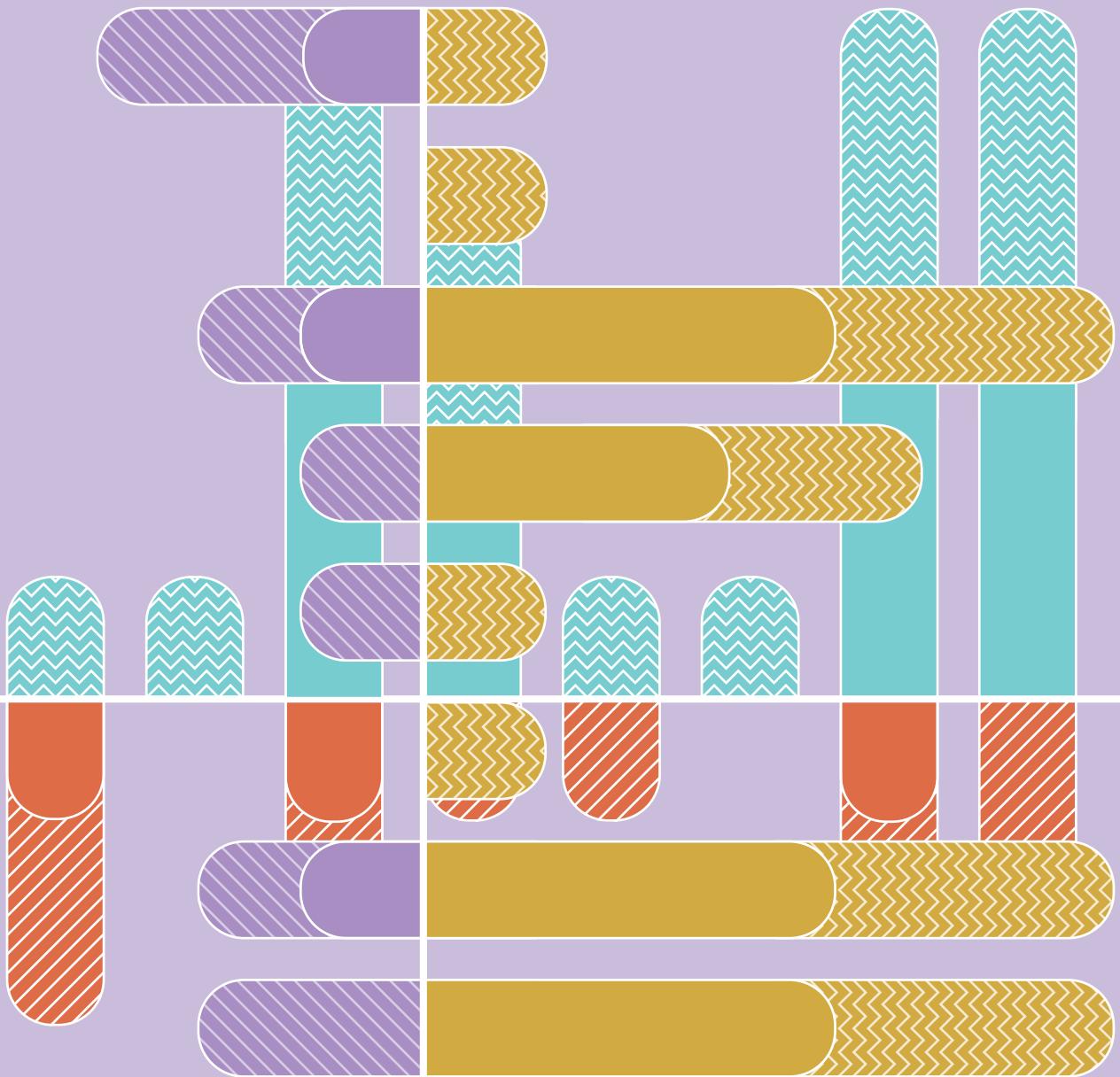


Business Statistics

Abridged

Australia / New Zealand
8th edition



Eliyathamby A. Selvanathan

Copyright 2021 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. WCN 02-200-202

Saroja Selvanathan

Gerald Keller

Business Statistics

Abridged

Australia / New Zealand
8th edition

Eliyathamby A. Selvanathan Saroja Selvanathan Gerald Keller

DEDICATION

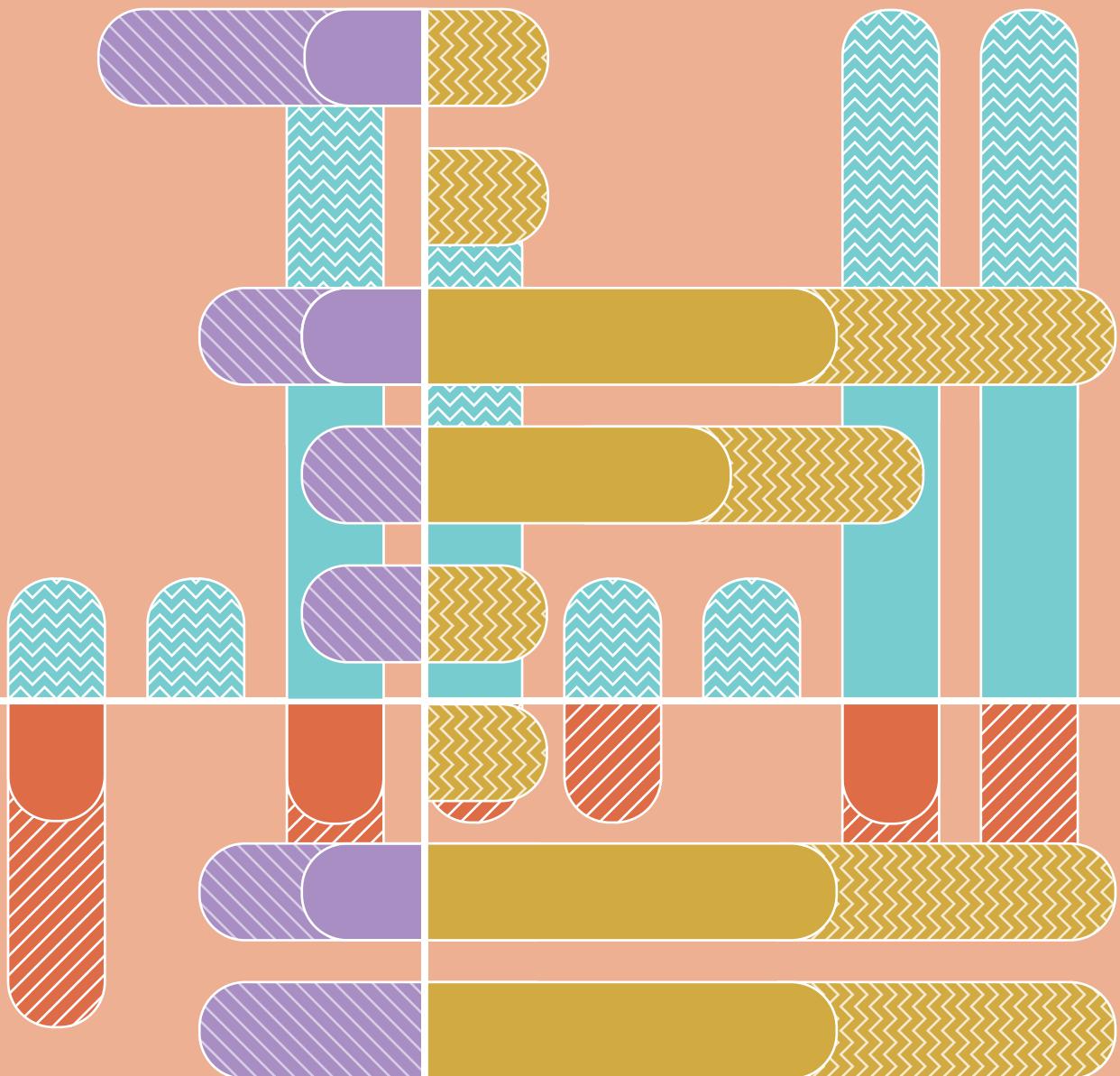
We would like to dedicate this book to our statistics gurus, the Foundation Lecturers of the Department of Mathematics and Statistics, University of Jaffna, Sri Lanka:

Late Professor Balan Selliah
Professor S Ganesalingam
Late Dr S Varatharajaperumal

Business Statistics

Abridged

Australia / New Zealand
8th edition



Eliyathamby A. Selvanathan Saroja Selvanathan Gerald Keller

Business Statistics, Abridged: Australia/New Zealand
8th Edition
Eliyathamby A. Selvanathan
Saroja Selvanathan
Gerald Keller

Head of content management: Dorothy Chiu
Senior content manager: Geoff Howard
Content developer: Emily Spurr
Senior project editor: Nathan Katz
Cover designer: Chris Starr (MakeWork)
Text designer: Watershed Art (Leigh Ashforth)
Permissions/Photo researcher: Wendy Duncan
Editor: Marta Veroni
Proofreader: James Anderson
Indexer: Max McMaster
Art direction: Nikita Bansal
Typeset by Cenveo Publisher Services

Any URLs contained in this publication were checked for currency during the production process. Note, however, that the publisher cannot vouch for the ongoing currency of URLs.

Adaptation of *Statistics for Management and Economics* 11e by Gerald Keller,
2017 ISBN: 9781337093453

This 8th edition published in 2021

© 2021 Cengage Learning Australia Pty Limited

Copyright Notice

This Work is copyright. No part of this Work may be reproduced, stored in a retrieval system, or transmitted in any form or by any means without prior written permission of the Publisher. Except as permitted under the Copyright Act 1968, for example any fair dealing for the purposes of private study, research, criticism or review, subject to certain limitations. These limitations include: Restricting the copying to a maximum of one chapter or 10% of this book, whichever is greater; providing an appropriate notice and warning with the copies of the Work disseminated; taking all reasonable steps to limit access to these copies to people authorised to receive these copies; ensuring you hold the appropriate Licences issued by the Copyright Agency Limited ("CAL"), supply a remuneration notice to CAL and pay any required fees. For details of CAL licences and remuneration notices please contact CAL at Level 11, 66 Goulburn Street, Sydney NSW 2000, Tel: (02) 9394 7600, Fax: (02) 9394 7601
Email: info@copyright.com.au
Website: www.copyright.com.au

For product information and technology assistance,

in Australia call 1300 790 853;

in New Zealand call 0800 449 725

For permission to use material from this text or product, please email
aust.permissions@cengage.com

National Library of Australia Cataloguing-in-Publication Data

ISBN: 9780170439541

A catalogue record for this book is available from the National Library of Australia.

Cengage Learning Australia

Level 7, 80 Dorcas Street
South Melbourne, Victoria Australia 3205

Cengage Learning New Zealand

Unit 4B Rosedale Office Park
331 Rosedale Road, Albany, North Shore 0632, NZ

For learning solutions, visit cengage.com.au

Printed in Singapore by 1010 Printing International Limited.

1 2 3 4 5 6 7 24 23 22 21 20



BRIEF CONTENTS

1	What is statistics?	1
2	Types of data, data collection and sampling	18

PART 1	Descriptive measures and probability	43
---------------	---	-----------

3	Graphical descriptive techniques - Nominal data	44
4	Graphical descriptive techniques - Numerical data	85
5	Numerical descriptive measures	135
6	Probability	211
7	Random variables and discrete probability distributions	260
8	Continuous probability distributions	309

PART 2	Statistical inference	349
---------------	------------------------------	------------

9	Statistical inference and sampling distributions	350
10	Estimation: Single population	373
11	Estimation: Two populations	429
12	Hypothesis testing: Single population	466
13	Hypothesis testing: Two populations	530
14	Chi-squared tests	582
15	Simple linear regression and correlation	624
16	Multiple regression	686

PART 3	Applications	747
---------------	---------------------	------------

17	Time series analysis and forecasting	748
18	Index numbers	799

CONTENTS

PREFACE	XII
GUIDE TO THE TEXT	XVI
GUIDE TO THE ONLINE RESOURCES	XX
ACKNOWLEDGEMENTS	XXII
ABOUT THE AUTHORS	XXIII

1 What is statistics?	1
Introduction to statistics	2
1.1 Key statistical concepts	5
1.2 Statistical applications in business	6
Case 3.6 Differing average weekly earnings of men and women in Australia	7
Case 4.2 Analysing the spread of the Global Coronavirus Pandemic	7
Case 5.5 Sydney and Melbourne lead the way in the growth in house prices	7
Case 14.1 Comparing salary offers for finance and marketing MBA majors - I	8
Case 16.1 Gold lotto	8
Case 17.3 Does unemployment affect inflation in New Zealand?	9
1.3 How managers use statistics	9
1.4 Statistics and the computer	11
1.5 Online resources	13
Appendix 1.A Introduction to Microsoft Excel	15
2 Types of data, data collection and sampling	18
Introduction	19
2.1 Types of data	20
2.2 Methods of collecting data	26
2.3 Sampling	30
2.4 Sampling plans	32
2.5 Sampling and non-sampling errors	39
Chapter summary	41

PART 1: DESCRIPTIVE MEASURES AND PROBABILITY 43

3 Graphical descriptive techniques – Nominal data	44
Introduction	45
3.1 Graphical techniques to describe nominal data	46
3.2 Describing the relationship between two nominal variables	68
Chapter summary	74
Case 3.1 Analysing the COVID-19 deaths in Australia by gender and age group	77
Case 3.2 Corporate tax rates around the world	77
Case 3.3 Trends in CO ₂ emissions	78
Case 3.4 Where is the divorce rate heading?	79

