110**	14,103	**4,110,584**	**15,269**	**103.9**
1997	**28,835,883**	25,602	**8,762,823**	**28,291**	**94.9**
1998	**52,224,945**	44,977	**14,573,224**	**51,948**	**91.9**
1999	**77,374,802**	70,158	**20,801,155**	**84,995**	**86**

Source: TUIK and TCMB

CPI (Consumer Price Index)
The price of consumer products, 1987 = 100.

The next step is the transformation of the data. All the variables are presented in nominal terms, where we must transform them into real variables. All the transformed variables are presented in Table 14.2.

Real imports are obtained by dividing imports in current prices by import prices. The real GDP is obtained by dividing GDP in current prices by the GDP deflator. The real Unit Value Index (UVI) denotes the real import prices which are obtained by import prices divided by consumer price index for all items.

Firstly, we can look at the diagram of the data. This will give some ideas about the relationship. If any unusual situation exists, that might avoid further mistakes. For this purpose, we are plotting the real import variable against the real GDP in Fig. 14.2. The scatter diagram shows us an increase in imporst as real GDP increases. Of course an increase in import is not only effected by GDP and the prices.: it has also got to do with some trade liberalization and the customs union with the European Union. But we will examine the effect of GDP and prices in this regression analysis.

The following step is the estimation. The model to be estimated is:

$$\left(\frac{imp}{P_m}\right)_t = b_0 + b_1\left(\frac{GDP}{P_{gdp}}\right)_t + b_2\left(\frac{P_m}{CPI}\right) + e_t$$

Table 14.2 Import data transformations

Year	Real imports M/Pm	Real GDP GDP/Pgdp	Real UVI Pm/P
1982	1346.28	54618.00	623.158
1983	2124.89	57333.00	442.000
1984	4164.31	61181.00	286.216
1985	6228.84	63776.00	202.830
1986	9274.14	68248.00	125.139
1987	13651.23	74722.00	97.200
1988	23481.37	76306.00	55.517
1989	40420.00	76498.00	35.211
1990	65614.43	83578.00	23.194
1991	102763.73	84353.00	13.510
1992	189456.54	89401.00	7.796
1993	408261.98	96591.00	4.396
1994	832661.03	91321.00	2.149
1995	1709075.95	97888.00	1.299
1996	3956288.74	104745.00	0.680
1997	9233743.94	112631.00	0.335
1998	15857697.50	116114.00	0.177
1999	24187389.53	110286.00	0.101

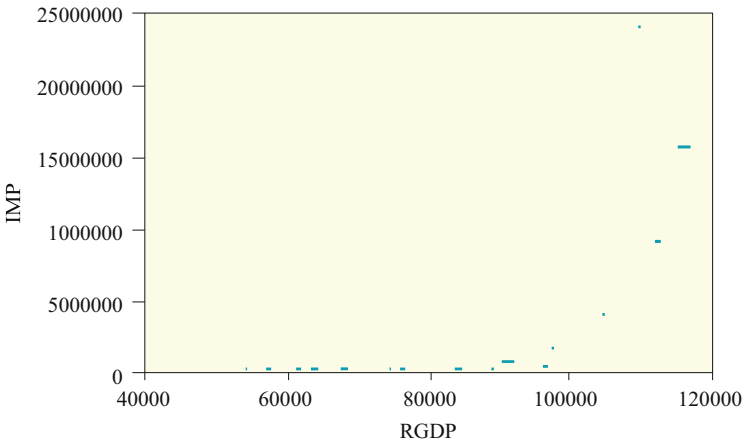


Fig. 14.2 Scatter diagram of imports against real GDP

This equation can be re-presented in simplified notation:

$$RM_t = b_0 + b_1 RGDP_t + b_2 RPM_t + e_i$$

The estimation of the above equation can be done using a number of different software packages, which provide the estimated results. As was mentioned earlier, we will be using Eviews for these estimations. The Eviews output of the estimated

Table 14.3 Regression results using Review

Dependent variable: IMP Method: Least squares Sample: 1982 1999 Observations: 18			
Variable	Coefficient	Std. Error t-Statistic	Prob.
C	−31,939,347	8,472,133. -3.769930	0.0019
RGDP	388.4154	89.66526 4.331839	0.0006
RIMP	21510.49	9801.968 2.194507	0.0444
R-squared	0.595338	Mean dependent var	3,146,869.
Adjusted R-squared	0.541383	S.D. dependent var	6,699,406.
S.E. of regression	4,536,925.	Akaike info criterion	33.64441
Sum squared resid	3.09E + 14	Schwarz criterion	33.79280
Log likelihood	−299.7997	F-statistic	11.03397
Durbin-Watson stat	0.601465	Prob(F-statistic)	0.001130

import demand equation is presented in Table 14.3 below. We can summarize the results presented on Table 14.3 in the following way:

$$RM_t = 31939347 + 388.42\ RGDP_t + 21510.5 RPM_t + e_i$$
(89.7) (9801.1)
$$R^2 = 0.59\ F = 11.03\ n = 18$$

The next step is to interpret these results. As expected (theoretically), we have obtained a positive coefficient on income, but not a negative coefficient on price as it should be. Two dimensions of the coefficient are important. These are the size and their significance.

In general, the sizes of the coefficients depend on the units used for measuring the variables. The GDP measured in 1987 prices and the import is measured in 1989 prices. This may make the use of the coefficient more difficult. It is perhaps better to express the GDP and import in proportionate terms rather than absolute values. For example, we can use the elasticity of import with respect to GDP. The mathematical formula for this elasticity is as follows:

$$\sum_{gdp} = \frac{\Delta RM/RM}{\Delta GDP/GDP} = \frac{\Delta RM}{\Delta GDP}\frac{GDP}{RM} = b\frac{\overline{GDP}}{\overline{RM}}$$

Substituting the values into the last equation gives:

$$\sum_{gdp} = 388.42\frac{84421.67}{3146869.14} = 10.4(?!)$$

The elasticity of GDP with respect to import is high. This shows us that imports are highly responsive to changes on GDP. For example a 6 % rise in real GDP (Turkish growth in the year 2000) leads to a 60 % rise in imports. This is of course

the case under the condition that the prices or other conditions are not changing. This is an extremely high proportional increase.

Similarly, we can obtain the elasticity for the price variable.

$$\sum_{rpm} = 21510.5\frac{106.72}{3146869.14} = 0.73(?!)$$

This suggests that the imports are inelastic with respect to price, contrary to what economics theory suggests. The numerical result suggest that a 10 % import price rise (relative to domestic prices) increases the import demand by 7 %. These results are probably not correct because of the remaining problems with the estimated regression equation. Hence the calculated elasticity do not match with the theoretical requirements.

The second dimension of the regression result are the significance tests. We can test each coefficient's significance by using the t tests and also test the significance of the regression as a whole by using the F tests. Next, we will perform these tests.

14.2.2 t Test

The *t* test in regression is about testing whether each coefficient and constant are significantly different from zero or not. Here, we need our previous knowledge about the t distribution and the tests for the constant term.

$H_0 : \alpha = 0$

$H_1 : \alpha > 0$

The test statistic for the first coefficient is:

$$t = \frac{a - \alpha}{s_a} = \frac{-31939347 - 0}{8472133} = 3.77$$

This is already provided in Table 14.3 (The Eviews results).

The degrees of freedom is $n - k - 1 = 18\text{-}2\text{-}1 = 15$ where k is the number of independent variables. The critical value for a one tail test at 5 % level is 2.131. This is taken from the t distribution table for 15 degrees of freedom. In simple regression, the degrees of freedom was $n - 2$, but in general we will use $n - K - 1$ where K is the number of explanatory variables in the regression. So we can conclude that the estimated coefficient is statistically significant (since the calculated value is greater than the table value). Ho is rejected. A similar test should be carried out for the other coefficients. For the income coefficient, the hypothesis is:

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 > 0 \end{aligned}$$

The test statistics:

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{388.42 - 0}{89.67} = 4.33$$

This is also greater than 2.131. We reject the Ho here again. For the price coefficient, the hypothesis is:

$$\begin{aligned} H_0 &: \beta_2 = 0 \\ H_1 &: \beta_2 > 0 \end{aligned}$$

The test statistics:

$$t = \frac{b_2 - \beta_2}{s_{b_2}} = \frac{21510.49}{9801.968} = 2.195$$

This is also marginally greater than 2.131. We reject the Ho here, too.

14.2.3 F-Test for Overall Significance of the Regression

After investigating the individual coefficient's significance, we shall also test the significance of the overall regression. By this, we mean to show that the relationship between the parameters of the equation namely RM, RGDP and RPM is not a random event. To do this, we have to remember the calculation of R^2 because the hypothesis here is whether the slope coefficients are simultaneously zero or not.

If we remember that TSS = ESS + RSS and $R^2 = RSS/TSS$ shows the goodness of fit.

Here we would like to test:

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = 0 \\ H_1 &: \beta_1 \neq \beta_2 \neq 0 \end{aligned}$$

In the general model:

$$Y_i = \alpha + \beta_1 x_{1i} + \beta_{2i} x_{2i} + \ldots\ldots\ldots\ldots\ldots\beta_K x_{Ki} + \varepsilon_i$$

Then

$H_0 : \beta_1 = \beta_2 = \ldots\ldots\ldots\ldots\ldots \beta_K = 0$

There is no relationship between the explanatory variables and the dependent variable. RSS is the part that is explained and ESS is the unexplained part. We use the ratios: RSS/ESS this ratio will be very small if there is not a good relation. If this ratio is big, then the relation is good.

We could also use the information from ^{2} discussion. Take for instance:

$$\frac{{}^{R}SS/_{K}}{{}^{E}SS/_{n-K-1}} = F$$

Rearranging this equation gives:

$$\frac{RSS(n-K-1)}{ESS\,K}$$

Dividing by TSS gives:

$$\frac{RSS/TSS}{ESS/TSS} \cdot \frac{(n-K-1)}{K}$$

So if $\frac{RSS}{TSS} = R^2$ then what is the $\frac{RSS}{TSS}$?
What we know from our previous argument is that TSS = RSS + ESS.

$$\frac{TSS}{TSS} = \frac{RSS}{TSS} + \frac{ESS}{TSS}$$
$$1 = R^2 + \frac{ESS}{TSS} = 1 - R^2 + \frac{ESS}{TSS}$$

The F statistics is; $\left(\frac{R^2}{1-R^2}\right)\left(\frac{(n-K-1)}{K}\right)$
Substituting the given values are: $\left(\frac{R^2}{1-R^2}\right)\left(\frac{(n-K-1)}{K}\right) = \frac{0.595338}{1-0.595338}\frac{15}{2} = 11.03$
The result is also provided in the Eviews printouts in Table 14.3.

At 5 % significance, the table value is 3.6823. The calculated F value is greater than the table value. We reject the Ho hypothesis.

The next step we need to consider is to look at the results. Are they satisfactory? The statistical tests appear to be satisfactory. A common practice for this purpose is to use the estimated regression equation and forecast.

For example, let us take the last two years:

Inserting the explanatory variables for these years into the regression equation provides these forecasts.

1998: RMF = −31939347 + 388.42 × 116114 − 21510.49 × (0.177) = 13157845.5
1999: RMF = −31939347 + 388.42 × 110286 − 21510.49 × (0.101) = 10895768.56

The actual value of the real import in 1998 is 15857697.5. Comparing this value with our regression estimate gives a 17 % error. Likewise in 1999, this error is much higher at 55 %. Both predicted years have got a great numbers of mistakes. The forecast of real imports with this regression equation is presented in Fig. 14.3. The

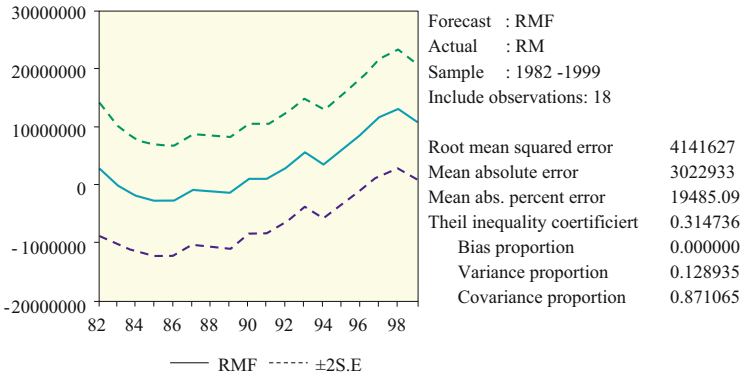


Fig. 14.3 Forecasting of real imports

errors are presented in the same figure. The errors appear to be very high, which gives an indication that the linear estimated model is perhaps not a good fit.

The accuracy of the forecasting can also be numerically tested. This test is called Chow test

14.2.4 Chow Test

This is a stability test. For this test, the first step is to divide the sample into two groups. The first group is the sample size up to the last two samples, which is in $n_1 = 16$ our example and the second sample size consists of the last two samples, that is $n_2 = 2$. The next step is to estimate the first sample size and obtain its Error Sum of Squares (ESS_1) and the second sample size with $n_1 + n_2$ also to be estimated. The second sample size ESS_P (pooled ESS) and the first sample size ESS_1 will be used for this test. The following step is to calculate F statistics with these ESSs.

$$F = \frac{(ESS_p - ESS_1)/n_2}{ESS_1/(n_1 - k - 1)} = \frac{(4536925 - 1500138)/2}{1500138/(16 - 2 - 1)} = 13.16$$

A further step is to compare the F statistics with the critical value of the F distribution. The degrees of freedom is n_2, $n_1 - k - 1 = 2$, 13. The table value is 3.6337 for 5 % significance level. The calculated value exceeds the critical value. Hence the model **fails** the prediction test. The model should be definitely improved.

The first problem is to look at the price variable. Theoretically, the expectation is that the price variable should not be positive and its error is very high in this example. We may do something about this. Importers may not be able to adjust price changes very quickly since the import contracts are for long term. We can try

Table 14.4 Lagged regression estimate

Dependent variable: RM Method: Least squares Sample (adjusted): 1983 1999 Included observations: 17 after adjusting endpoints				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	−36,931,460	9,611,138.	−3.842569	0.0018
RGDP	437.0564	99.94036	4.373172	0.0006
RPM(−1)	23012.07	10200.81	2.255906	0.0406
R-squared	0.620883	Mean dependent var	3,331,900.	
Adjusted R-squared	0.566723	S.D. dependent var	6,858,017.	
S.E. of regression	4,514,205.	Akaike info criterion	33.64214	
Sum squared resid	2.85E + 14	Schwarz criterion	33.78918	
Log likelihood	−282.9582	F-statistic	11.46395	
Durbin-Watson stat	0.674136	Prob(F-statistic)	0.001126	

the lagged value of the import price variable. The economical interpretation of this is that the demand might respond to price changes a year later.

The new model:

$$RM_t = b_0 + b_1 RGDP_t + b_2 RPM_{t-1} + \mathrm{e}_t$$

The Eviews output of the above estimated equation is presented in Table 14.4.

$RM_t = -36931460 + 437\ .05 RGDP_t + 23012\ .07\ RPM_{t\ -1} + e_t$
(99.94) (10200.81)
$R^2 = 0.62\ F_{2,15} = 11.46$

In this new regression, the R^2 has slightly improved, but there are no great changes in the rest of the estimated equation. Further details of the results are presented in Table 14.4 Eviews output.

The next obvious step is to try estimating the regression relationship in natural logs. There are two main reasons for taking the logarithmic form of the variables:

1. This can give a better functional form.
2. The coefficients of the logarithmic estimate are direct estimates of the elasticities, which can be interpreted easily.

Table 14.5 presents the data in natural logarithmic form. If we summarize the logarithmic estimated results of the regression in Table 14.6, the following estimated regression equation is obtained.

$LRM_t = 5.59 + 0.838\ LRGDP_t - 1.0972\ LRPM_{t\ -1} + e_t$
(0.7258) (0.06143)
$R^2 = 0.99\ F = 3428\ n = 18$

The elasticity are now easier to observe. The elasticity with respect to income is 0.84 and the elasticity with respect to price (lagged) is −1.097. The most important

Table 14.5 The data is transformed into the natural logarithmic form

T Year	LM	LGDP	LP
1982	7.21	10.91	6.43
1983	7.66	10.96	6.09
1984	8.33	11.02	5.66
1985	8.74	11.06	5.31
1986	9.13	11.13	4.83
1987	9.52	11.22	4.58
1988	10.06	11.24	4.02
1989	10.61	11.25	3.56
1990	11.09	11.33	3.14
1991	11.54	11.34	2.60
1992	12.15	11.40	2.05
1993	12.92	11.48	1.48
1994	13.63	11.42	0.77
1995	14.35	11.49	0.26
1996	15.19	11.56	−0.38
1997	16.04	11.63	−1.09
1998	16.58	11.66	−1.73
1999	17.00	11.61	−2.29

Table 14.6 Logarithmic results from the Eviews printout

Dependent variable: LOG(RM)
Method: Least squares
Sample (adjusted): 1983 1999
Included observations: 17 after adjusting endpoints

Variable	Coefficient	Std. error	t-Statistic	Prob.
C	5.593061	8.400430	0.665806	0.5164
LOG(RGDP)	0.838502	0.725878	1.155156	0.2674
LOG(RPM(−1))	−1.097181	0.061425	−17.86201	0.0000
R-squared	0.997962	Mean dependent var	12.03278	
Adjusted R-squared	0.997671	S.D. dependent var	3.015587	
S.E. of regression	0.145525	Akaike info criterion	−0.858159	
Sum squared resid	0.296484	Schwarz criterion	−0.711121	
Log likelihood	10.29435	F-statistic	3428.273	
Durbin-Watson stat	1.417966	Prob(F-statistic)	0.000000	

part of the last estimate is that apart from the lagged price coefficient, the other coefficients are **not** statistically significant. The R^2 is extremely high, which may indicate some other problems, such as **autocorrelation**.

Autocorrelation can occur when one observation is correlated with an earlier one. This problem does not matter in a cross-section data but it is a serious problem in time series data. Import equations are known to have autocorrelation because of the stickiness of imports. This means that if import at current time is higher than predicted, then they are likely to be higher in the future period.

14.2.5 Autocorrelation

If we need evidence for autocorrelation, we could examine the error term. The errors are the differences between the fitted values and the actual observations. This can be explained with the following equation:

$$e_t = Y_t - \hat{Y} = Y_t - b_0 - b_1 X_1 t - b_2 X_2 t$$

Before the calculation of the Durbin–Watson (DW) statistics, we can observe the time series diagram of the errors from the import demand equation. Figure 14.4 suggests a definite positive autocorrelation pattern. For example, negative errors tend to follow negative errors and positive errors tend to follow positive errors. It seems that there is a positive correlation between e_t and e_{t-1} . In fact what we need is a truly random series, which has a low or zero correlation but this is not the case here.

The numerical test of this randomness can be done by Durbin–Watson (DW) test statistics. More or less, all econometric software packages such as Eviews, Microfit, PCGive and TSP provide the standard DW numerical statistics. Our Eviews print out, Table 14.6, has the DW statistics for our example.

Although the Eviews results provide the calculated **Durbin–Watson (DW)**, we will now see how this calculation is done.

Firstly, the residuals need to be obtained. Table 14.7 presents these values. The first column is the actual import demand from the original data. The second column is the predicted values of the import demand. The third is the residual values which are obtained by subtracting the predicted values from the actual ones. The DW test statistics can be calculated with the following formula:

Fig. 14.4 Time series graph of the residuals

Table 14.7 Calculation of DW statistics

Year	Actual	Fitted	Residual	et-1	(et -et-1)	(et -et-1)^2	et^2
1982							
1983	7.66	7.72	−0.06				
1984	8.33	8.15	0.18	−0.06	0.2415	0.0583	0.0334
1985	8.74	8.66	0.07	0.18	−0.1090	0.0119	0.0055
1986	9.13	9.10	0.04	0.07	−0.0366	0.0013	0.0014
1987	9.52	9.70	−0.18	0.04	−0.2193	0.0481	0.0331
1988	10.06	10.00	0.07	−0.18	0.2476	0.0613	0.0043
1989	10.61	10.62	−0.01	0.07	−0.0735	0.0054	0.0001
1990	11.09	11.19	−0.10	−0.01	−0.0893	0.0080	0.0095
1991	11.54	11.65	−0.11	−0.10	−0.0171	0.0003	0.0131
1992	12.15	12.30	−0.14	−0.11	−0.0300	0.0009	0.0208
1993	12.92	12.96	−0.04	−0.14	0.0996	0.0099	0.0020
1994	13.63	13.55	0.09	−0.04	0.1312	0.0172	0.0075
1995	14.35	14.39	−0.04	0.09	−0.1244	0.0155	0.0014
1996	15.19	15.00	0.19	−0.04	0.2303	0.0530	0.0370
1997	16.04	15.77	0.27	0.19	0.0765	0.0059	0.0723
1998	16.58	16.57	0.01	0.27	−0.2615	0.0684	0.0001
1999	17.00	17.23	−0.23	0.01	−0.2346	0.0551	0.0517
T		Totals				**0.42040**	**0.2930**

$$DW = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

The calculation of the above formula is presented on the Table 14.7. Hence the result of DW is as follows:

$$DW = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = \frac{0.42040}{0.2930} = 1.43$$

The result is due to the rounding error and is therefore slightly different from the Eviews print out calculation. The calculated number will help us decide whether there is an autocorrelation problem in the estimated regression equation.

We will be testing the null hypothesis of no autocorrelation against the positive or negative autocorrelation as a one tail test. The test statistics always lie between 0 and 4. The obtained number needs to be compared with the critical values which are d_L and d_U. The rules about the decision are presented in Fig. 14.5 below. If the

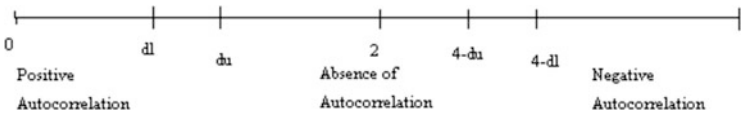


Fig. 14.5 The DW test statistics decision rule

values are lower than d_L, this suggest a positive autocorrelation. If the values are above the $4 - d_L$, this suggests a negative autocorrelation. If the obtained value is between d_U and $4 - d_U$ then the autocorrelation problem does not exist. For the other two regions, the test is inconclusive.

The next step is to look at the critical values table for the DW test. This is presented at the end of the text. At 5 % level of significance, the nearest sample size to 18 is 20 and two explanatory variables give the d_L as 1.100 and d_U as 1.537. The calculated DW, 1.43 is between d_L and d_U which suggests an inconclusive result. It may be possible to obtain a better regression estimate by taking the different auto correlated nature of the errors. If we look at our first estimated equation, the DW was 0.601465. This was clearly indicating a positive autocorrelation. The next model with the lagged price variable had DW = 0.674, which indicates a positive autocorrelation. The model with a logarithmic form turns the model into inconclusive DW result, which suggests that the last form is a better form. One of the most common ways of getting rid of the autocorrelation problem is to convert the data into logarithmic form. This is because of the fact that if the original relation is not linear, then the linear relation does not explain the relation well. Thus converting into a log form (nonlinear transformation) may solve the problem.

The other important point in Turkish import data is that Turkey had trade liberalization when (in particular) the country joined the custom union with EU. It has been argued for a long time that the country's import volume has had a dramatic shift since then. This effect can be examined by introducing a **dummy variable**.

14.2.6 Dummy Variables, Trends and Seasonal

A dummy variable takes a restricted range of values. It usually takes 0 and 1. For example, Turkey joined the EU Custom Union in 1996. If we create a new variable in the regression data called 'dummy' which takes the value 0 from 1982 to 1996 and which from 1996 until 1999 takes the value 1. The new regression equation then becomes:

$$LRM_t = b_1 + b_2(LRGDP)_t + b_3(RPM)_{t-1} + b_4 D_t + e_t$$

where D is the dummy variable. A dummy variable is entered into our sample and the regression re-estimated and the estimated equation is as follows:

$LRM_t = 0.9398 + 1.\ 2289\ LRGDP_t - 1.03567\ LRPM_{t\ -1} + 0.\ 228\ D_t + e_t$
(0.7287) (0.0689) (0.1333)
$R^2 = 0.99\ F = 2550\ n = 18$

The coefficient 0.228 gives the size of the shift in 1996 when Turkey joined the EU Custom Union. The dummy coefficient is also statistically not significant in our example (t test result).

These dummy variables can be applied to more than one structural break. When there are two breaks in the same term then the terms can be represented with numbers 0, 1, and 2. If there are differences between seasons (these may the case in quarterly or monthly data), we can apply these dummy variables for different seasons.

Another dummy variable which is also useful for time series data is the time trend. It takes the values 1, 2, 3, ...T. where T denotes the number of observations. For example, for our Turkish data, a general improvements in technical progress would make the import volume increase. (Conditions changed throughout time). Then it may be necessary to use time trend as another variable. After adding the time trend into the demand regression, we obtained the Eviews results presented on Table 14.8.

All coefficients appeared to change and again only the price coefficient is statistically significant. The autocorrelation still appeared to be inconclusive.

We will try to obtain a better regression estimate considering some further issues such as **Multicollinearity**.

14.2.7 Multicollinearity

Sometimes the explanatory variables are correlated and this situation creates the problem of not knowing which of the independent variables influence the dependent variable. The typical symptoms of multicollinearity can be observed if there is a high correlation between two or more independent variables, if high standard errors of the coefficients create low t-ratios and if there is a high value of R^2 with insignificant individual coefficients. This may be the case in our estimated import demand equation. In this case, low t-ratios do not mean that the corresponding coefficients are statistically insignificant: they mean that the multicollinearity problem is present. One way to solve this problem is to include some further data into the sample size, but this may not always solve the problem.

Another usual practice for multicollinearity is to take the first difference of the variables. In our regression equation, these are: DLM = log(RM) – log(RM(−1)) (real import differences), DLGDP = log(rgdp) – log(rgdp(−1)) (Real GDP difference) and DLP = log(RPM) – log(RPM(−1)) (Import price differences). All values are in the logarithmic form. The Eviews printout of the first difference-logarithmic equation is presented in Table 14.9.

Table 14.8 Import demand regression with dummy and time trend

Dependent variable: LM Method: Least squares Sample (adjusted): 1983 1999 Included observations: 17 after adjusting endpoints				
Variable	Coefficient	Std. error	t-Statistic	Prob.
C	8.423459	12.88302	0.653842	0.5255
LGDP	0.415370	1.279508	0.324632	0.7510
LP(−1)	−0.853531	0.243964	−3.498594	0.0044
DMY	0.334445	0.192299	1.739193	0.1076
T 0.120772		0.154966	0.779344	0.4509
R-squared	0.998385	Mean dependent var	12.03176	
Adjusted R-squared	0.997847	S.D. dependent var	3.016207	
S.E. of regression	0.139952	Akaike info criterion	−0.855101	
Sum squared resid	0.235040	Schwarz criterion	−0.610038	
Log likelihood	12.26836	F-statistic	1854.898	
Durbin-Watson stat	1.302470	Prob(F-statistic)	0.000000	

Table 14.9 Elimination attempt for multicollinearity

Dependent variable: DLM Method: Least squares Sample (adjusted): 1983 1999 Included observations: 17 after adjusting endpoints				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	−0.022696	0.121100	−0.187419	0.8540
DLGDP	1.689846	0.609707	2.771573	0.0150
DLP	−1.030612	0.208185	−4.950473	0.0002
R-squared	0.645390	Mean dependent var	0.576249	
Adjusted R-squared	0.594731	S.D. dependent var	0.156784	
S.E. of regression	0.099810	Akaike info criterion	−1.612312	
Sum squared resid	0.139468	Schwarz criterion	−1.465274	
Log likelihood	16.70465	F-statistic	12.73999	
Durbin-Watson stat	1.477229	Prob(F-statistic)	0.000705	

The last regression results in Table 14.9 appear to provide better results. The R^2 is not very high and the coefficients have similar characteristics as the theoretical requirements. The import price (DLP) has a negative relationship with import volume, that is when import prices go up, the import volume goes down. When GDP (DLGDP) goes up, the import goes up, too. Both of these coefficients are statistically significant and the DW is inconclusive.

14.3 Some Other Problems, Methods and Tests for Finding the Right Model

Finding the right model is not an easy task. Further tests and methods are beyond the scope of this book. If you would like to consult further, you could see, for example Koutsoyiannis (1977), Maddala (1992) and Thomas (1993) which are included in the reference list. Perhaps they have more suitable analysis for an econometrics course. However, we will now briefly mention some of these problems, methods and tests.

When we consider a regression equation, can we say that we find the right explanatory variables? Can you add another variable into the estimated regression? Good modelling is based on theoretical formations. A good model should be consistent with the theory, statistically significant and has to be simple. There are two approaches to construct a model. The first one is from general to specific and the second one is from specific to general. The general to specific approach is about starting with a comprehensive model which includes all possible parameters and gradually eliminates those who are not relevant. The specific to general is about beginning with a simple model and adding gradually more explanatory variables to the model. Both approaches have advantages and disadvantages. Omitting relevant variables can cause serious problems. There are also some other problems such as **simultaneity and measurement errors.**

Simultaneity
In our import demand model, quite often the price and the quantity of good consumed are simultaneously determined. For this problem, we can use the simultaneous equations model. (See other econometrics text for these methods).

Measurement Error
Measurement error has to be random for a good regression estimate. However, if the measurement error is systematic, then the estimate is not good. For example, if the transport cost excluded the import price, then systematic measurement error can exist.

In a regression estimation, large samples are always better than small samples. Make sure that the data is correctly typed and includes no errors before the computer estimation. It may be better to start with a simple model to estimate.

14.4 A Review of This Chapter

In this chapter, multiple regression was examined. Firstly, a standard multiple regression steps were explained and then more detailed analysis is carried out. The applied part of the chapter used Turkish macroeconomic data. We also examined further problems in econometrics such as autocorrelation, multicollinearity, but we mainly left the econometrics for another course.

14.4.1 Review Problems For Multiple Regression

14.1. The following table contains data on the price of gold (PG; weighted average price of gold (TRL/Gr) in Istanbul Gold Exchange) and the Istanbul Consumer Price Index (CPI; 1994 = 100).