Case 3.5	Geographic location of share ownership in Australia	80
Case 3.6	Differing average weekly earnings of men and women in Australia	80
Case 3.7	The demography of Australia	81
Case 3.8	Survey of graduates	82
Case 3.9	Analysing the health effect of the Coronavirus pandemic	82
Case 3.10	Australian domestic and overseas student market by states and territories	82
Case 3.11	Road fatalities in Australia	83
Case 3.12	Drinking behaviour of Australians	84
4 Graphical descriptive techniques – Numerical data		85
Introduction		86
4.1	Graphical techniques to describe numerical data	86
4.2	Describing time-series data	106
4.3	Describing the relationship between two or more numerical variables	111
4.4	Graphical excellence and deception	123
Chapter summary		131
Case 4.1	The question of global warming	133
Case 4.2	Analysing the spread of the global coronavirus pandemic	134
Case 4.3	An analysis of telephone bills	134
Case 4.4	An analysis of monthly retail turnover in Australia	134
Case 4.5	Economic freedom and prosperity	134
5 Numerical descriptive measures		135
Introduction		136
5.1	Measures of central location	136
5.2	Measures of variability	153
5.3	Measures of relative standing and box plots	169
5.4	Measures of association	179
5.5	General guidelines on the exploration of data	193
Chapter summary		195
Case 5.1	Return to the global warming question	199
Case 5.2	Another return to the global warming question	199
Case 5.3	GDP versus consumption	199
Case 5.4	The gulf between the rich and the poor	199
Case 5.5	Sydney and Melbourne leading the way in the growth in house prices	200
Case 5.6	Performance of managed funds in Australia: 3-star, 4-star and 5-star rated funds	200
Case 5.7	Life in suburbs drives emissions higher	201
Case 5.8	Aussies and Kiwis are leading in education	202
Case 5.9	Growth in consumer prices and consumption in Australian states	202
Appendix 5.A	Summation notation	203
Appendix 5.B	Descriptive measures for grouped data	206
6 Probability		211
Introduction		212
6.1	Assigning probabilities to events	212
6.2	Joint, marginal and conditional probability	224
6.3	Rules of probability	234
6.4	Probability trees	239

6.5	Bayes' law	244
6.6	Identifying the correct method	251
	Chapter summary	252
Case 6.1	Let's make a deal	255
Case 6.2	University admissions in Australia: Does gender matter?	255
Case 6.3	Maternal serum screening test for Down syndrome	255
Case 6.4	Levels of disability among children in Australia	256
Case 6.5	Probability that at least two people in the same room have the same birthday	257
Case 6.6	Home ownership in Australia	257
Case 6.7	COVID-19 confirmed cases and deaths in Australia II	259
7	Random variables and discrete probability distributions	260
	Introduction	261
7.1	Random variables and probability distributions	261
7.2	Expected value and variance	269
7.3	Binomial distribution	275
7.4	Poisson distribution	284
7.5	Bivariate distributions	290
7.6	Applications in finance: Portfolio diversification and asset allocation	296
	Chapter summary	303
Case 7.1	Has there been a shift in the location of overseas-born population within Australia over the 50 years from 1996 to 2016?	306
Case 7.2	How about a carbon tax on motor vehicle ownership?	306
Case 7.3	How about a carbon tax on motor vehicle ownership? – New Zealand	307
Case 7.4	Internet usage by children	307
Case 7.5	COVID-19 deaths in Australia by age and gender III	308
8	Continuous probability distributions	309
	Introduction	310
8.1	Probability density functions	310
8.2	Uniform distribution	313
8.3	Normal distribution	316
8.4	Exponential distribution	336
	Chapter summary	341
Case 8.1	Average salary of popular business professions in Australia	343
Case 8.2	Fuel consumption of popular brands of motor vehicles	343
	Appendix 8.A Normal approximation to the binomial distribution	344
PART 2:	STATISTICAL INFERENCE	349
9	Statistical inference and sampling distributions	350
	Introduction	351
9.1	Data type and problem objective	351
9.2	Systematic approach to statistical inference: A summary	352
9.3	Introduction to sampling distribution	354
9.4	Sampling distribution of the sample mean \bar{X}	354
9.5	Sampling distribution of the sample proportion \hat{p}	366
9.6	From here to inference	369
	Chapter summary	371

10 Estimation: Single population	373
Introduction	374
10.1 Concepts of estimation	375
10.2 Estimating the population mean μ when the population variance σ^2 is known	378
10.3 Estimating the population mean μ when the population variance σ^2 is unknown	391
10.4 Estimating the population proportion p	403
10.5 Determining the required sample size	410
10.6 Applications in marketing: Market segmentation	417
Chapter summary	422
Case 10.1 Estimating the monthly average petrol price in Queensland	426
Case 10.2 Cold men and cold women will live longer!	426
Case 10.3 Super fund managers letting down retirees	427
Appendix 10.A Excel instructions for missing data and for recoding data	428
11 Estimation: Two populations	429
Introduction	430
11.1 Estimating the difference between two population means ($\mu_1 - \mu_2$) when the population variances are known: Independent samples	431
11.2 Estimating the difference between two population means ($\mu_1 - \mu_2$) when the population variances are unknown: Independent samples	439
11.3 Estimating the difference between two population means with matched pairs experiments: Dependent samples	449
11.4 Estimating the difference between two population proportions, $p_1 - p_2$	453
Chapter summary	461
Case 11.1 Has demand for print newspapers declined in Australia?	463
Case 11.2 Hotel room prices in Australia: Are they becoming cheaper?	463
Case 11.3 Comparing hotel room prices in New Zealand	464
Case 11.4 Comparing salary offers for finance and marketing major graduates	464
Case 11.5 Estimating the cost of a life saved	465
12 Hypothesis testing: Single population	466
Introduction	467
12.1 Concepts of hypothesis testing	467
12.2 Testing the population mean μ when the population variance σ^2 is known	476
12.3 The p -value of a test of hypothesis	491
12.4 Testing the population mean μ when the population variance σ^2 is unknown	504
12.5 Calculating the probability of a Type II error	510
12.6 Testing the population proportion p	517
Chapter summary	524
Case 12.1 Singapore Airlines has done it again	527
Case 12.2 Australian rate of real unemployment	527
Case 12.3 The republic debate: What Australians are thinking	527
Case 12.4 Has Australian Business Confidence improved since the May 2019 election?	528
Case 12.5 Is there a gender bias in the effect of COVID-19 infection?	528
Appendix 12.A Excel instructions	529
13 Hypothesis testing: Two populations	530
Introduction	531
13.1 Testing the difference between two population means: Independent samples	531

13.2	Testing the difference between two population means: Dependent samples - matched pairs experiment	551
13.3	Testing the difference between two population proportions	562
	Chapter summary	573
Case 13.1	Is there gender difference in spirits consumption?	578
Case 13.2	Consumer confidence in New Zealand	578
Case 13.3	New Zealand Government bond yields: Short term versus long term	579
Case 13.4	The price of petrol in Australia: Is it similar across regions?	579
Case 13.5	Student surrogates in market research	579
Case 13.6	Do expensive drugs save more lives?	580
Case 13.7	Comparing two designs of ergonomic desk: Part I	580
	Appendix 13.A Excel instructions: Manipulating data	581
14	Chi-squared tests	582
	Introduction	583
14.1	Chi-squared goodness-of-fit test	583
14.2	Chi-squared test of a contingency table	593
14.3	Chi-squared test for normality	608
14.4	Summary of tests on nominal data	614
	Chapter summary	616
Case 14.1	Gold lotto	620
Case 14.2	Exit polls	620
Case 14.3	How well is the Australian Government managing the coronavirus pandemic?	620
	Appendix 14.A Chi-squared distribution	622
15	Simple linear regression and correlation	624
	Introduction	625
15.1	Model	626
15.2	Estimating the coefficients	628
15.3	Error variable: Required conditions	641
15.4	Assessing the model	643
15.5	Using the regression equation	659
15.6	Testing the coefficient of correlation	664
15.7	Regression diagnostics - I	667
	Chapter summary	677
Case 15.1	Does unemployment rate affect weekly earnings in New Zealand?	683
Case 15.2	Tourism vs tax revenue	683
Case 15.3	Does unemployment affect inflation in New Zealand?	683
Case 15.4	Does domestic market capital influence stock prices?	683
Case 15.5	Book sales vs free examination copies	683
Case 15.6	Does increasing per capita income lead to increase in energy consumption?	684
Case 15.7	Market model of share returns	684
Case 15.8	Life insurance policies	685
Case 15.9	Education and income: How are they related?	685
Case 15.10	Male and female unemployment rates in New Zealand – Are they related?	685

16	Multiple regression	686
Introduction		687
16.1	Model and required conditions	687
16.2	Estimating the coefficients and assessing the model	688
16.3	Regression diagnostics - II	714
16.4	Regression diagnostics - III (time series)	726
Chapter summary		736
Case 16.1	Are lotteries a tax on the poor and uneducated?	741
Case 16.2	Demand for beer in Australia	741
Case 16.3	Book sales vs free examination copies revisited	741
Case 16.4	Average hourly earnings in New Zealand	742
Case 16.5	Testing a more effective device to keep arteries open	742
Appendix 16.A	<i>F</i> -distribution	743
PART 3: APPLICATIONS		747
17	Time series analysis and forecasting	748
Introduction		749
17.1	Components of a time series	749
17.2	Smoothing techniques	753
17.3	Trend analysis	763
17.4	Measuring the cyclical effect	768
17.5	Measuring the seasonal effect	772
17.6	Introduction to forecasting	780
17.7	Time series forecasting with exponential smoothing	783
17.8	Time series forecasting with regression	785
Chapter summary		796
Case 17.1	Part-time employed females	798
Case 17.2	New Zealand tourism: Tourist arrivals	798
Case 17.3	Seasonal and cyclical effects in number of houses constructed in Queensland	798
Case 17.4	Measuring the cyclical effect on Woolworths' stock prices	798
18	Index numbers	799
Introduction		800
18.1	Constructing unweighted index numbers	801
18.2	Constructing weighted index numbers	808
18.3	The Australian Consumer Price Index (CPI)	812
18.4	Using the CPI to deflate wages and GDP	816
18.5	Changing the base period of an index number series	821
Chapter summary		824
Case 18.1	Soaring petrol prices in Australian capital cities	828
Case 18.2	Is the Australian road toll on the increase again?	828
APPENDIX A:	SUMMARY SOLUTIONS FOR SELECTED (EVEN-NUMBERED) EXERCISES	829
APPENDIX B:	STATISTICAL TABLES	843
GLOSSARY		864
INDEX		869

PREFACE

Managing a business is a very complex responsibility and requires effective management to succeed. Managing complexity requires many skills. There are more competitors, more places to sell products and more places to locate workers. As a consequence, effective decision making is more crucial than ever before. On the other hand, nowadays managers have more access to larger and more detailed data sets that are potential sources of information for making well-informed objective decisions. Business managers are increasingly using statistical techniques to convert data into meaningful information. However, to achieve this, potential managers need to know which statistical techniques they should use to extract useful information from the available data to make informed decisions. For students preparing for the business world, it is not enough to focus merely on mastering a diverse set of statistical techniques and calculations. A course and its recommended textbook must provide a complete picture of statistical concepts and their applications to the real world. *Business Statistics, Abridged – Australia and New Zealand* is designed to demonstrate that statistical methods are vital tools for today's businesses and business managers to improve their decision-making skills.

This book is a thorough Australasian adaptation of the most popular and best-selling US text, *Statistics for Management and Economics* (11th edition) by Gerald Keller. This edition is a further attempt to make the basic business and economics statistics subject a more effective and enjoyable learning experience for both instructors and students at Australasian universities. It uses familiar local terminology, together with examples, exercises and cases that draw upon Australasian data. To enhance flexibility, we have also rearranged a number of chapters from the US edition. For example, we have incorporated the data collection chapter with types of data at the start of the book, additional graphical techniques such as bubble chart and heat maps in the graphical methods chapter, introduce estimation and hypothesis testing in separate chapters, present inference about population variance in another chapter and single population and two or more populations in different chapters. Furthermore, we have included a chapter on index numbers, which includes some important topics such as the construction of the Australian Consumer Price Index, as well as comparison of Laspeyres and Paasche index numbers.

When solving problems, *Business Statistics, Abridged – Australia and New Zealand* uses its unique 'ICI' approach, which is renowned for its consistent, proven three-step method to solving problems. The ICI approach teaches you how to: (1) *Identify* the appropriate technique, (2) *Compute* the statistics and (3) *Interpret* the results, in the context of the problem at hand. The *compute* stage can be completed in any or all of three ways: manually (with the aid of a calculator) or using Excel or XLSTAT (on the computer). This book contains step-by-step instructions and commands to teach students how to use Microsoft Excel® or XLSTAT to solve statistical problems. For those courses that wish to use the computer extensively, manual calculations can be played down or omitted completely. Conversely, those that wish to emphasise manual calculations may easily do so, and the computer solutions can be selectively introduced or skipped entirely. This approach is designed to provide maximum flexibility, and it leaves to the instructor the decision of when to introduce the computer.

Additionally, most examples, exercises and cases feature raw data. These data sets are available to download from the companion website accessible through <https://login.cengagebrain.com/>.

Key features of our approach

1. Systematic approach

This edition retains the systematic approach introduced in the US edition, which teaches students how to recognise which statistical technique to use. We believe that this skill is the most important one to develop, yet it is the one students have the greatest difficulty in mastering. As each technique is introduced, we demonstrate how to recognise when its use is appropriate and when it is not. Our ICI approach divides the solution of statistical problems into three parts: (1) identify the technique; (2) calculate/compute the required sample statistics; and (3) interpret the results. Our focus has been on the first and third parts, as the sample statistics could be produced relatively easily with a computer.

When demonstrating examples, we start the solutions by reviewing the appropriateness of the method to be used. One of the main benefits of our approach is that it allows instructors to de-emphasise mathematical manipulation. Consequently, students can spend more time properly setting up the procedure and interpreting the statistical results, and less time grinding out the arithmetic.

For students without access to a computer and statistical software, we continue to teach how to calculate statistics manually (with the exception of the most complicated procedures), and most exercises can be solved in this way.

2. Cases

Recent academic conferences devoted to improving the teaching of applied statistics have advocated the use of cases to help motivate students. In practice, a statistician often has access only to raw data and the correct procedure to employ is not obvious; our approach allows us to offer more realistic applications. In fact, many of the cases are based on real studies that have been reported in newspapers, magazines, journals, on television and at academic conferences. Several from our own consulting projects have also been included. Such applications can motivate students, who unfortunately often believe that statistics is not very relevant to their future careers. We believe that our approach can change these attitudes. More than 80 cases are included in the book. Students are expected to analyse the cases and draw conclusions in the same way as the original authors did. These cases are neither summaries of what a particular statistician did to solve a problem, nor glorified exercises; rather, they give students the opportunity to see for themselves how statistical problem solving works.

3. Review chapter

The review chapter is included in the book to help students practise identifying the correct techniques. This chapter reviews all the statistical methods covered in the book and provides exercises and cases that require the use of several different statistical procedures. It, therefore, provides practice in the technique identification skills that are required for statistics exams and, ultimately, in any real-life application of statistics.

4. Use of Excel

Because the use of spreadsheets is so widespread, we believe that Microsoft® Excel is an important addition to this book. However, spreadsheets are not designed for use as statistical

software, although they are increasingly capable in data analysis. Because of this limitation, we offer workbooks that can be used to solve statistical problems beyond Excel's existing capabilities. In this edition we have introduced another Excel Add-in, *XLSTAT*, which is capable of sophisticated statistical analysis to complement Excel's menu of statistical procedures. All statistical techniques introduced in this book can be calculated using either Excel's *Data Analysis (Analysis ToolPak)* or *XLSTAT*, which is available to download from the companion website accessible through <https://login.cengagebrain.com/>. The website also contains another Excel add-in *Data Analysis Plus*, which can also be used to perform a number of statistical calculations that Excel *Data Analysis* is unable to perform.

The Excel spreadsheet package is used extensively and presented consistently throughout the book to calculate sample statistics. Most examples in the chapters present manual (*Calculating manually*) and computer (*Using the computer*) solutions, allowing students to see both methods together and to use the preferred method. This feature provides flexibility, allowing the instructor to decide when manual or (Excel) computer calculations should be emphasised. Detailed instructions and Excel commands provided for the examples make it easy for instructors and students to make use of the computer. They also eliminate the need for instructors to teach how to use the software.

Data files are provided in Excel format for most of the examples, exercises and cases. The eighth edition includes hundreds of data files, some consisting of hundreds of observations, which emphasise a central theme in the book – statistical techniques convert data into information. For students who will conduct statistical analyses manually, we have also provided the summary statistics (e.g. means and variances) for exercises, allowing most exercises to be solved manually.

5. Exercises

There are over 1500 exercises of varying levels of difficulty. At the end of most sections we supply, under the heading '*Learning the techniques*', exercises that help students to learn the arithmetic involved in a specific procedure. '*Applying the techniques*' exercises then stress when and why the technique is used and how the results assist in the decision-making process. '*Computer applications*' help students gain hands-on experience in applying the techniques to solve problems using real-world data and computer software. Supplementary exercises appear at the end of each chapter. As they cover all the topics presented in that chapter, they allow students to practise identifying which of the techniques encountered in that chapter should be employed. They also tend to be more realistic than the other types of exercises.

We are optimistic that the systematic approach used in this book will be successful in helping students to understand how, when and why statistics are used. We hope that the realistic examples, exercises and cases we present, wherever possible with Australasian data, will make the subject more interesting and will persuade students that statistics can play a vital role in managerial decision making.

This text is suitable for a one- or two-semester subject in a business program. Although various sections can be omitted, we strongly urge instructors to attempt to complete most of the statistical inference part of the book. Like a house under construction, the structure of the systematic approach is stronger when most of the various components are in place. Nonetheless, the book has been designed so that chapters can be omitted relatively easily.

Unique features

- *Chapter opening examples* illustrate the use of techniques introduced in that chapter. These examples are designed to help students learn the concepts in the chapters. These chapter opening examples are revisited at the relevant section of the chapter, where they are solved.

- In addition to the examples provided in each chapter, we have included 'Real-life applications' sections which illustrate the fundamental applications of statistics in finance, marketing, human resources management, operations management, accounting and economics.
 - For example, to illustrate graphical techniques, we use an example that compares the histograms of the returns on two different investments. To explain what financial analysts look for in the histograms requires an understanding that risk is measured by the amount of variation in the returns. The example is preceded by a 'Real-Life Applications' box that discusses how return on investment is computed and used.
 - Later when we present the normal distribution, we feature another 'Real-Life Applications' box to show why the standard deviation of the returns measures the risk of that investment.
 - Several 'Real-Life Application' boxes are scattered throughout the book.
- 'In Summary' boxes are included after each technique has been introduced. These boxes will allow students to see a technique's essential requirements, in addition to giving them a way to easily review their understanding. This is further enhanced by '*Chapter Summary*' at the end of each chapter, which also include individual summary flowcharts in most of the Statistical Inference chapters and an overall summary flowchart in the review chapter.
- Several new exercises are added to each chapter.
- In addition to updating the data in the Examples and Exercises as much as possible, several new data sets have been added to the existing computer exercises section of each chapter. For those students who wish to solve the computer exercises containing data sets manually, summary statistics to these data sets are provided within each exercise.
- A more sophisticated commercially available Excel Add-in XLSTAT has been incorporated, to enable you to use Excel for almost all statistical procedures introduced in this book.
- In addition to Excel Data Analysis and XLSTAT commands, we have included several Excel workbooks that feature worksheets for confidence interval estimators and test statistics. By changing one or more inputs, students can learn, for example, the effect of increasing sample sizes on confidence intervals or on test statistics.
- Appendix A: Summary solutions for selected exercises is available in the book.

Guide to the text

As you read this text you will find a number of features in every chapter to enhance your study of Business Statistics and help you understand how the theory is applied in the real world.

PART-OPENING FEATURES

Part outlines show the chapters that are included in each part.

PART ONE

Descriptive measures and probability

CHAPTER 3	Graphical descriptive techniques – Nominal data
CHAPTER 4	Graphical descriptive techniques – Numerical data
CHAPTER 5	Numerical descriptive measures
CHAPTER 6	Probability
CHAPTER 7	Random variables and discrete probability distributions
CHAPTER 8	Continuous probability distributions

To help you organise the material that you are about to learn, we have divided the rest of the book into three parts.

Part 1 covers descriptive statistics and probability. These topics constitute the foundation of statistical inference. Chapter 3 introduces the graphical techniques for nominal data and Chapter 4 deals with graphical techniques for numerical data. Chapter 5 presents numerical measures that are used to summarise data. The summary measures introduced in Chapter 5 will be used to make inferences about parameters in later chapters. In Chapters 6 to 8, we present probability and probability distributions that will provide the link between sample statistics and population parameters.

Everything we do in this book is mostly built upon these six chapters. However, Part 1 does much more than just lay the foundation. Both descriptive statistics and probability are subjects that are worth learning for their own intrinsic values.

We all make decisions on a daily basis, most of which are made under uncertainty. Consider an investor who must decide which investment to make, how much money to invest and for how long that investment should be held. There are a large number of events over which the investor has no control. All that the investor can do is attempt to assess the risks and returns associated with each investment. As you will discover, probability plays a central role in this assessment.

We believe that all business and economics graduates will have many opportunities to apply statistical inference techniques and concepts. However, not all of them will do so because of a lack of either knowledge (despite the best efforts of statistics lecturers) or confidence. Descriptive techniques are so common that it is virtually impossible to ignore them. Newspapers, magazines, company annual reports and presentations are filled with applications of descriptive statistics. Knowing how to use and interpret them is a critical skill for all of us.

43

Part-opening paragraphs introduce the chapters in each part to give you an overview of how the chapters relate to each other.

CHAPTER-OPENING FEATURES

Identify the key concepts that the chapter will cover with the **learning objectives** and get a clear sense of what you should be able to do after reading the chapter.

3

Graphical descriptive techniques – Nominal data

Learning objectives

This chapter discusses the graphical descriptive methods used to summarise and describe sets of nominal data.

At the completion of this chapter, you should be able to:

- L01 construct charts to summarise nominal data
- L02 use Excel to draw appropriate charts for nominal data
- L03 determine which chart is best for nominal data under a given circumstance
- L04 use charts to describe ordinal data
- L05 use various tabular and graphical techniques to analyse the relationships between two nominal variables.

CHAPTER OUTLINE

- Introduction
- 3.1 Graphical techniques to describe nominal data
- 3.2 Describing the relationship between two nominal variables

SPOTLIGHT ON STATISTICS

Break bail, go to jail?

XMAS-08 An overwhelming majority of Victorian electors (78%) say people charged with a criminal offence who are given bail and then break a bail condition should be immediately sent to jail, according to a special SMS Morgan Poll conducted on the eve of the last Victorian state election.

Victorian electors were asked: 'Many people charged with a criminal offence are given bail. If a person given bail then breaks a bail condition, should that person be immediately sent to jail or not?' This special SMS Morgan Poll was conducted on Thursday 22 November 2018 with a statewide cross-section of 961 Victorian electors aged 18 and over. The responses and party affiliation for a random sample of 200 respondents are stored in file **XMAS-08**. Some of the data are listed below. Determine whether the responses differ on the basis of party affiliation. On pages 71–72 we provide a possible answer.'

Source: www.roymorgan.com.au. Finding no: 7812

ID	Party	Response
1	LNP	Yes
2	LNP	Yes
3	LNP	Yes
4	Others	Yes
...
199	Others	No
200	ALP	Yes



Source: ©The Australian

44

Spotlight on Statistics are used at the beginning of every chapter. They highlight a specific problem that can be solved by using the statistical techniques that will be covered in the chapter. The problem is answered later in the chapter.

FEATURES WITHIN CHAPTERS

Examples are used throughout each chapter. These are designed to teach you to use the authors' unique three-step approach to problem solving, and to help you apply statistics to real business problems.

You will learn to:

- **identify** the right statistical technique to use by focusing on the relationship between the problem and data type and **calculate manually**.
- **compute** the answer either by hand calculation, or by calculating it in Microsoft Excel® when you see a 'Using the computer' heading with instructions and 'Command' boxes, which include step-by-step instructions on how to complete the examples using Microsoft Excel®.
- **interpret** the answer in the context of the problem.

EXAMPLE 11.2

LO7

Comparing the clothing purchasing habits of men and women

The manager of a major clothing manufacturer wants to compare the annual expenditure on clothing of men and women. She decided to estimate the difference in mean annual expenditure to within \$100 with 95% confidence. How large should the two sample sizes be, if we assume that the range of expenditure is \$800 for males and \$1200 for females and the populations of male and female expenditures on clothing are normal?

Solution

Identifying the technique

The problem objective is to determine the required sample sizes. The data type is numerical. Both populations of interest are normally distributed and the samples are independent.

Calculating manually

The error bound is \$100 and the confidence level is 0.95. Hence, $B = \$100$ and $z_{0.025} = z_{0.975} = 1.96$.

We approximate σ_1 and σ_2 by using the following formulas:

$$\sigma_1 = \frac{\text{Range}}{4} = \frac{800}{4} = \$200 \text{ (for males)}$$

$$\sigma_2 = \frac{\text{Range}}{4} = \frac{1200}{4} = \$300 \text{ (for females)}$$

Thus,

$$n_1 = n_2 = \left\lceil \frac{1.96^2(200^2 + 300^2)}{100} \right\rceil = 49.94 \approx 50$$

Interpreting the results

In order to estimate $\mu_1 - \mu_2$ to within \$100 with 95% confidence, we should take samples of 50 men and 50 women.

Using the computer

The required sample sizes can be calculated using the **Sample size-2Means** worksheet in the **Estimators** workbook. The output is shown below.

Excel output for Example 11.2

A	B	C
1 Calculating the sample sizes for a fixed width (2B)		
2	Sample 1	Sample 2
3 Variance	40000	90000
4 Width/2 = B	100	
5 Confidence level	0.95	
6 Sample size	50	50

COMMANDS

To estimate required sample sizes, open the **Sample size-2Means** worksheet in the **Estimators** workbook, then type in the values of the given population variances (or as calculated above), B and the confidence level.

Definition and **Formula** boxes enable you to review your understanding of each new concept.