Year	CPI	PG	PG*PG	CPI * CPI	PG * CPI
1995	188.06	65,721,715	4.31934E + 15	35,365.24719225	12,359,395,697
1996	339.08	251,985,461	6.34967E + 16	114,976.87398976	85,443,834,881
1997	630.82	405,989,156	1.64827E + 17	397,931.34912400	2.56105E + 11
1998	1164.88	616,780,937	3.80419E + 17	1,356,955.43238649	7.18478E + 11
1999	1921.97	935,772,713	8.75671E + 17	3,693,985.20986049	1.79853E + 12
2000	2971.06	1,367,034,954	1.86878E + 18	8,827,181.47988329	4.06154E + 12
2001	3501.10	125,817,459	1.583E + 16	12,257,701.21000000	4.405E + 11
Totals	10716.97	3,769,102,395.00	3,373,347,104,256,490,000.00	26,684,096.80	7,372,956,718,207.12

(a) An investment advisor claims that the price of gold rises faster than the CPI. What value would you expect the elasticity of gold prices with respect to the CPI to take if he/she is correct?

ch14data2.xls

(b) Estimate the following equation using the data in the table

$$GP_t = \beta_1 + \beta_2 CPI_t + e_t$$

(c) Evaluate the elasticity of PG with respect to CPI at the means.
(δ) Test the hypothesis H_0: $\beta_2 = 1$ against the alternative H_1:$\beta_2 > 1$

14.2 The following data is obtained from TUIK; Food = food, beverages and tobacco expenditure. Istanbul Consumer Price Index (CPI; 1987 = 100). Wage = Wages and salaries.

ch14data3.xls

Year	Food	CPI	Wage
1987	209,674,000	100	37,890,000
1988	213,245,000	174	38,594,000
1989	210,271,000	284	39,056,000
1990	225,850,000	454	40,189,000
1991	233,052,000	755	41,172,000
1992	235,407,000	1.284	42,586,000
1993	243,102,000	2.136	43,362,000
1994	241,514,000	4.406	43,708,000
1995	261,224,000	8.512	44,816,000
1996	267,524,000	15.269	44,688,000
1997	269,794,000	28.291	44,728,000
1998	269,472,000	51.948	47,385,000
1999	271,331,000	84.995	48,685,000

a) Estimate the following equation using the data in the table.

$$Food_t = \beta_1 + \beta_2 CPI_t + \beta_3 Wage_t + e_t$$

b) Calculate the estimated values by inputting an appropriate formula and calculate the residuals for the regression equation.
c) Obtain all the relevant tests. What can you do to improve your regression equation?

14.3 The following **Eviews** printout is obtained where IFM = infant mortality, C= Constant, GDPPC = Per capita gross domestic product, GR = Economic growth ratio, HEX = Per capita health expenditures.

Dependent variable: IFM
Method: Least squares
Sample: 1 29
Included observations: 29

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	19.23901	5.333742	3.607038	0.0013
GDPPC	−0.000428	0.000176	−2.435541	0.0223
GR	0.544935	0.626434	0.869901	0.3926
HEX	−0.567031	0.731047	−0.775643	0.4452
R-squared	0.406218	Mean dependent var	7.434483	
Adjusted R-squared	0.334964	S.D. dependent var	6.478276	
S.E. of regression	5.283018	Akaike info criterion	6.294314	
Sum squared resid	697.7569	Schwarz criterion	6.482906	
Log likelihood	−87.26755	F-statistic	5.700991	
Durbin-Watson stat	1.774595	Prob(F-statistic)	0.004076	

(a) Explain why each variable is included in the regression. Does each coefficient have the expected sign?
(b) Comment upon:

i. the sizes and signs of the coefficients
ii. the significance of the coefficients
iii. the overall significance of the regression.

(c) How would you simplify the model?
(d) Test for the autocorrelation

14.5 Computing Practical for Multiple Regression Using the 'Eviews'

C.14.1. Open a new Excel worksheet, use the data provided below and copy the table on your new sheet. Call it MACRO.XLS.

ch14data4.xls

Year	GDPFC	LTIR	Invest
1987	74721.8	52	18491.1
1988	76306.1	83.9	18298.6
1989	76498.1	58.8	18700.8
1990	83578.2	59.4	21669.9
1991	84352.7	72.7	21934.9
1992	89400.7	74.2	22882.2
1993	96590.5	74.7	28573.7
1994	91320.7	95.6	24026.8
1995	97887.8	92.3	26822.9
1996	104745.1	93.8	30597.8
1997	112631.2	96.6	35137.2
1998	116113.5	94.8	33768.2
1999	110,646	46.7	28369.6
2000	88804.7	45.6	15157.8

In Excel, highlight the cells containing the data for GDPFC (Gross Domestic Product with Factor Costs), LTIR (Long Term Interest Rates) and INVEST (Investment) with the column headings. Select **Edit/Copy** to copy the highlighted data to the clipboard. Start **Eviews** and create a workfile. To create a workfile to hold your data, select **File/New/Workfile**, which opens a dialog box where you will provide information about your data. You should set the work file frequency to annually, and specify the start date 1987 and the end date 2000. Click O.K. The next step is to select **Quick/Empty Group (Edit Series)** in Eviews. Place the cursor in the upper-

left cell, just to the right of the second obs label. Then select **Edit/Paste** from the main menu. The group spreadsheet will now contain the data from the clipboard.

Estimating a Regression Model

(a) We now estimate a regression model for investment using data over the period from 1987 to 2000. To estimate the model, we will create an equation object. Select **Quick** from the main menu and choose **Estimate Equation** to open the estimation dialogue. Enter the following in the Equation Specification box:
Invest c GDPFC LTIR
Click OK to estimate the equation using the least squares and display the regression results. Print your regression results. After completing the statistical tests (t-tests and F tests), comment about your investment regression. In your equation box, click on **View/Residual Graph**. What type of time-series graph of errors from the Investment do you obtain? Print your graph and comment on it.

(b) Now delete the equation box by clicking on the right top corner of the x signed icon. Select **Quick** again from the main menu and choose **Estimate Equation** to open the estimation dialogue. Enter the following in the Equation Specification box:
log(INVEST) c log(GDPFC) log(LTIR)
where log(INVEST) is the logarithm of the investment, log(GDPFC) is the log of GDP, log(LTIR) is the logarithm of the annual interest rates.
Click OK to estimate the equation using the least squares and display the regression results. Print your regression results.

(c) Use both of your estimated regression equations and forecast the investment levels for the year 1999 and 2000. Compare your estimated values with the actual investments for these years.

(d) Again by using your estimated regression equations, forecast the investment levels for the future years 2001, 2002 and 2003. (Assume that GDPFC increases by 5 % and LTIR stays the same as the year 2000.

(e) Now write a report about both of your estimated regression equations and forecasts. Make sure that you include the printouts at the end of your report.

Chapter 15
The Analysis of Time Series

15.1 Introduction

In this chapter, we will examine time series. The main aim of time series analysis is to try to predict the future by projecting the patterns identified in the past. It is assumed that these patterns might be the same in the near future. However, they may not be the same in the distant future. These patterns consist of various elements such as trend, seasonal factors, random factors and cyclical factors. We will firstly explain these factors and examine them within an example. We can predict the systematic components (trends, cycle and seasonal) of any time series data but not the random component.

15.2 Decomposition of a Time Series

Private consumption expenditure (excluding durable consumption) data will be used to illustrate our analysis of time series in this chapter since it is one of the most commonly used data in monthly or quarterly time series analysis. There is a large selection of data which can be used for this purpose, such as, monthly unemployment, money supply and the monthly or quarterly sales data. Time series analysis is about a set of techniques decomposing the time series data (hence it is not applicable to cross section data). If time series data has some random elements as we mentioned earlier, we cannot do very much about it, but we can do something about the systematic factors. The decomposition of time series is about considering the systematic component, such as trends, cycles and seasonal.

The online version of this chapter (doi:10.1007/978-3-319-26497-4_15) contains supplementary material, which is available to authorized users.

D. Özdemir, *Applied Statistics for Economics and Business*,
DOI 10.1007/978-3-319-26497-4_15

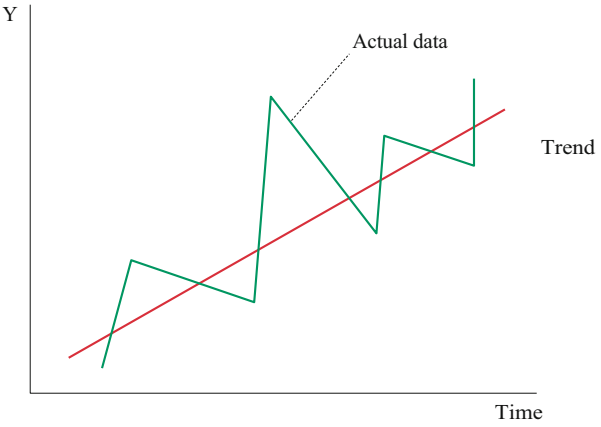


Fig. 15.1 Trend and the actual data

Trend (T)
A trend is a broad general direction of the movement in time series data. The trend analysis is important because a number of economic variables show trends over time. Figure 15.1 shows the difference between the trend and the actual data.

Cyclical Factors (C)
These factors are long term regular fluctuations. For example, in most economies, a rapid growth is quite often followed by a period of relative stagnation. These types of cycles can have different lengths. Trade cycles and growth cycles are the examples to name but a few. Since the time length of the cycles are not easy to predict, the analysis very difficult and hence quite often ignored.

Seasonal Factors (S)
The seasonal factors are short term cycles. For example, ice cream sales are higher du- ring the summer. Holiday sales also differ between summer and winter. These type of cycles are regular cycles, which make them easy to isolate.

Random Factors (R)
These cyclical factors are random, which makes them unpredictable. We will call these as the residual factors (ie the rest of the factors are apart from the other three factors). These four factors may be combined together in two ways; an additive model and a multiplicative model.

The additive model is about adding all these factors and obtaining the actual data (A). For example the additive model of M0 (currency issued) is

$$A = T + S + C + R$$

The multiplicative model is obtained by multiplying the four factors.

$$A = T \times C \times S \times R$$

The difficulty of identifying the cyclical factors and the unpredictability of the random element leaves us with the following models. For the additive model A = T + S and for the multiplicative model $A = T \times S$.

The analysis of the data follows three steps:

(a) Isolating the trend from the original data by using the method of moving averages.
(b) Comparing the actual figures to the trend and observing the times that the actual data is above the trend. This way the seasonal factors can be extracted from the data.
(c) Obtaining the seasonally adjusted data.

15.3 Isolating the Trend

Again the best way to explain this process is to use data and apply the method. For these exercises, quarterly data will be used. Constant price (1987) private consumption expenditures are presented in Table 15.1.

They are in Million Turkish Liras.

The same data is plotted in Fig. 15.2, which clearly shows fluctuations. There are a number of methods to smooth such data. The purpose of smoothing is to clear the short- term fluctuations in the data. The method is used called the **moving average trends**. Moving average smoothing is about looking at relatively small sections, finding an average, and then moving on to another section. Each size of the section depends on the size of the data used. If the data is monthly then the subset is 12: if the data is quarterly, use the subset of 4; if the data is daily, then use subset 5 or 7. For our quarterly data, we will be using the 4 point moving average trend for consumption expenditures.

ch15data1.xls

The centered 4 point moving average with the additive method is presented in Table 15.2. The first column shows the years considered in the data. The second column shows the corresponding quarters for each observed consumption expenditure. The third column indicates the observation number of the series. The fourth column denotes the actual consumption expenditures in billion Turkish Liras. Centered 4 point moving average calculation starts in column 5. For the moving average calculations, firstly we take the sum of the first four quarterly consumption expenditure values: these are the 1997Q1-Q4. The total value of the first four actual observations are entered in the second cell of the fifth column. That is 10793 + 11492.9 + 14644.3 + 14088.3 = 51018.5. The second sum of four quarters is calculated after dropping the first observation. That is 11492.9 + 14644.3 + 14088.3 + 11429.2 = 51654 (taking the second observation until the fifth actual observation). The rest of the fifth column is calculated in a similar way.

Table 15.1 Quarterly data on consumption (private) expenditures (excluding durable goods)

	Q1	Q2	Q3	Q4
1987	10793	11492.9	14644.3	14088.3
1988	11429.2	11888	15028.7	13292.1
1989	10592.8	11559.2	15246.2	13706.9
1990	12098.3	13335.9	16940.1	15428.9
1991	12373.8	13325.3	17635	15579.2
1992	13026.6	13726.9	18082	16027.4
1993	13698	15430.2	19492.3	17379.2
1994	14498.3	13855.9	18040.9	16098.8
1995	14008.8	15527.9	19678.1	16796.6
1996	15551.8	16998.1	20902.7	18160.9
1997	16832	18662	22766.8	19359.6
1998	18275.7	18555	23055.2	18227.4
1999	17032.1	18389.7	22384.1	17884.6
2000	17815.8	19677.2		

Source: Turkish Central Bank data

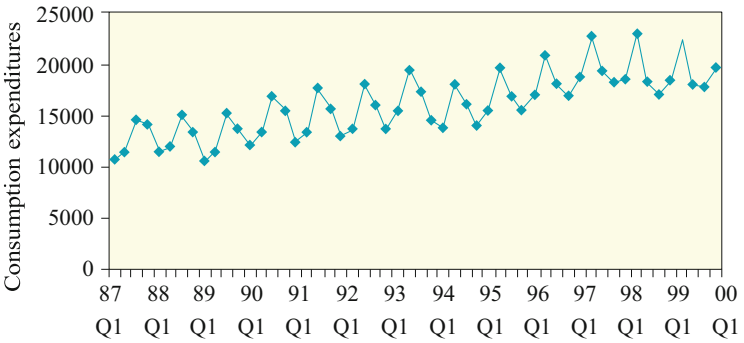


Fig. 15.2 Quarterly data on private consumption Expenditures. From the first quarter 1987 to the second quarter 2000 (million TL). *Source*: Turkish Central Bank data source

The sixth column, which is the sum of 8 quarters, is calculated as follows: The first two sums of the 4 quarter column are written in the sixth column as the first sum of the 8th quarter sum that is $51018.5 + 51654 = 102673.2$. The second sum of the 8th quarter is calculated after dropping the first sum of the 4th quarter, that is $51654 + 52049.8 = 103704.5$. The rest of the sixth column is calculated with a similar way. The trend column is the column 7. The trends are obtained by dividing the sum of the 8th quarter by 8, which give the average of the 8 quarters. This is the column which excludes the systematic components of the time series on consumption expenditure data. The last column Table 15.2 shows the seasonal variations, which are obtained after subtracting the trend column (Average (T)) from the actual observations (Actual (A) consumption expenditures in billion TL). This is the seasonal adjustment factor.