In Summary boxes are included after each technique has been introduced, as well as at the end of sections, to allow you to appreciate a technique's essential requirements and to enable you to review your understanding of each technique.

Geometric mean

The geometric mean of a population of N observations x_1, x_2, \dots, x_N is defined as:

$$\text{Population geometric mean: } \mu_g = \sqrt[N]{x_1 x_2 \dots x_N} = (x_1 x_2 \dots x_N)^{1/N}$$

The geometric mean of a sample of n observations x_1, x_2, \dots, x_n is defined as:

$$\text{Sample geometric mean: } \bar{x}_g = \sqrt[n]{x_1 x_2 \dots x_n} = (x_1 x_2 \dots x_n)^{1/n}$$

IN SUMMARY

Factors that identify when to calculate the range, variance, standard deviation and coefficient of variation

- 1 **Objective:** to describe a single set of data
- 2 **Type of data:** numerical
- 3 **Descriptive measurement:** variability

FEATURES WITHIN CHAPTERS

The text contains over 1500 **Exercises**, located at the end of each section in the chapters.

These include:

- **Learning the techniques** exercises help you to learn the arithmetic involved in a specific procedure
- **Applying the technique** exercises highlight when and why the techniques are used and how the results assist in the decision-making process
- **Computer applications** help you gain hands-on experience in applying the techniques to solve problems using real world data and Microsoft Excel®.

EXERCISES

Learning the techniques

- 3.29 XR03-29** The following table summarises the data from a survey on the ownership of iPads for families with different levels of income ($C_1 < C_2 < C_3$). Determine whether the two nominal variables are related.

Ownership	Level of income		
	C_1	C_2	C_3
No	40	32	48
Yes	30	48	52

Applying the techniques

- 3.30 XR03-30 Self-correcting exercise.** The trustee of a company's superannuation scheme has solicited the opinions of a sample of the company's employees regarding a proposed revision of the scheme. A breakdown of the responses is shown in the following table. Use an appropriate graphical presentation to determine whether the responses differ among the three groups of employees.

Responses	Blue-collar workers	White-collar workers	Managers
For	67	32	11
Against	63	18	9

Computer applications

- 3.31 XR03-31** The associate dean of a business school was looking for ways to improve the quality of applicants to its MBA program. In particular she wanted to know whether the undergraduate degrees of applicants to her school and the three nearby universities with MBA programs differed. She sampled 100 applicants of her program and an equal number from each of the other universities.

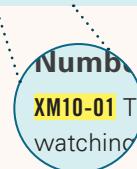
Data files, highlighted throughout the text, enable you to complete the Examples, Exercises and Case Studies in the text without having to spend time inputting raw data. These data files are available on the accompanying student companion website (<http://login.cengagebrain.com>), the MindTap platform or via your instructor.

EXAMPLE 10.1

LO4 LO5

Number of hours children spend watching television

XM10-01 The sponsors of television shows targeted at children wanted to know the amount of time children spend watching television, since the types and number of programs and commercials presented are greatly influenced by this information. As a result, a survey was conducted to estimate the average number of hours Australian children spend watching television per week. From past experience, it is known that the population standard deviation is 8.0 hours. The following are the data gathered from a sample of 100 children. Find the 95% confidence interval estimate of the average number of hours Australian children spend watching television.



Real-Life Applications are included throughout the text to demonstrate real-world applications of statistics in the areas of finance, marketing, human resource management, accounting and economics.

REAL-LIFE APPLICATIONS

Test marketing

In Chapter 13, we described test marketing, which is often used to assess consumer reaction to changes in one or more elements of the marketing mix. Marketing managers will conduct experiments to determine whether differences in sales exist between different prices for the product, different package designs or

different advertising strategies. Some of these experiments are carried out in small communities where it is easy to vary the particular elements that the manager wishes to investigate.

Source: Shutterstock.com/
AlenKadir



Commands boxes can be found throughout the text and include step-by-step instructions on how to complete exercises in Microsoft Excel®.

COMMANDS

- 1 Open the data file (**XM03-07**). Highlight the data including the column titles (**A1:C55**).
- 2 Click **INSERT** and **PivotTable** under the **Charts** submenu. Then select **Pivot Chart & Pivot Table**.
- 3 Make sure that the **Table/Range** is correct and click **OK**.
- 4 Drag the **Occupation** button from the menu that appears on the right of the screen to the **Drop Row Fields Here** section of the box. Drag the **Newspaper** button to the **Drop Column Fields Here** section. Drag the **Reader** button to the **Drop Value Fields Here** section. Right-click any number in the table, click **Summarize Values By**, and check **Count**. This will produce a contingency table for the counts (frequencies).
- 5 To convert to row percentages, right-click any number in the table, click **Summarize Values By**, **More options...**, and **Show values as**. Select **% of rows** from the drop-down menu and then click **OK**. Format the data into decimals by highlighting all cells with percentages and right-clicking. Select **Number Format...**, select **Number** under **Category** and select the number of **Decimal places**. Click **OK**. This will produce the contingency table for the row relative frequencies.

Definitions or explanations of important **key terms** are located in the margins for quick reference.

cross-classification table
A first step in graphing the relationship between two nominal variables.

To describe the relationship between two nominal variables, we must remember that we are only permitted to determine the frequency of the values. A variation of the bar chart introduced in Section 3.1 is used to describe the relationship between two nominal (categorical) variables in graphical form. As a first step, we need to produce a **cross-classification table** (also known as contingency table or cross-tabulation table), which lists the frequency of each combination of the values of the two variables. We will illustrate the use of graphs to describe the relationship between two nominal variables using data from the newspaper readership case in Example 3.7.

END-OF-CHAPTER FEATURES

At the end of each chapter you will find several tools to help you to review, practise and extend your knowledge of the key learning objectives including:

- A **summary** section that consolidates your knowledge of the content of the chapter by reviewing key concepts and drawing out their wider significance

- A recap of relevant **symbols**

- A **summary of formulas** from the chapter.

Study Tools

CHAPTER SUMMARY

The concept of a random variable permits us to summarise the results of an experiment in terms of numerically valued events. Specifically, a random variable assigns a numerical value to each simple event of an experiment. There are two types of random variables. A discrete random variable is one whose values are countable. A continuous random variable can assume an uncountable number of values. In this chapter we discussed discrete random variables and their probability distributions.

We defined the *expected value*, *variance* and *standard deviation* of a population represented by a discrete probability distribution. We also presented two most important discrete distributions: the *binomial* and the *Poisson*. Finally, in this chapter we also introduced *bivariate distributions* for which an important application in finance was discussed.

SYMBOL

Symbol	Pronounced	Represents
$\sum_{i=1}^n x_i$	Sum of x for all values of x	Summation
C_n^r	n -choose- x	Number of combinations
$n!$	n -factorial	$n(n-1)(n-2) \dots (3)(2)(1)$
e	exponential	2.718...

SUMMARY OF FORMULAS

Expected value (mean)	$E(X) = \mu = \sum_{x=1}^n xp(x)$
Variance	$V(X) = \sigma^2 = \sum_{x=1}^n (x - \mu)^2 p(x) = \sum_{x=1}^n x^2 p(x) - \mu^2$
Standard deviation	$SD(X) = \sigma = \sqrt{V(X)}$
Covariance	$COV(X, Y) = \sigma_{xy} = \sum_{x,y} (x - \mu_x)(y - \mu_y) p(x, y)$ $= \sum_{x,y} xy p(x, y) - \mu_x \mu_y$
Coefficient of correlation	$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

Supplementary Exercises at the end of each chapter give you the opportunity to further test your understanding of the key concepts covered.

SUPPLEMENTARY EXERCISES

9.33 The dean of a business school claims that the average Master of Business Management graduate is offered an annual starting salary of \$109 500. The standard deviation of the offers is \$9300. What is the probability that for a random sample of 38 Master of Business Management graduates, the mean starting annual salary is less than \$105 000?

9.34 Refer to Exercise 9.33. Suppose that a random sample of 38 Master of Business Management graduates report that their mean starting salary is \$105 000. What does this tell you about the dean's claim?

9.35 A restaurant in a large commercial building provides coffee for the building's occupants. The restaurateur has determined that the mean number of cups of coffee consumed in one day by all the occupants is 2.0, with a standard deviation of 0.6. A new tenant of the building intends to have a total of 125 new employees. What is the probability that the new employees will consume more than 240 cups of coffee per day?

9.36 The number of pages photocopied each day by the admin staff in a busy office is normally distributed with a mean of 550 and a standard deviation of 150. Determine the probability that in one business week (i.e. 5 days) more than 3000 pages will be copied.

9.37 A university bookstore claims that 50% of its

location they seek, rather than ask for directions. To examine this belief, he took a random sample of 350 male drivers and asked each what they did when lost. If the belief is true, determine the probability that less than 75% said they continue driving.

9.39 The Red Lobster restaurant chain regularly surveys its customers. On the basis of these surveys, management claims that 75% of customers rate the food as excellent. A consumer testing service wants to examine the claim by asking 460 customers to rate the food. What is the probability that less than 70% rate the food as excellent?

9.40 An accounting professor claims that no more than one-quarter of undergraduate business students will major in accounting.

a. What is the probability that in a random sample of 1200 undergraduate business students, 336 or more will major in accounting?

b. A survey of a random sample of 1200 undergraduate business students indicates that there are 336 students who plan to major in accounting. What does this tell you about the professor's claim?

9.41 Statisticians determined that the mortgages of homeowners in a city is normally distributed with a mean of \$500 000 and a standard deviation of \$100 000. A random sample of 100 homeowners

Case Studies are included at the end of each chapter to assist you in applying the statistical techniques you are learning to real-world problems.

Case Studies

CASE 8.1 Average salary of popular business professions in Australia

C8-01 Based on a recent publication (payscale.com/index/au/job), the average salaries of 10 popular business-related jobs in Australia are listed below. It is believed that salaries can be considered as following a normal distribution with a standard deviation of \$1500. A high school student would like to choose a degree that could lead to the profession which has a higher probability of gaining a reasonably good salary. What is the likelihood of him receiving an annual salary greater than \$60 000 for each of the 10 jobs listed in the table?

Average salary in Australia, 10 popular business jobs

Job title	Average salary (\$)
Accountant	57139
Business Data Analyst	69823
Finance Manager	93835
Financial Planner	76004
HR Manager	89328
Marketing Manager	78847
Personal Banker	53285
Retail Manager	51413
Supply Chain Manager	103724
Tax Accountant	56316

Source: payscale.com/index/au/job

Appendices throughout the text included step-by-step instructions on how to perform complex statistical calculations.

Appendix 5.A

Summation notation

This appendix offers an introduction to the use of summation notation. Because summation notation is used extensively throughout statistics, you should review this appendix even if you have had previous exposure to summation notation. Our coverage of the topic begins with an introduction to the necessary terminology and notation, follows with some examples, and concludes with four rules that are useful in applying summation notation.

Consider n numbers x_1, x_2, \dots, x_n . A concise way of representing their sum is:

$$\sum_{i=1}^n x_i$$

That is:

$$\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n$$

Terminology and notation

1 The symbol Σ is the capital Greek letter sigma, and means 'the sum of'.

2 The letter i is called the *index of summation*. The letter chosen to represent the index of summation is arbitrary.

3 The expression $\sum_{i=1}^n x_i$ is read as 'the sum of the terms x_i where i assumes the values from 1 to n inclusive'.

Guide to the online resources

FOR THE INSTRUCTOR

Cengage Learning is pleased to provide you with a selection of resources that will help you prepare your lectures and assessments. These teaching tools are accessible via cengage.com.au/instructors for Australia or cengage.co.nz/instructors for New Zealand.

MINDTAP

Premium online teaching and learning tools are available on the *MindTap* platform – the personalised eLearning solution.

MindTap is a flexible and easy-to-use platform that helps build student confidence and gives you a clear picture of their progress. We partner with you to ease the transition to digital – we're with you every step of the way.

The *Cengage Mobile App* puts your course directly into students' hands with course materials available on their smartphone or tablet. Students can read on the go, complete practice quizzes or participate in interactive real-time activities.

MindTap for Selvanathan's Business Statistics is full of innovative resources to support critical thinking, and help your students move from memorisation to mastery! Includes:

- Selvanathan's Business Statistics eBook
- Data sets for Examples, Exercises and Cases
- Downloadable Workbooks to perform some of the statistical estimations and tests discussed in the book
- Videos and animations
- Selected solutions for the self-correcting exercises
- Revision quizzes
- Interactive assignment quizzes
- Aplia problem sets
- Interactive flashcards to revise key terms.

MindTap is a premium purchasable eLearning tool. Contact your Cengage learning consultant to find out how *MindTap* can transform your course.



SOLUTIONS MANUAL

The **Solutions Manual** includes solutions for all the exercises and supplementary exercises in each chapter.

WORD-BASED TEST BANK

This bank of questions has been developed with the text for the creation of quizzes, tests and exams for your students. Deliver tests from your LMS and your classroom.

POWERPOINT™ PRESENTATIONS

Use the chapter-by-chapter **PowerPoint** presentations to enhance your lecture presentations and handouts to reinforce the key principles of your subject.

ARTWORK FROM THE TEXT

Add the digital files of graphs, pictures and flow charts into your course management system, use them in student handouts, or copy them in your lecture presentations

FOR THE STUDENT

This book is accompanied by a companion website that can be accessed via <https://login.cengagebrain.com>, which contains the data sets for Examples, Exercises and Cases.

MINDTAP

A new approach to highly personalised online learning, **MindTap** is designed to match your learning style and provides you with an engaging interface that allows you to interact with the course content and multimedia resources as well as with your peers, lecturers and tutors. In the **MindTap** Reader you can make notes, highlight text and even find a definition directly from the page.

To purchase your **MindTap** experience for **Business Statistics**, please contact your instructor.

MindTap also includes:

- Data sets for examples, exercises and cases
- Workbooks for examples, exercises and cases
- Excel estimator workbooks
- Animations plus assistance documentation
- Solutions to selected self-correcting exercises
- Data Analysis Plus with assistance documentation
- Revision quizzes.



ACKNOWLEDGEMENTS

We are grateful to the publishing, editorial and production teams at Cengage Learning for their help and support. We are particularly grateful to Cengage Learning's Senior Publishing Editor, Geoff Howard, for his patience, guidance and encouragement; to Emily Spurr for developmental work; to Marta Veroni for her thorough editing work; and to Nathan Katz, whose firm hand on the editorial process kept us on task and whose production process converted our manuscript into the polished final product. Their advice and suggestions have made our task much easier.

We would like to thank the many colleagues, survey participants, reviewers and students for their helpful suggestions, comments and criticisms on various chapters of the book, which have contributed to improving this edition and the previous editions. In particular, we would like to thank Professor Prasada Rao and Dr Mohammed Alauddin, University of Queensland; Professor Ken Clements, University of Western Australia; Professor Bill Griffiths, Melbourne University; Dr Patti Cybinski, Associate Professor Helen Higgs, Dr Suri Rajapakse, Dr Lucille Wong, Dr Tommy Soesmanto, Dr Ripon Mondal, Alexander Gardental and Danny Williams, Griffith University; Andrew Paulridge, Queensland University of Technology; Dr Christine Williams, Assistant Director General – Science Division, Queensland Government; Professor Eric Sowey, University of New South Wales; Associate Professor Brett Inder, Dr Ann Maharaj, Dr Mindi Nath, John Betts, Bruce Stephens and Lesley Tissera, Monash University; Associate Professor John MacFarlane, Dr Kevin Donegan, Dr Than Pe and Neil Hopkins, University of Western Sydney; Professor Alan Woodland, Tig Ihnatko and John Goodhew, University of Sydney; Dr Geoff Coates, Anne Arnold and Margaret Meyler, University of Adelaide; Dr G.V. Crockett and Dr Fay Rola-Rubzen, Curtin University of Technology; Damon Chernoff, Edith Cowan University; Dr Maneka Jayasinghe, Charles Darwin University; Dr Eddie Oczkowski, Kerrie Cullis, William Collen and Sue Moffatt, Charles Stuart University; Dr Jim Bates, Victoria University; Joy Ross and Lisa Yong, RMIT; Iain Fraser and Suzanne Sommer, La Trobe University; Dr S. Ganeshalingam, Massey University, New Zealand; Professor Christine Lim, Waikato University, New Zealand; Professor Brinda Viswanathan, Madras School of Economics, India; Dr N Ravinthirakumaran, University of Colombo; and Dr C. Elankumaran and Dr S. Kalamani, University of Jaffna, Sri Lanka.

Cengage and the authors would like to thank the following reviewers for their incisive and helpful feedback:

- Boris Choy (The University of Sydney)
- Joe Hirschberg (University of Melbourne)
- Diane Morien (Curtin University)
- Vivian Piovesan (University of Adelaide)
- Shahadat Uddin (University of Sydney).

Every effort has been made to trace and acknowledge copyright. However, if any infringement has occurred, the publishers tender their apologies and invite the copyright holders to contact them.

ABOUT THE AUTHORS

Professor Eliyathamby A Selvanathan

Eliyathamby 'Selva' Selvanathan is a Professor in Econometrics and Director of the Economics Policy Analysis Program (EPAP) at Griffith University, Queensland, Australia. He is also a Visiting Professor at the Madras School of Economics, Anna University, Chennai, India. Selva has also taught previously at the University of Jaffna, Murdoch University, The University of Western Australia and University of Queensland. He has held positions such as the Deputy Dean (Staffing) – Faculty of International Business, Director of Bachelor of International Business Program and Deputy Director of the Statistics and Research Design (STARDS) Unit at Griffith University. Selva was educated at the University of Jaffna, University of Bucharest, Murdoch University and The University of Western Australia. He is the recipient of several individual and group Excellence in Teaching Awards, such as *75th Anniversary Distinguished Teaching Award, The University of Western Australia, 1988, 2006 Minister's Awards for Outstanding Contribution to Improving Literacy and/or Numeracy, Queensland state winner, Griffith Awards for Excellence in Teaching 2005 – Innovation Across the Institution, Pro-Vice Chancellor's Research Excellence Award 2014 – Research Supervision*, and has received numerous competitive teaching and national research grants. Selva has published six research monographs and has published widely in international refereed journals such as *Journal of Econometrics, Review of Economics and Statistics, Journal of Business and Economic Statistics, Review of Economic Studies, Energy Economics, Economic Letters, Tourism Analysis, Tourism Analysis and Marketing Science*, several chapters in books, and several editions of two textbooks.

Professor Saroja Selvanathan

Dr Saroja Selvanathan is Professor in Econometrics at Griffith University and a Visiting Professor at the Madras School of Economics, Anna University, Chennai, India. She has held positions such as such as Deputy Dean (Research and Postgraduate Studies), Acting Dean, Deputy Head of Department, Head of Discipline, Higher Degree Research Convenor and the Director of the Statistics and Research Design Support Unit at Griffith University. Saroja was educated at the University of Jaffna, Murdoch University and The University of Western Australia, and has published several research monographs and a number of research papers in international refereed journals such as the *Review of Economics and Statistics, Transportation Research, Economics Letters, Empirical Economics, Economic Modelling, Tourism Economics and Applied Economics*. She has received several awards, including *Vice Chancellor's Research Excellence Award 2019 – Higher Degree Research Supervision, Pro-Vice Chancellor's Research Excellence Award 2018 – Research Supervision, Griffith Awards for Excellence in Teaching 2011 – Group Learning and Teaching Citation, and Griffith Awards for Excellence in Teaching 2005 – Innovation Across the Institution*.

Professor Gerald Keller

Dr Gerald Keller is Emeritus Professor of Business at Wilfred Laurier University, where he has taught statistics, management science and operations management since 1974. He has also taught at the

University of Toronto, the University of Miami, McMaster University, the University of Windsor and the Beijing Institute of Science and Technology. Gerald has consulted with banks on credit scoring and credit card fraud and has conducted market surveys for the Canadian government on energy conservation. The author of *Applied Statistics with Microsoft Excel*, *BSTAT*, *Essentials of Business Statistics (co-authored)*, *Australian Business Statistics (co-authored)*, and *Statistics Laboratory Manual Experiments Using Minitab*, Gerald has also been published in *Omega*, *IIE Transactions*, *Decision Sciences*, *INFOR*, *Economics Letters* and *Archives of Surgery*.

What is statistics?

Learning objectives

This chapter provides an introduction to the two general bodies of methods that together constitute the subject called statistics: descriptive statistics and inferential statistics.

At the completion of this chapter, you should be able to:

- L01** describe the two major branches of statistics – descriptive statistics and inferential statistics
- L02** understand the key statistical concepts – population, sample, parameter, statistic and census
- L03** provide examples of practical applications in which statistics have a major role to play and understand how statistics are used by business managers
- L04** understand the basics of the computer spreadsheet package Microsoft Excel and its capabilities in aiding with statistical data analysis for large amounts of data.

CHAPTER OUTLINE

Introduction to statistics

1.1 Key statistical concepts

1.2 Statistical applications in business

1.3 How managers use statistics

1.4 Statistics and the computer

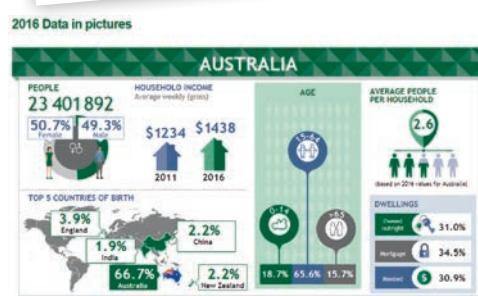
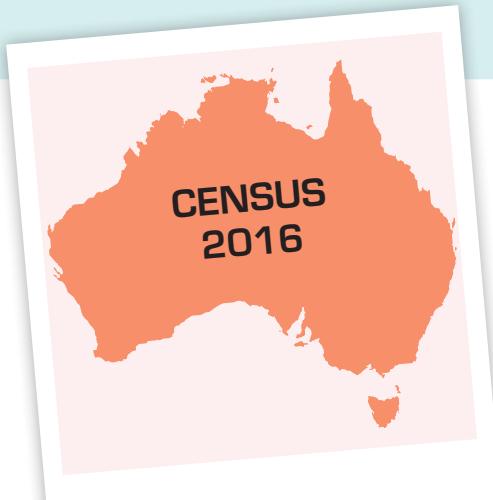
1.5 Online resources

SPOTLIGHT ON STATISTICS

Census in Australia

Information (or data) gathering on various characteristics of different populations of interest is an important part of statistics. When we use all units of the population to record information about the characteristics of interest, this type of data gathering is called a census. Due to cost and other resource implications, a census of the whole population is done once in a while in most countries. A census takes place only every five years in Australia. The peak Australian government statistics agency, the Australian Bureau of Statistics (ABS), carries out the census. The two most recent censuses were carried out on the nights of 9 August 2011 and 2016. A census provides a snapshot of the population characteristics in the year it is held. We will discuss census and other forms of data collection in detail in this and the next chapter.

Source: Australian Bureau of Statistics. © Commonwealth of Australia CC BY 4.0. <https://creativecommons.org/licenses/by/4.0/>



Introduction to statistics

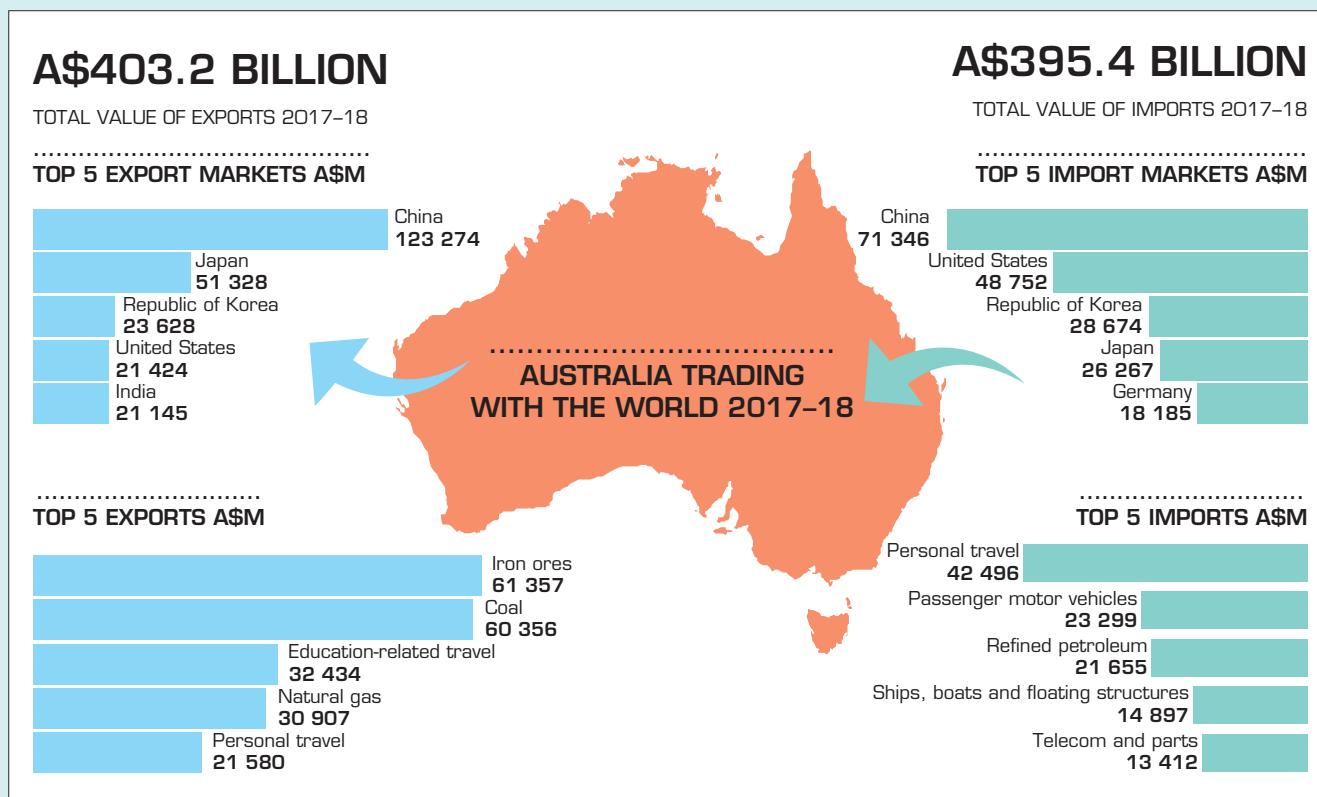
Statistics is a way to get information from data. That's it! Most of this textbook is devoted to describing how, when and why managers and statistics practitioners conduct statistical procedures. You might ask, 'If that's all there is to statistics, why is this book (and most other statistics books) so large?' The answer is that students of applied statistics will be exposed to different kinds of information and data. We demonstrate some of these with case studies and examples throughout this book.