Table 15.2 Centered four point moving average trend for consumption expenditures and its seasonal variations

		Observ.	Actual (A)	Sum of	Sum of	Average(T)	
Date	Quarts.	Number	Exp.	4 qtrs.	8 qtrs.	8qtrs/8	S = A - T
1987	Q1	1	10793				
	Q2	2	11492.9	51018.5			
	Q3	3	14644.3	51654.7	102673.2	12834.15	1810.15
	Q4	4	14088.3	52049.8	103704.5	12,963,063	11,252,375
1988	Q1	5	11429.2	52434.2	104484	13060.5	-1631.3
	Q2	6	11888	51638	104072.2	13,009,025	-1,121,025
	Q3	7	15028.7	50801.6	102439.6	12804.95	2223.75
	Q4	8	13292.1	50472.8	101274.4	12659.3	632.8
1989	Q1	9	10592.8	50690.3	101163.1	12,645,388	-20,525,875
	Q2	10	11559.2	51105.1	101795.4	12,724,425	-1,165,225
	Q3	11	15246.2	52610.6	103715.7	12,964,463	22,817,375
	Q4	12	13706.9	54387.3	106997.9	13,374,738	3,321,625
1990	Q1	13	12098.3	56081.2	110468.5	13,808,563	-17,102,625
	Q2	14	13335.9	57803.2	113884.4	14235.55	-899.65
	Q3	15	16940.1	58078.7	115881.9	14,485,238	24,548,625
	Q4	16	15428.9	58068.1	116146.8	14518.35	910.55
1991	Q1	17	12373.8	58763	116831.1	14,603,888	-22,300,875
	Q2	18	13325.3	58913.3	117676.3	14,709,538	-13,842,375
	Q3	19	17635	59566.1	118479.4	14,809,925	2,825,075
	Q4	20	15579.2	59967.7	119533.8	14,941,725	637,475
1992	Q1	21	13026.6	60414.7	120382.4	15047.8	-2021.2
	Q2	22	13726.9	60862.9	121277.6	15159.7	-1432.8
	Q3	23	18082	61534.3	122397.2	15299.65	2782.35
	Q4	24	16027.4	63237.6	124771.9	15,596,488	4,309,125
1993	Q1	25	13698	64647.9	127885.5	15,985,688	-22,876,875
	Q2	26	15430.2	65999.7	130647.6	16330.95	-900.75
	Q3	27	19492.3	66800	132799.7	16,599,963	28,923,375
	Q4	28	17379.2	65225.7	132025.7	16,503,213	8,759,875
1994	Q1	29	14498.3	63774.3	129000	16125	-1626.7
	Q2	30	13855.9	62493.9	126268.2	15,783,525	-1,927,625
	Q3	31	18040.9	62004.4	124498.3	15,562,288	24,786,125
	Q4	32	16098.8	63676.4	125680.8	15710.1	388.7
1995	Q1	33	14008.8	65313.6	128990	16123.75	-2114.95
	Q2	34	15527.9	66011.4	131325	16,415,625	-887,725
	Q3	35	19678.1	67554.4	133565.8	16,695,725	2,982,375
	Q4	36	16796.6	69024.6	136579	17,072,375	-275,775
1996	Q1	37	15551.8	70249.2	139273.8	17,409,225	-1,857,425
	Q2	38	16998.1	71613.5	141862.7	17,732,838	-7,347,375
	Q3	39	20902.7	72893.7	144507.2	18063.4	2839.3
	Q4	40	18160.9	74557.6	147451.3	18,431,413	-2,705,125
1997	Q1	41	16832	76421.7	150979.3	18,872,413	-20,404,125

(continued)

Table 15.2 (continued)

Date	Quarts.	Observ. Number	Actual (A) Exp.	Sum of 4 qtrs.	Sum of 8 qtrs.	Average(T) 8qtrs/8	S = A - T
	Q2	42	18662	77620.4	154042.1	19,255,263	-5,932,625
	Q3	43	22766.8	79064.1	156684.5	19,585,563	31,812,375
	Q4	44	19359.6	78957.1	158021.2	19752.65	-393.05
1998	Q1	45	18275.7	79245.5	158202.6	19,775,325	-1,499,625
	Q2	46	18555	78113.3	157358.8	19669.85	-1114.85
	Q3	47	23055.2	76869.7	154983	19,372,875	3,682,325
	Q4	48	18227.4	76704.4	153574.1	19,196,763	-9,693,625
1999	Q1	49	17032.1	76033.3	152737.7	19,092,213	-20,601,125
	Q2	50	18389.7	75690.5	151723.8	18,965,475	-575,775
	Q3	51	22384.1				
	Q4	52	17884.6				

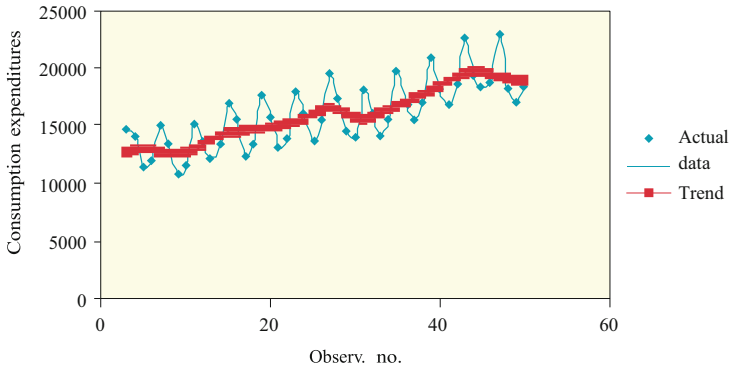


Fig. 15.3 Actual consumption expenditures and the trend which is obtained with moving average

Figure 15.3 plots the columns 3, 4 and 7 in Table 15.2. The centered 4 point moving average trend creates a smooth curve.

An alternative method to find the trend is regression. In Chap. 13, we showed how a straight line could be fitted to data. Similarly, we can obtain the regression equation and find a similar trend line. Both seasonal adjustment and the regression create a smooth line, but they will not be the same.

ch15data2.xls

The regression equation could be in the following form:

$$C = \beta 1 + \beta 2X$$

where X is the number of observations. The equation above is estimated and the following results are obtained (The calculations are not shown).

$$C = 11758 + 157.7918X$$

Forecasting with a regression equation has been examined before in the regression analysis chapters. Here again we are going to do forecasting with the consumption expenditure regression equation. The Consumption expenditure trend values which are estimated with the estimated regression equation are presented in the last column of Table 15.3. Clearly, there are some differences between these two trend calculations but not in a great extent. The regression method has the advantage of not losing observations at the beginning and end of the sample period. One of the main reasons of using regression and the moving average trend is that both can make future forecasts. Let us now examine these two forecast methods for some future periods.

For example in the year 2003 Q1 corresponds to the time period 65 (X = 65) Forecasting with the regression equation gives:

$$C = 11758 + 157.7918(65) = 22014.15$$

As Fig. 15.4 shows, there is a little difference between the regression trend and the moving average trend. It is clear that seasonal adjustment can help in interpreting figures but it may not be very good to use for further statistical analysis. Because the seasonal adjustment can introduce a cyclical component to the data series even though it may not initially have any cyclical factors. So for a regression analysis, it is often better not to do the seasonal adjustment when the F test for the joint significance of the seasonal coefficients shows that the seasonal effects are statistically not significant. If the seasonal effects are statistically significant, then it may be better to use seasonal dummy variables. The other important point about the seasonal adjustment is that the different methods can create different results.

15.4 A Review of This Chapter

In this chapter, we have examined the analysis of time series. We have focused mainly on the quarterly data which was obtained from the data source of the Central Bank. The systematic components of the time series such as the trend, cycle and seasonal cycles were examined.

Table 15.3 Centered four point moving average trend and the regression trend

		Observ.	Actual (A)	Sum of	Sum of	Average(T)		C = a + bX
Date	Quarts.	X	Exp	4 qtrs.	8 qtrs.	8qtrs/8	S = A - T	Est. Reg.
1987	Q1	1	10793					1714805.5
	Q2	2	11492.9	51018.5				1825244
	Q3	3	14644.3	51654.7	102673.2	12834.15	1810.15	2322509
	Q4	4	14088.3	52049.8	103704.5	12963.063	1125.2375	2234776.8
1988	Q1	5	11429.2	52434.2	104484	13060.5	-1631.3	1815192.6
	Q2	6	11888	51638	104072.2	13009.025	-1121.025	1887587.5
	Q3	7	15028.7	50801.6	102439.6	12804.95	2223.75	2383164.2
	Q4	8	13292.1	50472.8	101274.4	12659.3	632.8	2109143
1989	Q1	9	10592.8	50690.3	101163.1	12645.388	-2052.5875	1683215.6
	Q2	10	11559.2	51105.1	101795.4	12724.425	-1165.225	1835705.6
	Q3	11	15246.2	52610.6	103715.7	12964.463	2281.7375	2417483.9
	Q4	12	13706.9	54387.3	106997.9	13374.738	332.1625	2174595
1990	Q1	13	12098.3	56081.2	110468.5	13808.563	-1710.2625	1920771.1
	Q2	14	13335.9	57803.2	113884.4	14235.55	-899.65	2116054.3
	Q3	15	16940.1	58078.7	115881.9	14485.238	2454.8625	2684767.5
	Q4	16	15428.9	58068.1	116146.8	14518.35	910.55	2446312.5
1991	Q1	17	12373.8	58763	116831.1	14603.888	-2230.0875	1964242.8
	Q2	18	13325.3	58913.3	117676.3	14709.538	-1384.2375	2114381.7
	Q3	19	17635	59566.1	118479.4	14809.925	2825.075	2794417
	Q4	20	15579.2	59967.7	119533.8	14941.725	637.475	2470028.6
1992	Q1	21	13026.6	60414.7	120382.4	15047.8	-2021.2	2067249.3
	Q2	22	13726.9	60862.9	121277.6	15159.7	-1432.8	2177750.8
	Q3	23	18082	61534.3	122397.2	15299.65	2782.35	2864949.9
	Q4	24	16027.4	63237.6	124771.9	15596.488	430.9125	2540750.9
1993	Q1	25	13698	64647.9	127885.5	15985.688	-2287.6875	2173190.7

	Q2	26	15430.2	65999.7	130647.6	16330.95	-900.75	2446517.6
	Q3	27	19492.3	66800	132799.7	16599.963	2892.3375	3087483.7
	Q4	28	17379.2	65225.7	132025.7	16503.213	875.9875	2754053.8
1994	Q1	29	14498.3	63774.3	129000	16125	-1626.7	2299471.4
	Q2	30	13855.9	62493.9	126268.2	15783.525	-1927.625	2198106
	Q3	31	18040.9	62004.4	124498.3	15562.288	2478.6125	2858464.7
	Q4	32	16098.8	63676.4	125680.8	15710.1	388.7	2552017.2
1995	Q1	33	14008.8	65313.6	128990	16123.75	-2114.95	2222232.4
	Q2	34	15527.9	66011.4	131325	16415.625	-887.725	2461933.9
	Q3	35	19678.1	67554.4	133565.8	16695.725	2982.375	3116801.4
	Q4	36	16796.6	69024.6	136579	17072.375	-275.775	2662124.3
1996	Q1	37	15551.8	70249.2	139273.8	17409.225	-1857.425	2465705.1
	Q2	38	16998.1	71613.5	141862.7	17732.838	-734.7375	2693919.4
	Q3	39	20902.7	72893.7	144507.2	18063.4	2839.3	3310033.2
	Q4	40	18160.9	74557.6	147451.3	18431.413	-270.5125	2877399.7
1997	Q1	41	16832	76421.7	150979.3	18872.413	-2040.4125	2667710.2
	Q2	42	18662	77620.4	154042.1	19255.263	-593.2625	2956469.2
	Q3	43	22766.8	79064.1	156684.5	19585.563	3181.2375	3604172.9
	Q4	44	19359.6	78957.1	158021.2	19752.65	-393.05	3066544.7
1998	Q1	45	18275.7	79245.5	158202.6	19775.325	-1499.625	2895514.2
	Q2	46	18555	78113.3	157358.8	19669.85	-1114.85	2939585.4
	Q3	47	23055.2	76869.7	154983	19372.875	3682.325	3649680.1
	Q4	48	18227.4	76704.4	153574.1	19196.763	-969.3625	2887892.8
1999	Q1	49	17032.1	76033.3	152737.7	19092.213	-2060.1125	2699284.3
	Q2	50	18389.7	75690.5	151723.8	18965.475	-575.775	2913502.5
	Q3	51	22384.1					3543786
	Q4	52	17884.6					2833801.8

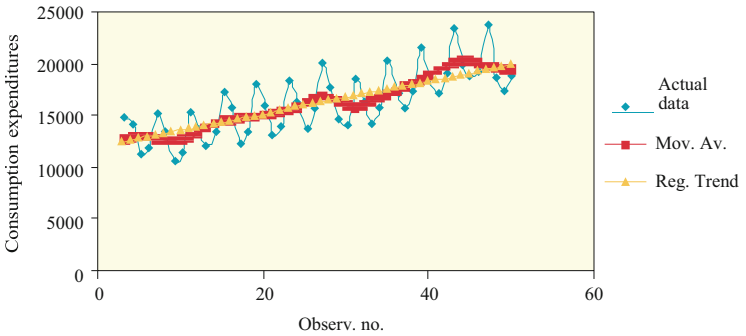


Fig. 15.4 Comparing the two trends

15.4.1 Review Problems for Time Series

15.1 The following table contains quarterly data on M0 (Currency Issued).

Date	Quarters	Observation Number	M0 Million TL
1997	Q1	3	459439.1
	Q2	4	567310.6
	Q3	5	702320
	Q4	6	758878
1998	Q1	7	897888
	Q2	8	1005988.3
	Q3	9	1195291.2
	Q4	10	1328542.4
1999	Q1	11	1926965.9
	Q2	12	1559694.1
	Q3	13	1870281.2
	Q4	14	2390748.4
2000	Q1	15	2371428.7
	Q2	16	2754486.7
	Q3	17	3120654.8
	Q4	18	3013451.7
2001	Q1	19	2926101.3

Source: Turkish Central Bank data

(a) Do you think that M0 varies seasonally? Why?
(b) Graph the series and comment upon any apparent seasonal pattern
(c) Construct a centered 4 point moving average trend for M0
(d) Using this trend construct your own seasonally adjusted series for M0
(e) Regressing M0 on X (the observation number) gives

$$M0 = 87309 + 178857.6X$$

where X is the observation numbers. Use this information to construct a different trend for M0

(f) Compare your regression trend to the centered moving average trend. Comment.

ch15data3.xls

Answers to Selected Review Problems and Computing Practicals

Chapter 1: Selected Answers to the Review Problems

1.3

As a statistical term, the population refers to all measurements or observations which are of interest to us. A sample is a smaller subset of population. It is essential to define the population because the ultimate objective is not to make statements about the sample but rather to draw conclusions about the population from which the sample has been drawn.

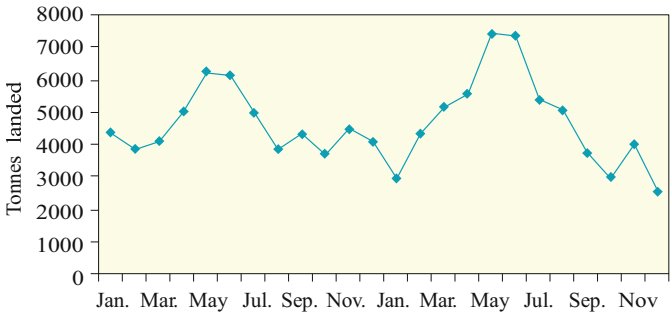
1.4

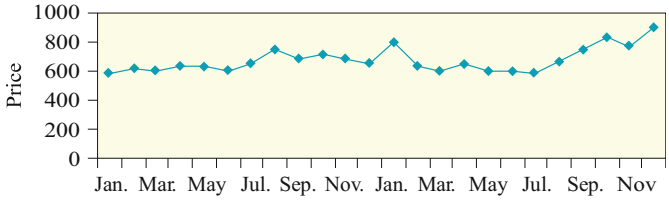
Postal surveys, because as far as the response rate is concerned they have the lowest up- take compared to other methods. However, a postal survey is the cheapest.