Statistics is a body of principles and methods concerned with extracting useful information from a set of data to help people make decisions. In general, statistics can be subdivided into two basic areas: *descriptive statistics* and *inferential statistics*.

EXAMPLE 1.1

Australia's trade with the world

The total value of imports into Australia continues to exceed the total value of its exports. The Australian government has signed various trade agreements with our major trading partners to relax certain trade restrictions and increase our exports. Over the past decade, successive governments have pledged to reduce the trade deficit (= Imports – Exports) and have only been able to experience the turnaround recently. The picture below shows some useful information on Australian imports and exports in 2017–18 with its major trading partners. For example, the United States (US) is one of our major trading partners and our total imports from the US are about two and a half times our exports to the US. We will discuss various descriptive graphical and numerical techniques to summarise such data in the following chapters.



Source: Department of Foreign Affairs and Trade CC BY 3.0 AU <https://creativecommons.org/licenses/by/3.0/au/legalcode>

EXAMPLE 1.2

L01

Summarising the business statistics marks

A student enrolled in a business program is attending his first lecture of the compulsory business statistics course. The student is somewhat apprehensive because he believes the myth that the course is difficult. To alleviate his anxiety, the student asks the lecturer about the exam marks for last year's business statistics course. Because, like all statistics lecturers, this one is friendly and helpful, the lecturer obliges and provides a list of the final marks. The overall marks are compiled from all the within-semester assessment items plus the end-of-semester final exam. What information can the student obtain from the list?

This is a typical statistics problem. The student has the data (marks) and needs to apply statistical techniques to get the information he requires. This is a function of *descriptive statistics*.

Descriptive statistics

Descriptive statistics deals with methods of organising, summarising and presenting data in a convenient and informative form. One form of descriptive statistics uses graphical techniques that allow statistics practitioners to present data in ways that make it easy for the reader to extract useful information. The main attraction of a graphical presentation is that the message can be easily understood by any layperson. In Chapters 3 and 4 we will present a variety of graphical methods.

Another form of descriptive statistics uses numerical techniques to summarise data. One such method you would have already used frequently is calculating an average or mean. Chapter 5 introduces several numerical statistical measures that describe different features of the data.

The actual descriptive statistical technique we use depends on what specific information we would like to extract from a given data set. In Example 1.2, there are at least three important pieces of summary information. The first is the 'typical' mark. We call this a *measure of central location*. The average is one such measure we will introduce in Chapter 5. The student calculated the average mark of the business statistics course, from the list of marks provided by the lecturer, by summing all the marks and dividing the total by the number of students, which was 74. In Chapter 5 we will also introduce another useful measure of central location, the median. In the above example, the median is 81, which is the middle mark of the class when the marks are arranged in ascending or descending order. That is, 50% of the students obtained marks less than the median mark, while 50% received marks greater than the median value.

Now the student knows that the average mark was 74. Is this enough information to reduce his anxiety? The student would likely respond 'no' because he would like to know whether most of the marks were close to the average mark of 74 or were scattered far below and above the average. He needs a *measure of variability*. The simplest such measure is the *range* (discussed further in Chapter 5), which is calculated by subtracting the smallest mark from the largest. The student noticed that the largest mark is 100 and the smallest is 11, thus the range is $(100 - 11) = 89$. Unfortunately, this provides little information, as the range doesn't indicate where most of the marks are located. Whether most data are located near 11 or near 100 or somewhere in the middle, the range is still 89. He needs other measures of variability, such as the variance and standard deviation, to reflect the true picture of the spread of the data. These will be introduced in Chapter 5. Moreover, the student must determine more about the marks. In particular he needs to know how the marks are distributed between 11 and 100. The best way to do this is to use a graphical technique, the histogram, which is introduced in Chapter 4.

descriptive statistics

Methods of organising, summarising and presenting data in ways that are useful, attractive and informative to the reader.

EXAMPLE 1.3

L01 L03

Comparing weekly sales between two outlets

A fast-food franchisee wishes to compare the weekly sales level over the past year at two particular outlets. Descriptive statistical methods could be used to summarise the actual sales levels (perhaps broken down by food item) in terms of a few numerical measures, such as the average weekly sales level and the degree of variation from this average that weekly sales may undergo. Tables and charts could be used to enhance the presentation of the data so that a manager could quickly focus on the essential differences in sales performance at the two outlets.

There is much more to statistics, however, than these descriptive methods. Decision-makers are frequently forced to make decisions based on a set of data that is only a small subgroup (sample) of the total set of relevant data (population).

inferential statistics

Methods used to draw conclusions about a population based on information provided by a sample of the population.

Inferential statistics

Inferential statistics is a body of methods for drawing conclusions (i.e. making inferences) about characteristics of a population, based on data available in a sample taken from the population. The following example illustrates the basic concepts involved in inferential statistics.

EXAMPLE 1.4

L01 L02 L03 L04

Profitability of a new life insurance policy

An Australia-wide automobile club (consisting of about 2 million members) is contemplating extending its services to its members by introducing a new life insurance policy. After some careful financial analysis, the club has determined that the proposed insurance policy would break even if at least 10% of all current members subscribing to the club also purchase the policy. The question here is how can inferential statistics be used by the automobile club to make a decision about introducing their new life insurance policy?

To answer this question, we need to first obtain additional information before reaching a decision on whether or not to proceed with the new insurance policy, the automobile club has decided to conduct a survey of 500 randomly selected current members. The collection of all its current 2 million or so members is called the *population*. The 500 members selected from the entire population for the analysis are referred to as a *sample*. Each member in the sample is asked if they would purchase the policy if it were offered at some specified price. Suppose that 60 of the members in this sample reply positively. While a positive response by 60 out of 500 members (12%) is encouraging, it does not assure the automobile club that the proposed insurance policy will be profitable. The challenging question here is how to use the response from these 500 sampled members to conclude that at least 10% of all 2 million or so members would also respond positively. The data are the proportion of positive responses from the 500 members in the sample. However, we are not so much interested in the response of the 500 members as we are in knowing what the response would be from all of the club's 2 million current members. To accomplish this goal we need another branch of statistics – *inferential statistics*.

If the automobile club concludes, based on the sample information, that at least 10% of all its members in the population would purchase the proposed insurance policy, the club is relying on inferential statistics. The club is drawing a conclusion, or making a statistical inference, about the entire population of its 2 million or so members on the basis of information

provided by only a sample of 500 members taken from the population. The available data tell us that 12% of this particular sample of members would purchase the policy; the inference that at least 10% of all its members would purchase the new insurance policy may or may not be correct. It may be that, by chance, the club selected a particularly agreeable sample and that in fact no more than 5% of the entire population of members would purchase the new policy.

Whenever an inference is made about an entire population on the basis of evidence provided by a sample taken from the population, there is a chance of drawing an incorrect conclusion. Fortunately, other statistical methods allow us to determine the reliability of the statistical inference. They enable us to establish the degree of confidence we can place in the inference, assuming the sample has been properly chosen. These methods would enable the automobile club in Example 1.4 to determine, for example, the likelihood that less than 10% of the population of its members would purchase the policy, given that 12% of the members sampled said they would purchase. If this likelihood is deemed small enough, the automobile club would probably proceed with its new venture.

1.1 Key statistical concepts

Statistical inference problems involve three key concepts: the population, the sample and the statistical inference. We now discuss each of these concepts in more detail.

1.1a Population

A **population** is the group of all items of interest to a statistics practitioner. It is frequently very large and may, in fact, be infinitely large. In the language of statistics, the word ‘population’ does not necessarily refer to a group of people. It may, for example, refer to the population of diameters of ball bearings produced at a large plant. In Example 1.4, the population of interest consists of all 2 million or so members.

A descriptive measure of a population is called a **parameter**. The parameter of interest in Example 1.4 was the proportion of all members who would purchase the new policy.

population

The set of all items of interest.

parameter

A descriptive measure of a population.

1.1b Sample

A **sample** is a subset of data drawn from the target or studied population. In Example 1.4, the sample of interest consists of the 500 selected members.

A descriptive measure of a sample is called a **statistic**. We use sample statistics to make inferences about population parameters. In Example 1.4, the proportion (\hat{p}) of the 500 members who would purchase the life insurance policy would be a sample statistic that could be used to estimate the corresponding population parameter of interest, the population proportion (p) who would purchase the life insurance policy. Unlike a parameter, which is a constant, a statistic is a variable whose value varies from sample to sample. In Example 1.4, 12% is a value of the sample statistic based on the selected sample.

sample

A set of data drawn from the studied population.

statistic

A descriptive measure of a sample.

1.1c Statistical inference

Statistical inference is the process of making an estimate, prediction/forecast or decision about a population parameter based on the sample data. Because populations are usually very large, it is impractical and expensive to investigate or survey every member of a population. (Such a survey is called a census.) It is far cheaper and easier to take a sample from the population of interest and draw conclusions about the population parameters based on sample statistics. In Example 1.4, we make a conclusion about the population proportion, p , based on the sample proportion \hat{p} .

For instance, political pollsters predict, on the basis of a sample of about 1500 voters, how the entire 16 million eligible voters from the Australian population will cast their ballots; and quality-control supervisors estimate the proportion of defective units being produced in a massive production process from a sample of only several hundred units.

Because a statistical inference is based on a relatively small subgroup of a large population, statistical methods can never decide or estimate with certainty. As decisions involving large amounts of money often hinge on statistical inferences, the reliability of the inference is very important. As a result, each statistical technique includes a measure of the reliability of the inference. For example, if a political pollster predicts that a candidate will receive 40% of the vote, the measure of reliability might be that the true proportion (determined on election day) will be within 3% of the estimate on 95% of the occasions when such a prediction is made. For this reason, we build into the statistical inference a measure of reliability. There are two such measures: the confidence level and the significance level. The *confidence level* is the proportion of times that an estimating procedure would be correct if the sampling procedure was repeated a very large number of times. For example, a 95% confidence level would mean that for a very large number of repeated samples, estimates based on this form of statistical inference will be correct 95% of the time. When the purpose of the statistical inference is to draw a conclusion about a population, the *significance level* measures how frequently the conclusion will be wrong in the long run. For example, a 5% significance level means that in repeated samples this type of conclusion will be wrong 5% of the time. We will introduce these terms in Chapters 10 and 12. The outcome of the 2019 Australian federal election is an example of how conclusions or predictions can go wrong if the samples selected do not represent the target population.

1.2 Statistical applications in business

Throughout the text, you will find examples, exercises and cases that describe actual situations from the business world in which statistical procedures have been used to help make decisions. For each example, exercise or case, you will be asked to choose and apply the appropriate statistical technique to the given data and to reach a conclusion. We cover such applications in accounting, economics, finance, management and marketing. Below is a summary of some of the case studies we analyse in this textbook, with partial data, to illustrate additional applications of inferential statistics. But you will have to wait until you work through these cases in the relevant chapters (where some data are also presented) to find out the conclusions and results.

The objective of the problem described in Case 3.6 is to use the descriptive graphical and numerical techniques to analyse the differences in weekly earnings of men and women in Australia. Case 4.2 uses graphical techniques to depict the cross-country spread of the global coronavirus pandemic. In Case 5.5 the objective is to compare the central location and variability of the house prices in Sydney and Melbourne. In Case 14.1 it is to compare two populations, the variable of interest being the salary of MBA graduates specialising in marketing and finance. Case 16.1 is a day-to-day real-life application. The objective of the problem is to see how statistical inference can be used to determine whether some numbers in a lotto draw occur more often than others. Case 17.3 illustrates another statistical objective. In this case, we need to analyse the relationship between two variables: rate of unemployment and the rate of inflation in New Zealand. By applying the appropriate statistical technique, we will be able to determine whether the two variables are related. As you will discover, the technique also permits statistics practitioners to include other variables to analyse the relationship between the rate of unemployment and inflation.

Case Studies

CASE 3.6 Differing average weekly earnings of men and women in Australia

Although a lot has been achieved in Australia to reduce the difference between men and women in a number of social status indicators, wage differences are still a matter of concern. The following table presents the average weekly cash earnings of male and female adults for each Australian state and territory and for Australia as a whole. Present the information using appropriate graphical techniques.

Average weekly (all full-time employees total) cash earnings (A\$), 2018

State/Territory	Males	Females
New South Wales	1807.80	1502.00
Victoria	1696.20	1490.30
Queensland	1777.60	1409.50
South Australia	1585.40	1393.30
Western Australia	2040.00	1495.30
Tasmania	1565.50	1320.00
Northern Territory	1877.90	1552.60
ACT	1977.90	1671.60
Australia	1823.70	1534.10

Source: Australian Bureau of Statistics, *Average Weekly Earnings*, February 2019, cat. no. 6302.0, ABS, Canberra.

CASE 4.2 Analysing the spread of the Global Coronavirus Pandemic

The coronavirus (COVID-19) pandemic outbreak was first identified in Wuhan, China, in December 2019 and by June 2020 it had spread to more than 200 countries, infected more than 10 million people and resulted in the death of more than 500 000 people globally. Daily data for the number of confirmed cases and deaths are recorded.¹ Depict graphically the number of confirmed cases and number of deaths during the six-month period for the top 10 affected countries and globally.

CASE 5.5 Sydney and Melbourne lead the way in the growth in house prices

A number of recent reports, including those published by the Commonwealth Bank, and two other national housing reports indicated that there is some improvement in house sales activity in Australia combined with a sharp increase in house prices. Some economists attributed this housing market recovery to the low interest rates and first-home owner grants and exemption from stamp duties in some states. The data below show the median (established) house prices in the capital cities of the six states and two territories in Australia for the March quarter of 2002–19. Prepare a summary report using the data below, analysing the movement in house prices in Australia as a whole and in the six states and two territories.

¹ <https://ourworldindata.org/coronavirus-source-data>, accessed 29 June 2020

Median house prices (\$'000) in eight Australian capital cities, 2002–19, March quarter

Year	Sydney	Melbourne	Brisbane	Adelaide	Perth	Hobart	Darwin	Canberra
2002	365.0	241.0	185.0	166.0	190.0	123.3	190.0	245.0
2003	434.5	270.0	225.0	209.4	216.0	145.0	198.0	300.0
2004	523.0	305.0	302.7	250.0	255.0	200.0	239.5	375.0
2005	486.0	310.0	312.0	270.0	290.0	240.0	275.0	374.7
.
.
2016	850.0	568.0	483.5	436.0	518.0	360.0	570.0	590.0
2017	930.0	657.5	510.0	451.3	510.0	389.5	530.0	654.5
2018	958.0	730.3	526.0	465.0	510.0	450.0	495.0	699.0
2019	875.0	685.0	533.5	474.0	499.0	471.5	480.0	690.0

Source: Australian Bureau of Statistics, March 2018, *Residential Property Price Indexes: Eight Capital Cities, March Quarter 2018*, cat. no. 6416.0, ABS, Canberra, www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/6416.0 Dec %202018?opendocument.

CASE 14.1 Comparing salary offers for finance and marketing MBA majors – I

In the last few years, there has been an increase in the number of web-based companies that offer job placement services. The manager of one such company wanted to investigate the job offers recent MBAs were obtaining. In particular, she wanted to know whether finance majors were being offered higher salaries than marketing majors. In a preliminary study, she randomly sampled 50 recently graduated MBAs, half of whom majored in finance and half in marketing. For each, she recorded the highest salary offer (including benefits). Can we infer that MBAs with finance majors obtain higher salary offers than do MBAs with marketing majors? Verify the underlying assumptions.

CASE 16.1 Gold lotto

Gold lotto is a national lottery that operates as follows. Players select eight different numbers (six primary and two supplementary numbers) between 1 and 45. Once a week, the corporation that runs the lottery selects eight numbers (six primary and two supplementary numbers) at random from 1 to 45. Winners are determined by how many numbers on their tickets agree with the numbers drawn. In selecting their numbers, players often look at past patterns as a way to help predict future drawings. A regular feature that appears in the newspaper identifies the number of times each number has occurred since Draw 413 (Saturday 6 July 1985). The data recorded in the following table appeared in the *Saturday Gold Lotto* website (<https://www.thelott.com/saturday-gold-lotto/results>) after the completion of draw 4047 (2 May 2020). What would you recommend to anyone who believes that past patterns of the lottery numbers are useful in predicting future drawings?

Drawing frequency of lotto numbers since draw 413 (as at 2 May 2020)

Lotto number	Number of times drawn	Lotto number	Number of times drawn	Lotto number	Number of times drawn
1	361	16	324	31	318
2	312	17	301	32	320
.
14	301	29	320	44	290
15	333	30	301	45	309

CASE 17.3 Does unemployment affect inflation in New Zealand?

A social science research student is interested in investigating the relationship between unemployment and inflation in New Zealand. The student found quarterly data on the rate of unemployment and the quarterly rate of inflation for the period 2015(1) to 2019(1) on the government website, *Statistics New Zealand*. Can you help him to establish the relationship between the two variables using the data he has collected?

1.3 How managers use statistics

As we have already pointed out, statistics is about acquiring and using information. However, the statistical result is not the end product. Managers use statistical techniques to help them make decisions. In general, statistical applications are driven by the managerial problem. The problem creates the need to acquire information. This in turn drives the data-gathering process. When a manager acquires data, they must convert the data into information by means of one or more statistical techniques. The information then becomes part of the decision process.

Many business students will take or have already taken a subject in marketing. In the introductory marketing subject, students are taught about market segmentation. Markets are segmented to develop products and services for specific groups of consumers. For example, the Coca-Cola Company produces several different cola products.

There is Coca-Cola Classic, Coca-Cola Vanilla, Coca-Cola Zero, Coca-Cola Life, Coke, Diet Coke and Caffeine-Free Diet Coke. Each product is aimed at a different market segment. For example, Coca-Cola Classic is aimed at people who are older than 30, Coca-Cola Vanilla is aimed primarily at women, Coke is aimed at the teen market, Coca-Cola Life and Diet Coke are marketed towards individuals concerned about their weight or sugar intake, and Caffeine-Free Diet Coke is for people who are health-conscious. In order to segment the cola market, the Coca-Cola Company had to determine that all consumers were not identical in their wants and needs. The company then had to determine the different parts of the market and ultimately design products that were profitable for each part of the market. As you might guess, statistics plays a critical but not exclusive role in this process.

Because there is no single way to segment a market, managers must try different segmentation variables. Segmentation variables include geographic (e.g. states, cities, country towns), demographic (e.g. age, gender, occupation, income, religion), psycho-graphic (e.g. social class, lifestyle, personality) and behaviouristic (e.g. brand loyalty, usage, benefits sought). Consumer surveys are generally used by marketing researchers to determine which segmentation variables to use. For example, the Coca-Cola Company used age and lifestyle. The age of consumers generally determines whether they buy Coca-Cola Classic or Coke. Lifestyle determines whether they purchase regular, diet or caffeine-free cola. Surveys and statistical techniques would tell the marketing manager that the 'average' Coca-Cola Classic drinker is older than 30, whereas the 'average' Coke drinker is a teenager. Census data and surveys are used to measure the size of the two segments. Surveys would also inform about the number of cola drinkers who are concerned about kilojoules and/or caffeine. The conversion of the raw data in the survey into statistics is only one part of the process. The marketing manager must then make decisions about which segments to pursue (not all segments are profitable), how to sell and how to advertise.

With the mass availability of data these days, decision making based on data analysis involves advanced data analytic techniques. Below we briefly discuss the topic business data analytics (BDA).

1.3a Business data analytics

Business data analytics provides methods by which companies explore, collect, organise and analyse data using statistical methods. When the decision making is data driven, business data analytics becomes an important tool that businesses use to make their informed decisions. As the data are unique to the company, a company can employ business data analytics to analyse their data and to make informed decisions in a competitive business environment. Data are the backbone of decision making, so the quality of the data is crucial for the successful application of business data analytics.

Business data analytics methods can be subdivided into two major areas: (a) business intelligence and (b) statistical analysis. **Business intelligence** involves a company analysing historical data related to its own organisation; for example, analysing the performance of its own staff members.

Statistical analysis comprises three major areas of analytical techniques: (1) descriptive analytics, (2) predictive analytics and (3) perspective analytics.

1.3b Descriptive analytics

Descriptive analytics involves a number of techniques that can be used to describe historical data to track the path and key performance indicators to provide understanding of the current state of a business. Such techniques could take the form of data queries, reports, summary statistics, graphical presentations, data dashboards and data mining.

- A data query, for example, can be the details in raw data form of the starting and finishing times, travel times and distances travelled by all taxi fleets owned by a taxi company. A report can be prepared using descriptive summary statistics techniques (such as mean, median, mode, standard deviation) or in graphical summary presentation form (such as tables, charts and maps) and showing a pattern of the movement of a variable (such as trends, swings etc.) or showing the relationship between variables (e.g. linear, non-linear etc.).
- Data dashboards are specific tools used by managers to make decisions in a dynamic fashion. For example, a report on the dashboard shows the information on the past data in the form of summary statistics and graphical representations. If some new data (or information) is received, the analytic techniques would update the summary statistics and

graphical representations on the dashboard instantaneously so that managers can use the additional information in their decision making. For example, if there is a sudden traffic jam due to an accident in a particular area, the manager in charge of the operations in the taxi office can give new instructions to all taxi drivers on the run about the changes they need to follow with their assignments.

- Data mining techniques are designed to identify patterns of individual variables and the relationship between variables when analysing large-scale data sets. For example, in Section 1.3, the Coca-Cola company can use data mining techniques to identify the different parts of its soft drink market using data from various sources such as sample surveys, social media, web-based reviews etc.

1.3c Predictive analytics

Predictive analytics involves applying statistical algorithms to past or historical data to analyse the trend in the data, to assess the likelihood of future outcomes and make predictions about the activities of a business, such as predicting the demand for a product for a future time period. Another example is using advanced predictive analytics techniques such as cluster analysis to identify similarities across various consumer groups in order to target marketing campaigns.

1.3d Prescriptive analytics

Prescriptive analytics uses the knowledge gained from descriptive and predictive analysis to generate recommendations for a course of action to take when similar situations arise in the future.

In this book, we will address the part of the process that collects the raw data and produces the statistical result. By necessity, we must leave the remaining elements of the decision process to the other subjects that constitute business programs. We will demonstrate, however, that all areas of business and management can, and do, use statistical techniques as part of the information system.

1.4 Statistics and the computer

In almost all applications of statistics, the statistics practitioner must deal with large amounts of data. For example, Case 4.1 (Global warming) involves hundreds of observations. To estimate the average temperature anomalies, the statistics practitioner would have to perform calculations on the data. Although the calculations do not require any great mathematical skill, the sheer number of calculations makes this aspect of the statistical method time-consuming and tedious. Fortunately, there are numerous commercially available computer programs to perform these calculations. We have chosen to use Microsoft Excel in the belief that almost all university graduates use it now and will in the future. We have also used XLSTAT to perform any calculation that cannot be performed by Excel.

In most of the examples used to illustrate statistical techniques in this book, we will provide two methods for answering the question: calculating manually and using Microsoft Excel on a computer.

1.4a Calculating manually

Except where doing so is prohibitively time-consuming, we will show how to answer the questions using hand calculations (with only the aid of a calculator). It is useful for you to produce some solutions in this way, because by doing so you will gain insights into statistical concepts.

1.4b Using a computer

Excel

Excel can perform statistical procedures in several ways.

- 1 Statistical (which includes probability) and other functions *fx*:** We use some of these functions to help create graphical techniques in Chapters 3 and 4, calculate statistics in Chapter 5, and to compute probabilities in Chapters 7 and 8.
- 2 Spreadsheets:** We use statistical functions to create Excel workbooks that calculate statistical inference methods in Chapters 10–14. These can be downloaded from the companion website (accessible through <https://login.cengagebrain.com/>). Additionally, the spreadsheets can be used to conduct what-if analyses.
- 3 Analysis ToolPak:** This group of procedures comes with every version of Excel. The techniques are accessed by clicking 'Data' and 'Data Analysis'. One of its drawbacks is that it does not offer a complete set of the statistical techniques we introduce in this book.
- 4 XLSTAT:** This is a commercially created add-in that can be loaded onto your computer to enable you to use Excel for almost all statistical procedures introduced in this book. XLSTAT commands are provided for those examples that cannot be completed using the *Data Analysis* tool in Excel.

Data Analysis Plus

We have offered *Data Analysis Plus* in the past five editions of this book. Unfortunately, we have encountered problems with some of the combinations of Excel versions and operating systems. As a result, it is no longer possible to offer *Data Analysis Plus* as a universal tool for all Excel users of this book. Section 1.5 provides further information and lists the combinations that do work. Data Analysis Plus can be downloaded from the companion website.

1.4c Our approach

We expect that most instructors will choose both methods described above. For example, many instructors prefer to have students solve small-sample problems involving few calculations manually, but turn to the computer for large-sample problems or more complicated techniques.

To allow as much flexibility as possible, most examples, exercises and cases are accompanied by a set of data that you can obtain from the companion website or your course instructor.

You can solve these problems using a software package. In addition, we have provided the required summary statistics of the data so that students without access to a computer can solve these problems using a scientific calculator.

Ideally, students will solve the small-sample, relatively simple problems by manual calculation and use the computer to solve the others. The approach we prefer to take is to minimise the time spent on manual calculations and to focus instead on selecting the appropriate method for dealing with a problem, and on interpreting the output after the computer has performed the necessary calculations. In this way, we hope to demonstrate that statistics can be as interesting and as practical as any other subject in your curriculum.

1.4d Excel spreadsheets

Books written for statistics courses taken by mathematics or statistics students are considerably different from this one. Not surprisingly, such courses feature mathematical

proofs of theorems and derivations of most procedures. When the material is covered in this way, the underlying concepts that support statistical inference are exposed and relatively easy to see. However, this book was created for an applied course in statistics. Consequently, we will not address directly the mathematical principles of statistics. As we have pointed out above, one of the most important functions of statistics practitioners is to properly interpret statistical results, whether produced manually or by computer. And, to correctly interpret statistics, students require an understanding of the principles of statistics.

To help students understand the basic foundation, we have created several Excel spreadsheets that allow for what-if analyses. By changing some of the inputs, students can see for themselves how statistics works. (The name derives from ‘*What* happens to the statistics if I change this value?’)

1.5 Online resources

This book is accompanied by a companion website that can be accessed via <https://login.cengagebrain.com>.