Chapter 2. Selected Answers to the Review Problems

A.2.1

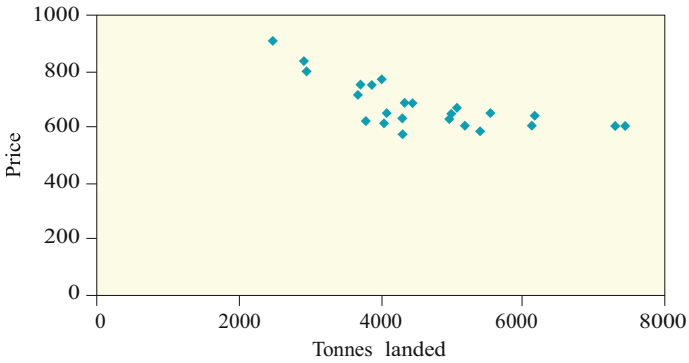
(a)





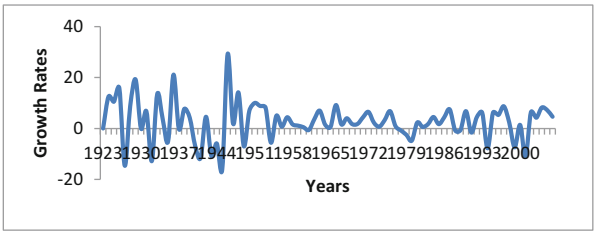
(b) i. (7313-6144)/6144 = 19.0 % increase.
ii. (7313-4082)/4082 = 79 % increase
iii. 7313-7449)/7449 = -1.8 % (decrease)

(c) A scatter diagram with quantity landed on the x axis and price on the y axis. It is a typical price/quantity relationship



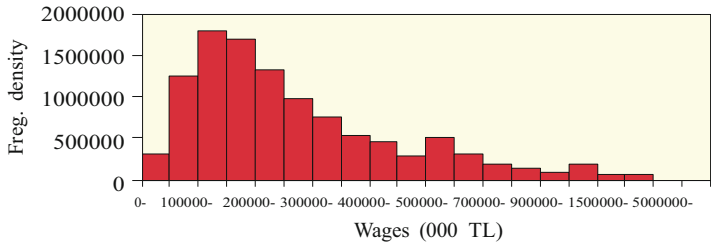
2.2

Turkish GNP growth rates at producer's prices 1923-1995: The effect of population growth can be observed by looking at the text version of the diagram, which has a bit more moderate growth.

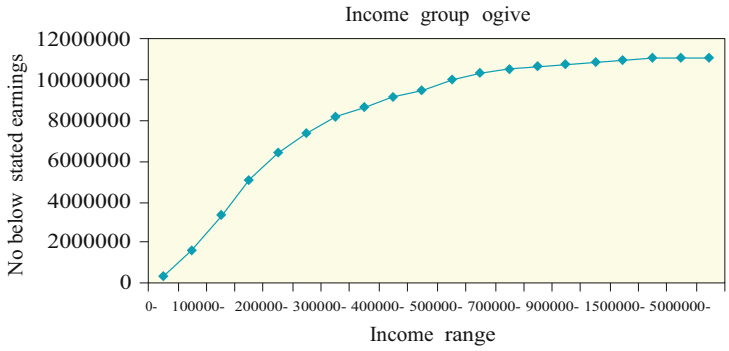


2.4

Histogram:



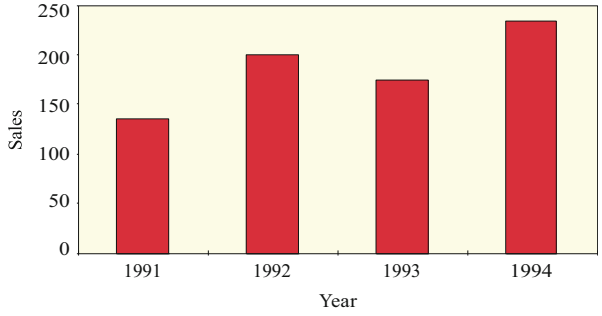
The ogive:



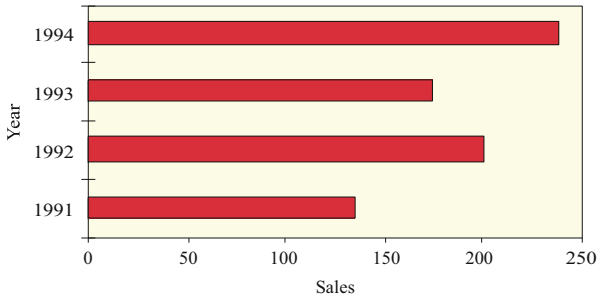
Chapter 2: Selected Computing Practical Solutions

C.2.2

(a) Column chart (vertical bars):

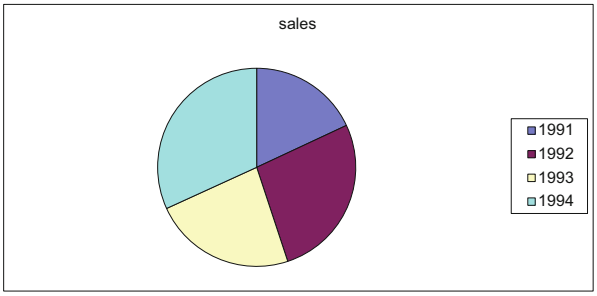


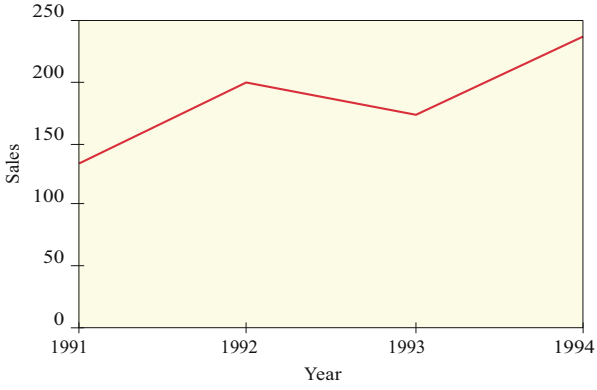
(b) Column chart (horizontal bars):



(c) Pie chart:

A line chart:





15.4.2 Chapter 3 Selected Answers to the Review Problems

3.1
For a) and b) see the basic descriptions from the chapter. c) Mean = 2.84, Median = 3, Mode = 4

3.2
Mean = 5.756, median = 5.1854, Mode = 1.92

3.4
Mean (pre-tax) = 14.815 and (post-tax) = 72.371
Modal Class = 5000–7500, Median = 11.976

Chapter 4: Selected Answers to the Review Problems

4.1
The standard deviation of car ages is 3.9071
The quartiles: Q1 = 2.4701, Q2 = 5.1851, Q3 = 8.4322
The quartile deviation is 2.9811:

4.2
The standard deviation: (pre-tax) = 12.345 and (post-tax) = 8.755
Coefficient of variation = (pre-tax)0.833288 and (post-tax) 0.70773
The coefficient of variation (CV) is a relative measure of dispersion (relative to the mean). Both the CV and the standard deviation decrease after tax. This indicates that the tax is stabilizing.

Chapter 5: Selected Answers to the Review Problems

5.1
a) and b)

a)		b)	
Year	Profits	In. 97	In. 00
1997	61750	100	82.99
1998	69180	112.03	92.98
1999	73892	119.66	99.31
2000	74405	120.49	100.00
2001	78063	126.42	104.92
2002	77959	126.25	104.78

c) Profit increase: 26.25 % between 1997 and 2002
4.78 % between 2000 and 2002

5.2

(a) Laspeyre: 100, 117.72 and 122.72
(b) Paasche: 100, 111.76, 117.63
(c) The Laspeyre is greater than The Paasche because one is a current weighted index while the other is a base weighted index.

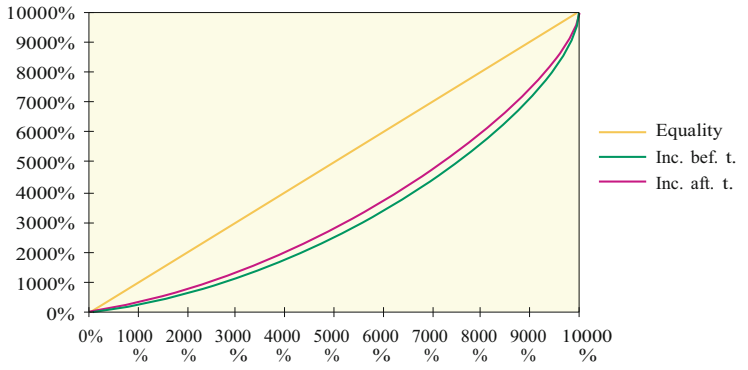
5.3

The Laspeyres index gives a 26.06 % increase in inflation and the Paasche index gives a 25.71 % increase.

Chapter 6: Selected Answers to the Review Problems

6.1

Income tax has reduced income inequality as the Lorenz curve indicates below. (But be careful: other taxes may not reduce inequality)



6.4

The post-tax Gini coefficient is 0.30, which is lower than the pre-tax Gini coefficient. This confirms the similar result about the effect of taxes on income distribution as in question 6.1.

Chapter 6: Selected Computing Practical Solutions

C.6.2

(a) The calculated Gini coefficients are as follows: For original income, the measure is 0.45, for gross income it is 0.35, for disposable income it is 0.33 and for post tax income it is 0.34.
(b) A smooth Lorenz curve increases the Gini coefficient.

15.4.3 Chapter 7: Selected Answers to the Review Problems

7.1

(a) 4/52 = 0.077
(b) 13/52 = 0.25
(c) 1/6
(d) Event is independent (1/6) (1/6) = (1/36)
(e) (4/52) (3/51) = 0.0045

7.3

(a) True
(b) True
(c) True
(d) False

7.4
b is more probable

7.5

(a) Independent
(b) Dependent
(c) Dependent
(d) Independent
(e) Dependent

Car insurance is higher

7.6
The two-engine boat is safer because the probability of engine failure is lower.

7.7
Expected winnings: $E(w) = (0.5)15\ 2000000000 + (1 - 0.5)15\ 10000000$
It depends on risk averse, risk neutral or risk lover attitudes.

7.8

P (I or R) = (I) + P (R) – P (I and R) Mutually exclusive event.
P (I or R) = 0.6 + 0.4 − 0.12 = 0.88

15.4.4 Chapter 8: Selected Answers to the Review Problems

8.2

(a) There are only two mutually exclusive outcome independent n trials: probability of occurrence or success constant.

(b)
$$P(3) = \frac{7}{3!(7-3)!} 0.5^3 0.5^{5-3} = 0.027344$$

(c) P(0) = 0.007813, P(1) = 0.54688, P(2) = 0.164063

P(X < 3) = 0.007813 + 0.54688 + 0.164063 = 0.226564

8.3

c) P(3) = 0.000842

8.4

(a) P (4) = 0.018917

(b) P(0) = 0.00004539, P(1) = 0.0004539, P(2) = 0.002269, P(3) = 0.007566. Therefore, 0.00004539 + 0.0004539 + 0.002269 + 0.007566 + 0.018917 = 0.0292513

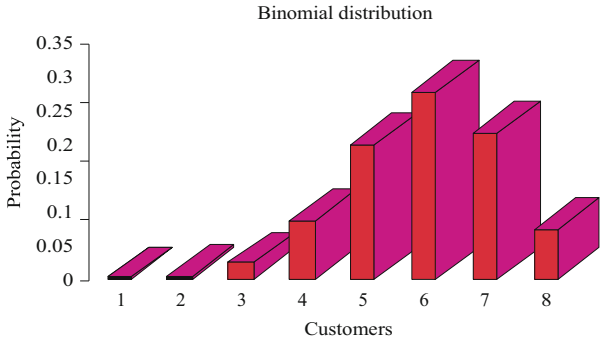
(c) The expected value, or mean, of this Poisson distribution is 10 customers, and the standard deviation is $\sqrt{10} = 3.162$

15.4.5 Chapter 8: Selected Computing Practical Solutions

C.8.1

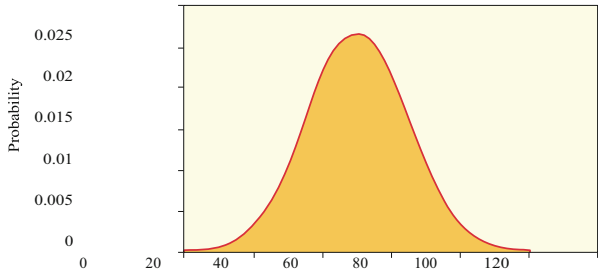
X	X!	n = (n-X)	7 (n-X)!	p= nCxForm. nCx	0.7 p^x		1-p= (1-p)^n-X	0.3 binomp	bformula
0	1	7	5040	1	1	1	0.000219	0.00022	2E-04
1	1	6	720	7	7	0.7	0.000729	0.00357	0.004
2	2	5	120	21	21	0.49	0.00243	0.025	0.025
3	6	4	24	35	35	0.343	0.0081	0.09724	0.097
4	24	3	6	35	35	0.24	0.027	0.22689	0.227
5	120	2	2	21	21	0.168	0.09	0.31765	0.318
6	720	1	1	7	7	0.118	0.3	0.24706	0.247
7	5040	0	1	1	1	0.082	1	0.08235	0.082

n=7
p=0.7
1-p=0.3 nCxForm



8.2

The normal distribution diagram:



15.4.6 Chapter 9: Selected Answers to the Review Problems

9.2

Since n > 25, the theoretical sampling distribution of the mean is approximately normal. $\overline{X} - Z\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + Z\frac{\sigma}{\sqrt{n}}$ Substitution of the given values for 95 % confidence interval gives:

$$\overline{X} - Z\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + Z\frac{\sigma}{\sqrt{n}} = 180 - 1.96\frac{25}{\sqrt{90}} < \mu < 180 + 1.96\frac{25}{\sqrt{90}}$$

$$174.84 < \mu < 185.165$$

9.3

$P = \frac{80}{120} = 0.67$. The sampling distribution of P is: $p \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$

$$\left[p - 1.96\sqrt{\frac{0.67(1-0.67)}{120}} < \pi < p + 1.96\sqrt{\frac{0.67(1-0.67)}{120}}\right]$$
$$= [0.586 < \pi < 0.754]$$

Thus, the proportion of all the workers in the factory who prefer their own retirement is between 0.586 and 0.754 with 95 % degree of confidence.

15.4.7 Chapter 10: Selected Answers to the Review Problems

10.1

(a) The acceptance or rejection of an assumption made about an unknown characteristic of a population, such as a parameter or the shape or form of the population distribution.
(b) The rejection of a true hypothesis is considered as a type one error. The acceptance of a false hypothesis is a type two error.

10.4
a) 5 % significance:

$$\begin{array}{lll} H_0: \mu_1 = \mu_2 & \text{Or} & H_0: \mu_1 - \mu_2 = 0 \\ H_1: \mu_1 \neq \mu_2 & \text{Or} & H_1: \mu_1 - \mu_2 \neq 0 \end{array}$$

$$\begin{array}{lll} \overline{x}_1 = 1050h & S_1 = 95h & n_1 = 120 \\ \overline{x}_2 = 1200h & S_2 = 145h & n_2 = 120 \end{array}$$

This is a two-tail test with an acceptance region within ± 1.96 under the standard normal curve. Therefore,

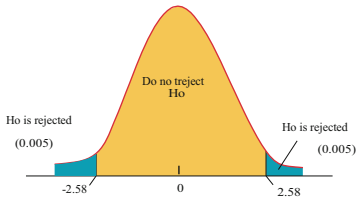
$$Z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$\frac{(1050 - 1200) - 0}{\sqrt{\frac{95^2}{120} + \frac{145^2}{120}}} = \frac{1050 - 1200}{\sqrt{75.21 + 175.21}} = -9.48$$

Since the calculated value of Z falls within the rejection region for, the buyer should accept H1, that μ1 ≠ μ2, at the 5 % level of significance (and presumably decide to purchase brand two)

b) 1 % significance:

At the 1 % level of significance, the calculated Z value would also not fall within the acceptance region for Ho (2.58 < 9.48). This would indicate that there is a significant difference between μ1 and μ2 and at the 1 % level, so the buyer could buy brand two.



10.5

a) The 5 % level of significance:

H0 : π1 = π2	*and*	H1 : π1 > π2
$p1 = 0.55$	*and*	$n1 = 70$
$p_2 = 0.45$	*and*	$n_2 = 50$

This is a right-tail test and the acceptance region for Ho with a = 0.05 lies to the left of 1.64 under the standard normal curve:

$$Z = \frac{(p_1 - p_2)}{\sqrt{\frac{p'(1-p')}{n_1} + \frac{p'(1-p')}{n_2}}}$$

where p' is $\frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{70(0.55)+50(0.45)}{70+50} = 0.51$

Thus $Z = \frac{(0.55-0.45)}{\sqrt{\frac{0.51(1-0.51)}{70} + \frac{0.51(1-0.51)}{50}}} = 1.08$, we accept Ho that , with $\alpha = 0.05$

b) The 10 % level of significance:

With a = 0.10, the acceptance region for Ho lies to the left of 1.28 under the standard normal curve. Since the calculated Z falls within the acceptance region, we accept H_0 at a = 0.10 as well.

10.7

Mapana and Disgate test:

$$H_0 : \mu_1 = \mu_2 \quad \text{and} \quad H_1 : \mu_1 \neq \mu_2$$

$$\bar{x}_1 = 30 \qquad S_1 = 5 \qquad n_1 = 18$$

$$\bar{x}_2 = 28 \qquad S_2 = 4 \qquad n_2 = 18$$

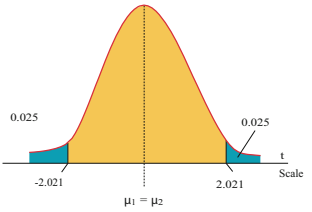
Since the two populations are normally distributed but n_1 and $n_2 < 30$ and it is assumed that σ2 = σ2 (but unknown), the sampling distribution of the difference

between the means has a t distribution with $n_1 + n_2 - 2$ degrees of freedom. Since it is assumed that $\sigma_1^2 = \sigma_2^2$ (and we can use S_1^2 as an estimate of σ_1^2 and S_2^2 as an estimate of σ_2^2), we get $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$ S^2 is a weighted average of S_1^2 and S_2^2. The weights are $n_1 - 1$ and $n_2 - 1$ for S_1^2 and S_2^2, in order to get 'unbiased' estimates for σ_1^2 and σ_2^2. This is a two-tail test with the acceptance region for H_o within $\pm$ 2.021 under the t distribution with $\alpha = 5\,\%$ and $n_1 + n_2 - 2 = 18 + 18 - 2 = 34df$:

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{(18 - 1)5^2 + (18 - 1)4^2}{18 + 18 - 2} = 20.5$$

$$t = \frac{(30 - 28) - 0}{\sqrt{\frac{20.5}{18} + \frac{20.5}{18}}} = \frac{30 - 28}{1.51} \cong 1.32$$

Since the calculated value of t falls within the acceptance region, we cannot reject H_0, that $\mu_1 = \mu_2$



15.4.8 Chapter 11: Selected Answers to the Review Problems

11.1

The expected number of sick days are equal to the proportion of the age group in the labour force (considering that age is not a factor for taking sick days)

$$\chi^2 = \frac{(O_i - E_i)^2}{E_i} = \frac{(25 - 35)^2}{35} + \frac{(35 - 38)^2}{38} + \frac{(26 - 16)^2}{16} + \frac{(14 - 11)^2}{11}$$

$$= 10.165$$

d.f. = 4 - 1 = 3. Since the calculated value of the chi-squared value (10.165) is greater than the tabular value (7.81) with 95 % confidence interval and 3 d.f. We reject the Ho: that the age is a factor for taking sick days. Here the chi-squared test is used as a right tail test only.

11.2

The expected frequencies $E\ i$ s are estimated as follows:

$$E_1 = \frac{\sum r \sum c}{n} = \frac{(80)(90)}{150} = 48$$

$$E_2 = \frac{\sum r \sum c}{n} = \frac{(20)(90)}{150} = 12$$

After similar calculations the following table is obtained:

Expected Frequencies of traffic accidents

Age group	Male	Female	Total
Below 35	48	32	80
From 35–60	12	8	20
Above 60	30	20	50
Total	90	60	150

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(50-48)^2}{48} + \frac{(30-32^2}{32} + \frac{(10-12)^2}{12} + \frac{(30-30)^2}{30} + \frac{(20-20)^2}{20} = 0.542$$

For 1 % level of significance and 2 degrees of freedom, the tabular chi-squared value (9.21) is greater than the calculated (0.542) value. The hypothesis is Ho: that age is in- dependent of sex in the occurrence of traffic accidents. To be sure, males seem to have more traffic accidents but this tendency does not differ significantly with age at the 1 % level of significance.

15.4.9 Chapter 11: Selected Computing Practical Solutions

C.11.1

Ho: No association between income groups

H1: An association between the two.

The calculated $\chi 2 = 12.59$ The result is dependent on the significance level. A higher significance level means the rejection of the null hypothesis.

C.11.2

One-way ANOVA:

SUMMARY

Groups	Count	Sum	Average	Variance
Talkswagen	5	1075	215	62.5
Mofas	5	925	185	59.5
Sayota	5	975	195	54.5
Sum	5	2975	595	519.5

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between groups	592375	3	197458.3	1134.818	7.64E-19	3.238867
Within groups	2784	16	174			
Total	595159	19				

Ho is rejected: the population servicing cost is not likely to be same for 5 % level of significance.

C.11.4

Anova: Two-factor without replication

SUMMARY

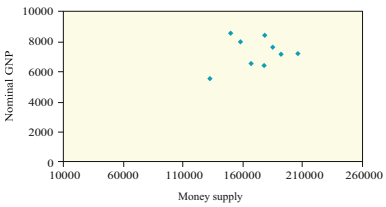
		Count	Sum	Average	Variance	
	225	3	1015	338.3333	59532.33	
	220	3	1006	335.3333	57836.33	
	215	3	979	326.3333	54981.33	
	210	3	950	316.6667	52033.33	
	205	3	925	308.3333	49433.33	
Mofas	5	925	185	59.5		
Sayota	5	975	195	54.5		
Sum	5	2975	595	519.5		
ANOVA						
Source of variation	SS	df	MS	F	P-value	F crit
Rows	1900.667	4	475.1667	6.002105	0.01561	3.837854
Columns	547000	2	273500	3454.737	1.79E-12	4.458968
Error	633.3333	8	79.16667			
Total	549534	14				

The hypothesis of equal average population servicing cost as well is rejected: the population values of mean servicing cost are not the same for each different level of trained service mechanics.

15.4.10 Chapter 12: Selected Answers to the Review Problems

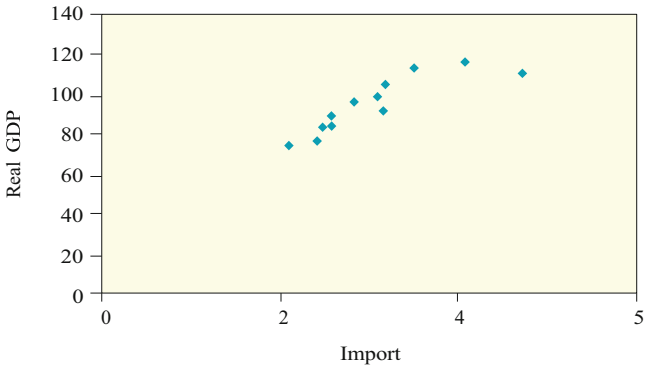
12.1
The rank correlation coefficient is 0.72. This indicates a positive rank correlation between the marks obtained from assignments and the exam. E.g. High assignment marks are associated with high exam performance.

12.2
The scatter diagram:



The correlation coefficient r is equal to 0.21 and $r2$ is 0.044. There appears to be a very loose correlation between the money supply and nominal GNP for the period between 1991 and 1999.

12.3
The scatter diagram of the data:

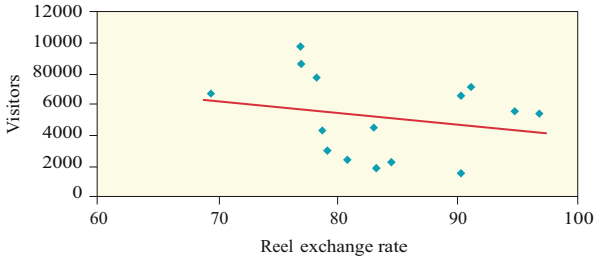


From the scatter diagram, there appears to be an upward relationship between Import and the Real GDP. This should have a positive correlation coefficient, as is suggested by the theory. Imports are dependent on the real GDP because whenever income rises, the demand for all domestic and foreign products rises too. The calculated correlation coefficient is 0.88 and $r2$ is 0.77. The goodness of fit is high and there is a positive and significantly high correlation between imports and real GDP.

15.5 Chapter 13: Selected Answers to the Review Problems

13.2

The relationship between the number of tourist arrivals and the real exchange rate in Turkey. The scatter diagram and the regression line of the figure below shows us this relationship:



As can be seen from the diagram, there is a negative relationship between the real exchange rate and the number of visitors. As the real exchange rate goes down (i.e. Turkey becomes cheaper for foreigners) the number of tourists visiting goes up. The estimated regression equation is:

$$\text{Visitors} = 11322.56 - 74,08\ *(\text{real exchange rate})$$

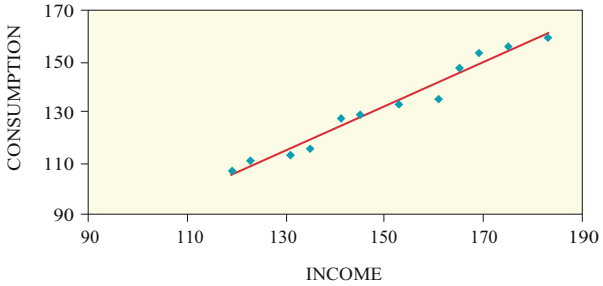
However, the calculated $R2$ is 0.05, which is very low. Hence the estimation with this regression may not be a very good estimate. If we make a prediction for the real exchange rate, 50.50, the number of visitors would be 7581 and a further decrease in the exchange rate such as 45 would bring the number of visitors up to 7989.

13.3

(a) The 'r' is not changed for changes in units. A change in the measure of income (an increase or a decrease in the numerical value) would change the coefficient.
(b) If milk is measured in gallons, the numbers predicting it would be divided by 8, so the constant (1500 /8) = 187.5 and coefficient (600/8) = 75.
(c) The numbers for milk are multiplied by 52 (there are 52 weeks in a year) and income numbers are also multiplied by 52. So just the constant must change and multiplied by 52.

13.4

a) The regression on consumption is presented on the following graph, which clearly proves the theoretical argument (if disposable income goes up, consumption goes up too).



The estimated regression equation is: $C = 2.82 + 0.86Yd$

The estimator a (=2.82) is the autonomous part of the consumption and the estimator b (=0.86) is the marginal propensity to consume.

b)

$$S^2 = \frac{\sum e_i^2}{n-k} = \frac{115.27}{12-2} = 11.53$$

$$= \frac{115.27}{12-2} = 11.53, \qquad S_a^2 = 54.86, \qquad S_a = 7.41$$

$$S_b^2 = 0.002395, \qquad S_b = 0.0489$$

c) To test for the statistical significance of 'a' and 'b', we set the following null hypothesis, Ho and alternative hypothesis, H1

$H_0 : \beta_1 = 0$	versus	$H_1 : \beta_1 \neq 0$
$H_0 : \beta_2 = 0$	versus	$H_1 : \beta_2 \neq 0$

The hope in regression analysis is to reject Ho and to accept H1, that 'a' and 'b' are not equal to zero with a two-tail test. Testing at the 5 % level of significance for 'a' :

$$t_a = \frac{a - \beta_1}{S_a} \cong \frac{2.82 - 0}{7.41} = 0.38$$

Since t_a is smaller than the tabular value of t = 2.228 at the 5 % level (two-tail test) and with 10 degrees of freedom, we conclude that β_1 is not statistically significant at the 5 % level (i.e., we cannot reject Ho, that a = 0).

$$t_b = \frac{b - \beta_2}{S_b} \cong \frac{0.86 - 0}{0.0489} = 17.59$$

So β2 is statistically significant at the 5 % (and 1 %) level, that is we cannot reject H1, that $\beta 1 \neq 0$.

d) Construct the 95 % confidence interval for parameters 'a' and 'b': The 95 % confidence interval for is given by,

$\beta_1 = a \pm 2.228 S_a = 2.823 \pm 2.228(7.41) = 2.823 \pm 16.51$

So β_1 is between -13.687 and 19.3 with 95 % confidence. Note how wide (and meaningless) the 95 % confidence interval β_1 is, reflecting the fact that 'a' is highly in- significant.

The 95 % confidence interval for β_2 is given by,

$\beta_2 = b \pm 2.228 S_b = 0.86 \pm 2.228(0.0489) = 0.86 \pm 0.1086$

So β_2 is between 0.75 and 0.97 (that is $0.75 < \beta_2 < 0.97$) with 95 % confidence.

e) $R2 = 0.9687$

f) n = 12, k = 2. $H_o : R^2 = 0$ $H_1 : R^2 \neq 0$

$$F_{k-1,n-k} = \frac{0.9687/(2-1)}{(1-0.9687)/(12-2)} = \frac{0.9687}{0.0313/10} = 309.\ 48$$

Since the calculated value of F exceeds the tabular value of F (F1,10 = 4.96). Ho is rejected. The regression as a whole is significant.

15.5.1 Chapter 13: Selected Computing Practical Solutions

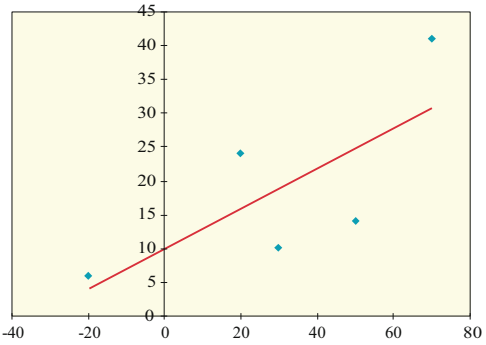
C.13.1 a) The completed table is as follows:

	Investment	Change in sales	Reg. line			
Year	Y	X	X*Y	X^2	Y^2	Y = 10 + 0.3*X
1988	10	30	300	900	100	19
1989	24	20	480	400	576	16
1990	41	70	2870	4900	1681	31
1991	14	50	700	2500	196	25
1992	6	-20	-120	400	36	4
Total	95		4230	9100	2589	
Mean	19					$r2 = 0.53$

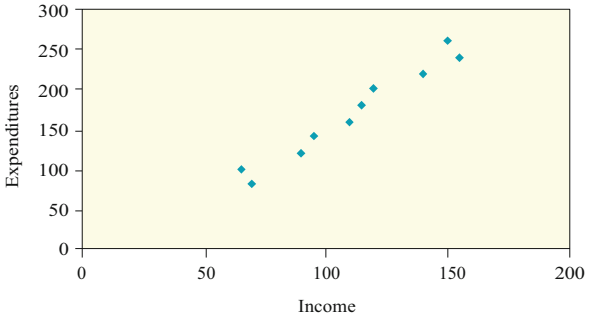
b) The 'linest' output is provided the same coefficient and the correlation values. The- se are presented in the following linest output.

0.3	10
0.163742	6.985492
0.528061	11.10555
3.356757	3
414	370

c) The scatter diagram and the regression line.



C.13.3 a) The scatter diagram:



An explanation of the relationship can be given by an estimation of a regression equation. The estimated equation is $Y = 24 + 0.5\ X$. The prediction of expenditures for different levels of income is as follows:

The 'linest' printout is:

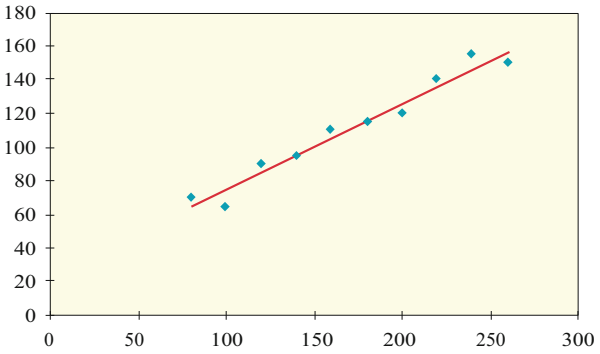
0.509091	24.45455
0.035743	6.413817
0.962062	6.493003
202.8679	8
8552.727	337.2727

Income	Expenditure.	Y(estimate)
80	70	65.18181818
100	65	75.36363636
120	90	85.54545455
140	95	95.72727273
160	110	105.9090909
180	115	116.0909091
200	120	126.2727273

(continued)

220	140	136.4545455
240	155	146.6363636
260	150	156.8181818

The fitted regression line:



15.5.2 Chapter 14: Selected Answers to the Review Problems

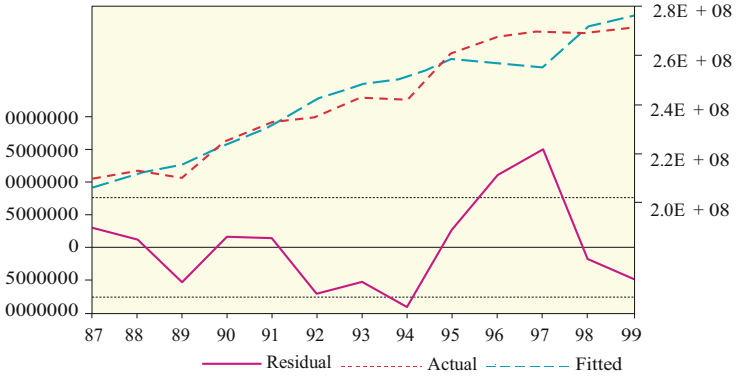
14.2

a) The regression analysis on food expenditures is as follows: Food expenditures are determined by two factors, the CPI (general price level) and the wages as suggested. Food = -84905154 - 158.187*CPI* + 1.697*WAGE*

The relevant Eviews print out is:

Dependent variable: FOOD Method: Least squares Sample: 1987 1999 Included observations: 13				
Variable	Coefficient	Std. error	t-Statistic	Prob.
C	-84905154	46153591	-1.839622	0.0957
CPI	-158.1874	144.1100	-1.097685	0.2981
WAGE	7.697899	1.117226	6.890191	0.0000
R-squared	0.911815	Mean dependent var	2.42E + 08	
Adjusted R-squared	0.894178	S.D. dependent var	23577223	
S.E. of regression	7669740.	Akaike info criterion	34.74264	
Sum squared resid	5.88E + 14	Schwarz criterion	34.87301	
Log likelihood	-222.8271	F-statistic	51.69898	
Durbin-Watson stat	1.203229	Prob(F-statistic)	0.000005	

b) The summary graph of the residuals:



c) Relevant tests: The t test, F test and the DW test

The t test:

The t test is based on n - p - 1 degrees of freedom. The degrees of freedom is v =29-2-1 = 26. For 5 % significance level, the two-tail test gives the table value for t t 26* = ±2.045.

By comparing the calculated t values in the Eviews printout, we can say that all para- meters are not statistically significant.

Since the DW statistic, 1.203, is less than the *dL* we can say that there is a positive autocorrelation. Due to this fact all parameters appeared to be statistically insignificant in t tests. One of the most common ways of solving the autocorrelation problem is to convert the data into the logarithmic form.The following Eviews printout provides the logarithmic transformation of the same regression:

Dependent variable: LOG (FOOD)
Method: Least squares
Date: 05/01/01 Time: 17:34
Sample: 1987 1999
Included observations: 13

Variable	Coefficient	Std. error	t-Statistic	Prob.
C	17.99386	7.750238	2.321718	0.0426
LOG(CPI)	0.040955	0.015666	2.614347	0.0258
LOG(WAGE)	0.056058	0.447986	0.125132	0.9029
R-squared	0.947821	Mean dependent var	19.30176	
Adjusted R-squared	0.937386	S.D. dependent var	0.098134	
S.E. of regression	0.024556	Akaike info criterion	-4.376556	
Sum squared resid	0.006030	Schwarz criterion	-4.246183	
Log likelihood	31.44761	F-statistic	90.82474	
Durbin-Watson stat	1.560884	Prob(F-statistic)	0.000000	

The logarithmic estimation result passes the DW test, and the t tests for the constant term and the log (CPI) indicate that these parameters are statistically significant. The wage parameter, however, is not.

The F^* 1,26 = 4.242 for 5 % level of significance. The test statistic (90.824) exceeds this, so regression as a whole is significant.

15.5.3 Chapter 14: Selected Computing Practical Solutions

C.14.1

a) The estimated Eviews printout is as follows:

Dependent variable: INVEST
Method: Least squares
Sample: 1987 2000
Included observations: 14

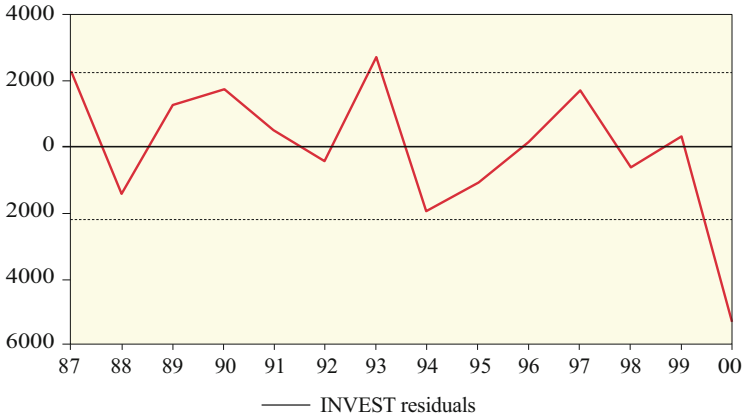
Variable	Coefficient	Std. error	t-Statistic	Prob.
C	-14307.39	4322.846	-3.309715	0.0070
GDPFC	0.343000	0.050355	6.811619	0.0000
LTIR	93.74772	36.57427	2.563215	0.0264
R-squared	0.884074	Mean dependent var	24602.25	
Adjusted R-squared	0.862997	S.D. dependent var	6114.001	
S.E. of regression	2263.030	Akaike info criterion	18.47421	
Sum squared resid	56334369	Schwarz criterion	18.61115	
Log likelihood	-126.3194	F-statistic	41.94425	
Durbin-Watson stat	1.758209	Prob(F-statistic)	0.000007	

Thus the summary of the estimated regression equation:

The t test is based on n - p - 1 degrees of freedom where n is the number of observation and p the number of independent variables. Hence the degrees of freedom is v =14-2-1 = 11. For 5 % significance level, the two-tail test gives the table value for t t 11* = ± 2.201.

By comparing the calculated t values in the Eviews output we can say that all parameters are statistically significant.

The significance of the regression equation as a whole can be tested by using the F distribution. For 5 % significance level, $v1 = 1, v2 =$ n-p-1 = 11 or $F^* = 4.844$ or The calculated F statistics (41.944) exceeds this. So the regression as a whole is significant or the relationship between dependent and independent variables is significant.



b) The logarithmic investment estimation is presented in the following Eviews printouts.

Dependent variable: LOG (INVEST)
Method: Least squares
Sample: 1987 2000
Included observations: 14

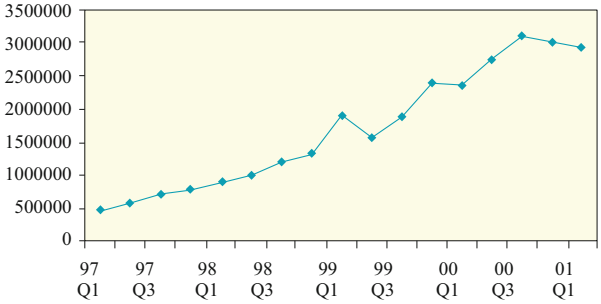
Variable	Coefficient	Std. error	t-Statistic	Prob.
C	-5.714628	2.343805	-2.438184	0.0329
LOG(GDPFC)	1.268652	0.217817	5.824383	0.0001
LOG(LTIR)	0.302545	0.117137	2.582835	0.0255
R-squared	0.847608	Mean dependent var	10.08143	
Adjusted R-squared	0.819901	S.D. dependent var	0.252250	
S.E. of regression	0.107050	Akaike info criterion	-1.443628	
Sum squared resid	0.126057	Schwarz criterion	-1.306688	
Log likelihood	13.10540	F-statistic	30.59122	
Durbin-Watson stat	1.879151	Prob(F-statistic)	0.000032	

Chapter 15: Selected Answers to the Review Problems

15.1

a) Mo (monetary base) may vary seasonally because demand for Mo may be higher in some parts of the years. For example it could be high in the west in January due to the New Year increase in demand.

b) In Turkey however the first point that we can observe an ever-increasing monetary base, which may explain the high inflation. Although Mo does not appeared to fluctuate seasonally in any great sense, an observation of the diagram below shows us a recent January variation, (First quarters of years 1999, 2000 and 2001) which is a monetary base increase in the first quarters of those years which then declines.



c) A centered 4 point moving average trend for Mo: The additive model is about adding all factors:

A (Actual data) = T (Trend) + S(Seasonal factors) + C(Cyclical factors) + R (Random factors).

The difficulty in identifying data for cyclical elements, and the unpredictability of the random element leave us with A = T + S. Similarly, the 'multiplicative' method is A = T×S.

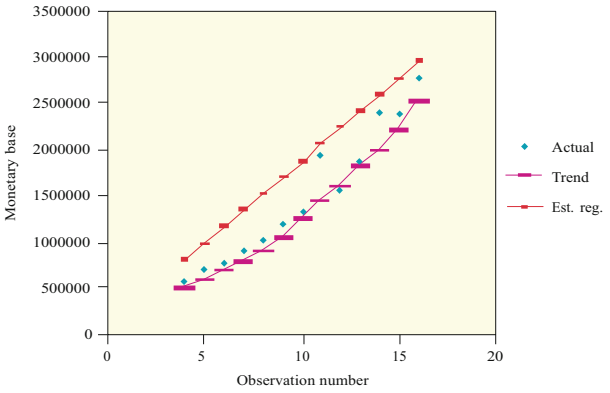
We will be using the additive method for this exercise. The data in the equation are quarterly and the appropriate subset size will be four. The following Excel table shows the 4 point moving average calculations.

Date	Quarters	Observation number(X)	Actual (A) M0 billion TL	Sum of 4 quarters	Sum of 8 quarters	Average (T8 qrts/8)	S = A - T	Mo=a +bX Est. reg.
1996	Q3	1	379578.1					266166
	Q4	2	382242.9					445023
1997	Q1	3	459439.1	1788571				623880
	Q2	4	567310.6	2111313	3899883.3	487485.4	79825.2	802737
	Q3	5	702320	2487948	4599260.3	574907.5	127412	981594
	Q4	6	758878	2926397	5414344.3	676793	82085	1160451
1998	Q1	7	897888	3365074	6291470.9	786433.9	111454	1339308
	Q2	8	1005988	3858046	7223119.8	902890	103098	1518165
	Q3	9	1195291	4427710	8285755.4	1035719	159572	1697022
	Q4	10	1328542	5456788	9884497.7	1235562	92980.2	1875879
1999	Q1	11	1926966	6010494	11467281.4	1433410	493556	2054736
	Q2	12	1559694	6685484	12695977.2	1586997	-27303	2233593
	Q3	13	1870281	7747690	14433173.2	1804147	66134.6	2412450
	Q4	14	2390748	8192152	15939842	1992480	398268	2591307
2000	Q1	15	2371429	9386945	17579097.4	2197387	174042	2770164
	Q2	16	2754487	10637319	20024263.6	2503033	251454	2949021
	Q3	17	3120655					3127878

d) The seasonally adjusted series are presented in the table above in column 7.

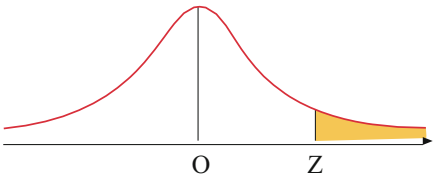
e) Column 9 in the table gives a different trend by using the estimated regression equation.

f) The following graph presents both the estimated regression trend and the seasonally adjusted series



Appendix Tables

Table 1 The standard normal distribution



x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4364	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233

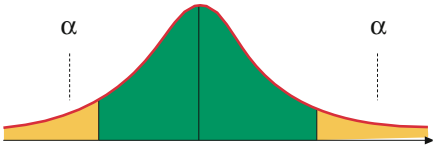
(continued)

D. Özdemir, *Applied Statistics for Economics and Business*,
DOI 10.1007/978-3-319-26497-4

Table 1 (continued)

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010

Table 2 Percentage points of the t distribution



v	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0250.0	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.656	127.32	318.29	636.58
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.328	31.600
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.364	4.032	4.773	5.894	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819

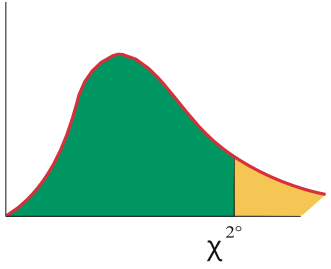
(continued)

Table 2 (continued)

v	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0250.0	0.001	0.0005
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.689
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.660
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
35	0.255	0.682	1.306	1.690	2.030	2.438	2.724	2.996	3.340	3.591
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	0.255	0.679	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

The table gives critical values of the *t* distribution cutting off an area α in each tail, shown by the top row of the table
Area (α) in each tail

Table 3 Critical values of the $\chi 2$ distribution



v	0.995	0.990	0.975	0.950	0.900	0.750	0.500	0.250	0.100	0.050	0.025	0.010	0.005	0.001
1	39.3E−6	1.57E−4	982e−6	3.93E−3	0.016	0.102	0.455	1.323	2.706	3.841	5.024	6.635	7.879	10.827
2	0.010	0.020	0.051	0.103	0.211	0.575	1.386	2.773	4.605	5.991	7.378	9.210	10.597	13.815
3	0.072	0.115	0.216	0.352	0.584	1.213	2.366	4.108	6.251	7.815	9.348	11.345	12.838	16.266
4	0.207	0.297	0.484	0.711	1.064	1.923	3.357	5.385	7.779	9.488	11.143	13.277	14.860	18.466
5	0.412	0.554	0.831	1.145	1.610	2.675	4.351	6.626	9.236	11.070	12.832	15.086	16.750	20.515
6	0.676	0.872	1.237	1.635	2.204	3.455	5.348	7.841	10.645	12.592	14.449	16.812	18.548	22.457
7	0.989	1.239	1.690	2.167	2.833	4.255	6.346	9.037	12.017	14.067	16.013	18.475	20.278	24.321
8	1.344	1.647	2.180	2.733	3.490	5.071	7.344	10.219	13.362	15.507	17.535	20.090	21.955	26.124
9	1.735	2.088	2.700	3.325	4.168	5.899	8.343	11.389	14.684	16.919	19.023	21.666	23.589	27.877
10	2.156	2.558	3.247	3.940	4.865	6.737	9.342	12.549	15.987	18.307	20.483	23.209	25.188	29.588
11	2.603	3.053	3.816	4.575	5.578	7.584	10.341	13.701	17.275	19.675	21.920	24.725	26.757	31.264
12	3.074	3.571	4.404	5.226	6.304	8.438	11.340	14.845	18.549	21.026	23.337	26.217	28.300	32.909
13	3.565	4.107	5.009	5.892	7.041	9.299	12.340	15.984	19.812	22.362	24.736	27.688	29.819	34.527
14	4.075	4.660	5.629	6.571	7.790	10.165	13.339	17.117	21.064	23.685	26.119	29.141	31.319	36.124
15	4.601	5.229	6.262	7.261	8.547	11.037	14.339	18.245	22.307	24.896	27.488	30.578	32.801	37.698
16	5.142	5.812	6.908	7.962	9.312	11.912	15.338	19.369	23.542	26.296	28.845	32.000	34.267	39.252

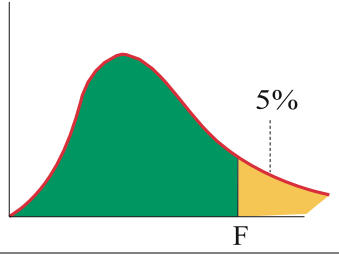
(continued)

Table 3 (continued)

v	0.995	0.990	0.975	0.950	0.900	0.750	0.500	0.250	0.100	0.050	0.025	0.010	0.005	0.001
17	5.697	6.408	7.564	8.672	10.085	12.792	16.338	20.489	24.769	27.587	30.191	33.409	35.718	40.791
18	6.265	7.015	8.231	9.390	10.865	13.675	17.338	21.605	25.989	28.869	31.526	34.805	37.156	42.312
19	6.844	7.633	8.907	10.117	11.651	14.562	18.338	22.718	27.204	30.144	32.852	36.191	38.582	43.819
20	7.434	8.260	9.591	10.851	12.443	15.452	19.337	23.828	28.412	31.410	34.170	37.566	39.997	45.314
21	8.034	8.897	10.283	11.591	13.240	16.344	20.337	24.935	29.615	32.671	35.479	38.932	41.401	46.796
22	8.643	9.542	10.982	12.338	14.041	17.240	21.337	26.039	30.813	33.924	36.781	40.289	42.796	48.268
23	9.260	10.196	11.689	13.091	14.848	18.137	22.337	27.141	32.007	35.172	38.076	41.638	44.181	49.728
24	9.886	10.856	12.401	13.848	15.659	19.037	23.337	28.241	33.196	36.415	39.364	42.980	45.558	51.179
25	10.520	11.524	13.120	14.611	16.473	19.939	24.337	29.339	34.382	37.652	40.646	44.314	46.928	52.619
26	11.160	12.198	13.844	15.379	17.292	20.843	25.336	30.435	35.563	38.885	41.923	45.642	48.290	54.051
27	11.808	12.878	14.573	16.151	18.114	21.749	26.336	31.528	36.741	40.113	43.195	46.963	49.645	55.475
28	12.461	13.565	15.308	16.928	18.939	22.657	27.336	32.620	37.916	41.337	44.461	48.278	50.994	56.892
29	13.121	14.256	16.047	17.708	19.768	23.567	28.336	33.711	39.087	42.567	45.722	49.588	52.335	58.301
30	13.787	14.953	16.791	18.493	20.599	24.478	29.336	34.800	40.256	43.773	46.979	50.892	53.672	59.702
40	20.707	22.164	24.433	26.509	29.051	33.660	39.335	45.616	51.805	55.758	59.342	63.691	66.766	73.403
50	27.991	29.707	32.357	34.764	37.689	42.942	349.335	56.334	63.167	67.505	71.420	76.154	79.490	86.660
60	35.534	37.485	40.482	43.188	46.459	52.294	59.335	66.981	74.397	79.082	83.298	88.379	91.952	99.608
70	43.275	45.442	48.758	51.739	55.329	61.698	69.334	77.577	85.527	90.531	95.023	100.425	104.215	112.317
80	51.172	53.540	57.153	60.391	64.278	71.145	79.334	88.130	96.578	101.879	106.629	112.329	116.321	124.839
90	59.196	61.754	65.647	69.126	73.291	80.625	89.334	98.650	107.565	113.145	118.136	124.116	128.299	137.208
100	67328	70.065	74.222	77.929	82.358	90.133	99.334	109.141	118.498	124.342	129.561	135.807	140.170	149.449

The values in the table give the critical values of $\chi 2$ cutting off the area in the right-hand tail given at the top of the columns

Table 4 (A) Critical values of the F distribution (upper 5 % points)



	v_1																		
v_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.785	8.745	8.703	8.660	8.638	8.617	8.594	8.572	8.549	8.527
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.912	5.858	5.803	5.774	5.746	5.717	5.688	5.658	5.628
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735	4.678	4.619	4.558	4.527	4.496	4.464	4.431	4.398	4.365
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060	4.000	3.938	3.874	3.841	3.808	3.774	3.740	3.705	3.669
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637	3.575	3.511	3.445	3.410	3.376	3.340	3.304	3.267	3.230
8	5.318	4.459	4.066	3.838	3.688	3.581	3.500	3.438	3.388	3.347	3.284	3.218	3.150	3.115	3.079	3.043	3.005	2.967	2.928
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137	3.073	3.006	2.936	2.900	2.864	2.826	2.787	2.748	2.707
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978	2.913	2.845	2.774	2.737	2.700	2.661	2.621	2.580	2.538
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854	2.788	2.719	2.646	2.609	2.570	2.531	2.490	2.448	2.405
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753	2.687	2.617	2.544	2.505	2.466	2.426	2.384	2.341	2.296
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671	2.604	2.533	2.459	2.420	2.380	2.339	2.297	2.252	2.206
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602	2.534	2.463	2.388	2.349	2.308	2.266	2.223	2.178	2.131
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544	2.475	2.403	2.328	2.288	2.247	2.204	2.160	2.114	2.066
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494	2.425	2.352	2.276	2.235	2.194	2.151	2.106	2.059	2.010

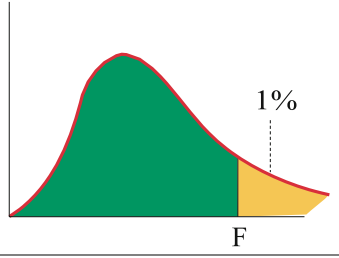
(continued)

Table 4 (continued)

	v_1																		
$\mathbf{v_2}$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450	2.381	2.308	2.230	2.190	2.148	2.104	2.058	2.011	1.960
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412	2.342	2.269	2.191	2.150	2.107	2.063	2.017	1.968	1.917
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378	2.308	2.234	2.155	2.114	2.071	2.026	1.980	1.930	1.878
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348	2.278	2.203	2.124	2.082	2.039	1.994	1.946	1.896	1.843
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321	2.250	2.176	2.096	2.054	2.010	1.965	1.916	1.866	1.812
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297	2.226	2.151	2.071	2.028	1.984	1.938	1.889	1.838	1.783
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275	2.204	2.128	2.048	2.005	1.961	1.914	1.865	1.813	1.757
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255	2.183	2.108	2.027	1.984	1.939	1.892	1.842	1.790	1.733
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236	2.165	2.089	2.007	1.964	1.919	1.872	1.822	1.768	1.711
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165	2.092	2.015	1.932	1.887	1.841	1.792	1.740	1.683	1.622
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077	2.003	1.924	1.839	1.793	1.744	1.693	1.637	1.577	1.509
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026	1.952	1.871	1.784	1.737	1.687	1.634	1.576	1.511	1.438
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993	1.917	1.836	1.748	1.700	1.649	1.594	1.534	1.467	1.389
120	3.920	3.072	2.680	2.447	2.290	2.175	2.087	2.016	1.959	1.910	1.834	1.750	1.659	1.608	1.554	1.495	1.429	1.352	1.254
∞	3.842	2.996	2.605	2.372	2.214	2.099	2.010	1.939	1.880	1.831	1.752	1.666	1.571	1.517	1.459	1.394	1.318	1.222	1.010

The entries in the table give the critical values of F cutting off 5 % in the right, hand tail of the distribution . v1 gives the degrees of freedom in the numerator, v2 those in the denominator

Table 4 (B) Critical values of the F distribution (upper 1 % points)



	v_1																		
v_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6107	6157	6209	6234	6260	6286	6313	6340	6366
2	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.48	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.888	9.722	9.553	9.466	9.379	9.291	9.202	9.112	9.021
6	13.75	10.92	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874	7.718	7.559	7.396	7.313	7.229	7.143	7.057	6.969	6.880
7	12.25	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620	6.469	6.314	6.155	6.074	5.992	5.908	5.824	5.737	5.650
8	11.26	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814	5.667	5.515	5.359	5.279	5.198	5.116	5.032	4.946	4.859
9	10.56	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257	5.111	4.962	4.808	4.729	4.649	4.567	4.483	4.398	4.311
10	10.04	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849	4.706	4.558	4.405	4.327	4.247	4.165	4.082	3.996	3.909
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539	4.397	4.251	4.099	4.021	3.941	3.860	3.776	3.690	3.603
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296	4.155	4.010	3.858	3.780	3.701	3.619	3.535	3.449	3.361
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100	3.960	3.815	3.665	3.587	3.507	3.425	3.341	3.255	3.165
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939	3.800	3.656	3.505	3.427	3.348	3.266	3.181	3.094	3.004
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805	3.666	3.522	3.372	3.294	3.214	3.132	3.047	2.959	2.869
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691	3.553	3.409	3.259	3.181	3.101	3.018	2.933	2.845	2.753

(continued)

Table 4 (continued)

	v_1																		
$\mathbf{v_2}$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
17	8.400	6.112	5.185	4.669	4.336	4.101	3.927	3.791	3.682	3.593	3.455	3.312	3.162	3.083	3.003	2.920	2.835	2.746	2.653
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508	3.371	3.227	3.077	2.999	2.919	2.835	2.749	2.660	2.566
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434	3.297	3.153	3.003	2.925	2.844	2.761	2.674	2.584	2.489
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368	3.231	3.088	2.938	2.859	2.778	2.695	2.608	2.517	2.421
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310	3.173	3.030	2.880	2.801	2.720	2.636	2.548	2.457	2.360
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258	3.121	2.978	2.827	2.749	2.667	2.583	2.495	2.403	2.306
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211	3.074	2.931	2.780	2.702	2.620	2.536	2.447	2.354	2.256
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168	3.032	2.889	2.738	2.659	2.577	2.492	2.403	2.310	2.211
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129	2.993	2.850	2.699	2.620	2.538	2.453	2.364	2.270	2.170
30	7.562	5.390	4.510	4.018	3.699	3.473	3.305	3.173	3.067	2.979	2.843	2.700	2.549	2.469	2.386	2.299	2.208	2.111	2.006
40	7.314	5.178	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801	2.665	2.522	2.369	2.288	2.203	2.114	2.019	1.917	1.805
50	7.171	5.057	4.199	3.720	3.408	3.186	3.020	2.890	2.785	2.698	2.563	2.419	2.265	2.183	2.098	2.007	1.909	1.803	1.683
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632	2.496	2.352	2.198	2.115	2.028	1.936	1.836	1.726	1.601
120	6.851	4.787	3.949	3.480	3.174	2.956	2.792	2.663	2.559	2.472	2.336	2.191	2.035	1.950	1.860	1.763	1.656	1.533	1.381
∞	6.635	4.605	3.782	3.319	3.017	2.802	2.640	2.511	2.408	2.321	2.185	2.039	1.878	1.791	1.697	1.592	1.473	1.325	1.015

The entries in the table give the criticalvalues of F cutting off 1 % in the right, 1 %hand tail of the distribution . v1 gives the degrees of freedom in thenumerator, v2 those in the Fdenominator

Table 5 Critical values for the Durbin Watson test at 5 % significance level

Number of explanatory variables										
Sample size	1		2		3		4		5	
n	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822
11	0.927	1.324	0.758	1.604	0.595	1.928	0.444	2.283	0.316	2.645
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.379	2.506
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	0.445	2.390
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.160	1.803
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.230	1.786
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.335	1.771
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.780
200	1.758	1.778	1.748	1.789	1.738	1.799	1.728	1.810	1.718	1.820

References

Barrow M (1996) Statistics for economics accounting and business studies, 2nd edn. Longman, Wallingford
Benson PG, McClave JT, Sincich T (1998) Statistics for business and economics, 7th edn. Prentice Hall, Upper Saddle River, NJ
Berenson ML, Levine DM (1996) Basic business statistics concepts and applications, 6th edn. Prentice Hall, Upper Saddle River, NJ
Brase CH, Brase CP (1995) Understandable statistics, 5th edn. D.C. Health and Company, Lexington, MA
Cochran WG (1977) Sampling techniques, 3rd edn. Wiley, New York
Curwin J, Slater R (1996) Quantitative methods for business decisions, 4th edn. Thomson, London
Deming WE (1960) Sample design in business research. Wiley, New York
Francis A (1996) Business mathematics and Statistics, 4th edn. DP Publications Ltd, London
Haeussler E Jr, Paul RS (1999) Introductory mathematical analysis, 9th edn. Prentice Hall, Upper Saddle River, NJ
Hansen MH, Hurwitz WN, Madow WG (1953) Sample survey methods and theory, vol 1 and 2. Wiley, New York
Koutsoyiannis A (1977) Theory of econometrics, 2nd edn. Macmillan, London
Lind DA (2000) Basic statistics for business and economics, 3rd edn. McGraw-Hill, Boston, MA
Maddala GS (1992) Introduction to econometrics, 2nd edn. Macmillan, Frederick, MD
Middleton MR (1997) Data analysis using Microsoft Excel. Duxbury, Belmont, CA
Newbold P (1995) Statistics for business and economics, 4th edn. Prentice Hall, Upper Saddle River, NJ
Thomas RL (1993) Introductory econometrics, 2nd edn. Longman, Wallingford

D. Özdemir, *Applied Statistics for Economics and Business*,
DOI 10.1007/978-3-319-26497-4