You will find the following available for download:

Data sets: There are approximately 800 data sets stored in folders for examples, exercises and cases.

Solutions for self-correcting exercises

Excel workbooks: There are several workbooks containing spreadsheets that perform many of the Excel procedures.

Excel instructions for Mac users

Formula card that lists every formula in the book.

DATA ANALYSIS PLUS:

Data Analysis Plus is a collection of macros we created to augment Excel’s list of statistical procedures. **Data Analysis Plus (stats.xls)** can be downloaded from the **companion** website and can be accessed via **Excel Add-ins**.

Here are the Excel Versions and Operating Systems that work with Data Analysis Plus.

(Some other combinations may work.) Excel Version Operating System:

2016 Windows

2013 Windows

2007 Windows

2003 Windows

2011 Mac Mac OS

2004 Mac Mac OS

2001 Mac Mac OS

Note that in the 2013 and 2016 versions of Excel, Data Analysis Plus conflicts with *Data Analysis*. As a result, in order to use Data Analysis Plus, it is necessary to temporarily remove the Analysis ToolPak (i.e. disable Data Analysis).

EXERCISES

- 1.1** In your own words, define and give an example of each of the following statistical terms:
- population
 - sample
 - parameter
 - statistic
 - statistical inference
- 1.2** Briefly describe the difference between descriptive statistics and inferential statistics.
- 1.3** The manager of a bank with 12000 customers commissions a survey to gauge customer views on internet banking, which would incur lower bank fees. In the survey, 21% of the 300 customers interviewed said they are interested in internet banking.
- What is the population of interest?
 - What is the sample?
 - Is the value 21% a parameter or a statistic?
- 1.4** A light bulb manufacturer claims that fewer than 5% of his bulbs are defective. When 1000 bulbs were drawn from a large production run, 1% were found to be defective.
- What is the population of interest?
 - What is the sample?
 - What is the parameter?
 - What is the statistic?
 - Does the value 5% refer to the parameter or to the statistic?
 - Is the value 1% a parameter or a statistic?
 - Explain briefly how the statistic can be used to make inferences about the parameter to test the claim.
- 1.5** Suppose you believe that, in general, the salaries offered to graduates of business programs are higher than those offered to graduates of arts and science programs. Describe a statistical experiment that could help test your belief.
- 1.6** You are shown a coin that its owner says is fair in the sense that it will produce the same number of heads and tails when flipped repeatedly.
- Describe an experiment to test this claim.
 - What is the population in your experiment?
 - What is the sample?
 - What is the parameter?
 - What is the statistic?
 - Describe briefly how statistical inference can be used to test the claim.
- 1.7** Suppose that in Exercise 1.6 you decide to flip the coin 100 times.
- What conclusion would you be likely to draw if you observed 95 heads?
 - What conclusion would you be likely to draw if you observed 55 heads?
 - Do you believe that, if you flip a perfectly fair coin 100 times, you will always observe exactly 50 heads? If you answered 'no', what numbers do you think are possible? If you answered 'yes', how many heads would you observe if you flipped the coin twice? Try it several times, and report the results.

Appendix 1.A

Introduction to Microsoft Excel

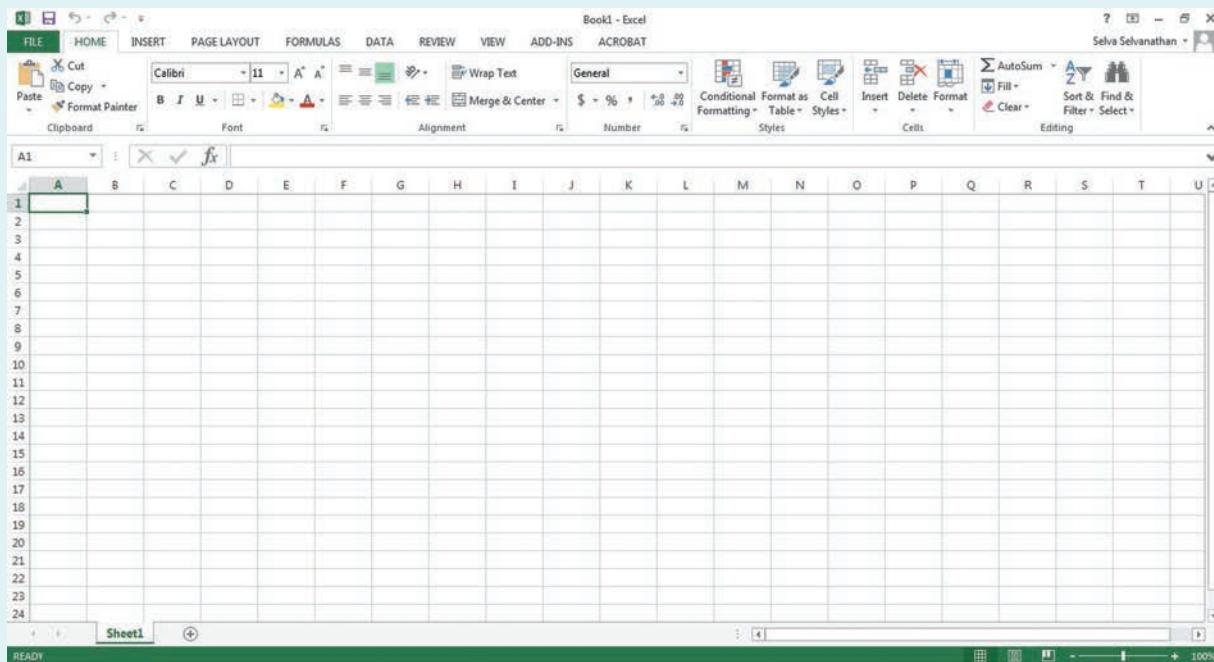
The purpose of this appendix is to introduce you to Microsoft Excel and to provide enough instructions to allow you to use Excel to produce the statistical results discussed in this textbook. We suggest that you obtain an Excel instruction book to help you learn about the software in more detail. (e.g. for Office 2016, use Selvanathan et al, *Learning Statistics and Excel in Tandem*, 5th edition, Cengage, 2020).

Opening Excel

Usually an Excel icon will be on your computer desktop. To execute Excel, just double-click the icon. Alternatively, to run Microsoft Excel, just click **Start, All Programs, Microsoft Office 2016** and **Excel 2016**. A blank workbook will appear, or if a menu appears, click on **Blank workbook**.²

The Excel screen depicted in **Figure A1.1** will then appear. (Unless stated otherwise, ‘clicking’ refers to tapping the left button on your mouse once; ‘double-clicking’ means quickly tapping the left button twice.) The screen that appears may be slightly different from that illustrated, depending on your version of Excel.

FIGURE A1.1 Blank Excel screen



The Excel workbook and worksheet

Excel files are called workbooks, which contain worksheets. A worksheet consists of rows and columns. The rows are numbered, and the columns are identified by letters.

² Instructions and screen view may be different for different versions of Excel. Excel can be accessed by typing Excel in 'Type here to search' and the icon with Excel will appear. Click on the icon and access Excel.

If you click the **FILE** tab on the top left-hand corner of the Excel screen and select **New** and then **Blank workbook**, you will see a worksheet like that in **Figure A1.1**. Notice that the cell in row 1 column A is active, which means that you can type in a number, word or formula in cell A1. To designate any cell as active, move the cursor (or the mouse pointer), which now appears as a large plus sign) over the cell and click. Alternatively, you can use any of the four **Up, Down, Left** or **Right** arrow keys to move to a different cell. (The keys appear on your keyboard as arrows pointing up, down, left and right respectively.)

At the bottom left-hand corner of the screen you will see the word **Ready**. As you begin to type something into the active cell, the word **Ready** changes to **Enter**. Above this word you will find the tab **Sheet1**, the worksheet that comprises this workbook. You can click the '+' icon next to this tab to add more worksheets to the workbook. To change worksheet, use your cursor and click the tab for the sheet you wish to move to. To change the name of the worksheet, for example 'Sheet2' to 'mydata', double-click on 'Sheet2' and type 'mydata', then hit the **Enter** key.

Inputting data

To input data, open a new workbook by clicking the **FILE** tab and then selecting **New** and then **Blank workbook**. Data are usually stored in columns. Activate the cell in the first row of the column in which you plan to type the data. If you wish, you may type the name of the variable. For example, if you plan to type your assignment marks in column A you may type 'Assignment marks' into cell A1. Hit the **Enter** key and cell A2 becomes active. Begin typing the marks. Follow each entry by **Enter** to move to the next row. Use the arrow keys or mouse to move to a new column if you wish to enter another set of numbers.

Importing data files

Data files accompany many of the examples, exercises and cases in this book. These data files can be downloaded from the companion website, which can be accessed at <https://login.cengagebrain.com>.

The file names reflect the chapter number and the example, exercise or case number. For example, the data set accompanying Example 4.1 contains 200 electricity bills. These data are stored in a file called **XM04-01**, which is stored in a directory (or folder) called **CH04**, which refers to the chapter number. (The XM refers to files used for eXamples, XR is used for eXercises, and C refers to Cases.) To open this file in Excel 2016, click the FILE tab and then select Open. You can then browse your computer to select the required file.

Performing statistical procedures

There are several ways to conduct a statistical analysis in Excel. These include **Data Analysis**, **XLSTAT** and the **Insert function** f_x (f_x is displayed on the formula bar). There are also some workbooks, available for download from the companion website, that we have created that will allow you to perform some of the statistical estimations and tests discussed in this book.

Data Analysis/Analysis ToolPak

The **Analysis ToolPak** is a group of statistical functions that comes with Excel. The **Analysis ToolPak** can be accessed through the **Menu bar**. Click the **DATA** tab on the menu bar and then select **Data Analysis** from the Analysis submenu. If the Analysis submenu does not appear, click on the **FILE** tab (on the top left-hand corner of the Excel screen), select **Options** and, on the window that

appears, click on the **Add-Ins** tab. At the bottom of the window, select **Excel Add-ins** in the drop-down menu that appears next to **Manage**, and then click Go. Select **Analysis ToolPak** and **Analysis ToolPak-VBA** by ticking them and then click OK. Now, if you go back to DATA on the menu bar, you will find Data Analysis added to the Analysis submenu. (Note that **Analysis ToolPak** is not the same as **Analysis ToolPak-VBA**.) If **Analysis ToolPak** did not show up as an option in the Add-ins menu, you will need to install it from the original Excel or Microsoft Office CD by running the setup program and following the instructions.

If you click on Data Analysis, there are 19 menu items available. Click the one you wish to use, and follow the instructions described in this book. For example, one of the techniques in the menu, **Descriptive Statistics**, is described in Chapter 5.

Formula bar and insert function f_x

On the formula bar you will find the f_x insert function. Clicking this button produces other menus that allow you to specify functions that perform various calculations.

Saving workbooks

To save a file (either one you created or one we created that you have altered) click the **FILE** tab on the top left-hand corner of the screen and select the option **Save as** on the drop-down menu. Select the folder in which you would like to save the file, then enter the new file name and click **Save**. If you want to save changes to an already saved file with the same name, choose **Save** on the drop-down menu. Caution, the original file will now be overwritten.

Types of data, data collection and sampling

Learning objectives

This chapter discusses the types of data and various methods of collecting data.

At the completion of this chapter, you should be able to:

- L01** describe different types of data
- L02** understand the primary and secondary sources of statistical data
- L03** explain the various methods of collecting data
- L04** explain the basic sampling plans
- L05** identify the appropriate sampling plan for data collection in a particular experiment
- L06** understand the main types of errors involved in sampling.

CHAPTER OUTLINE

Introduction

2.1 Types of data

2.2 Methods of collecting data

2.3 Sampling

2.4 Sampling plans

2.5 Sampling and non-sampling errors

SPOTLIGHT ON STATISTICS

Sampling and the census

The census, which is conducted by the Australian Bureau of Statistics (ABS) every five years in Australia, performs an important function. The two most recent censuses were carried out on the nights of 9 August 2011 and 2016. On census night, every person present in Australia, excluding foreign diplomats and their families, should be included on a census form at the place where they stayed. The federal, state and territory governments in Australia use various aspects of the census information for economic planning. Businesses often use the information derived from the census to help make decisions about products, advertising and business locations.

One of the problems with the census is the issue of undercounting. In the census, some people are missed and some are counted more than once. Usually, more people are missed than are overcounted. The difference between the



Source: Shutterstock.com/Vlada Young

census count and the true population is called the net undercount of the census. For example, the 2016 census missed 137 750 Aboriginal and Torres Strait Islander persons or 17.5% of the Aboriginal and Torres Strait Islander population who were present in Australia on census night. The extent of the net undercount varied for different population groups. Here are some of the most interesting observations about the 2016 Census:

- The net undercount rate in the 2016 Census for Australia as a whole is 1%.
- Among the states and territories, the undercount rate was highest for the Northern Territory (5%) and lowest for the Australian Capital Territory (-1.1%, i.e. net overcount of 1.1%).
- Among the various age groups, children (0–4 years) had the highest undercount rate (5.1%), while older people (70–74 years) had the lowest rate (-3.9%, i.e. net overcount of 3.9%).
- Widowed, divorced or separated people had higher undercount rates (0.6%) than people who were never married or married (0.3%).
- The undercount rate for people born in Australia was higher (8.1%) than that for people born in India (5%), which was the lowest.¹
- Males are more likely to be missed in the Census compared to females, with net undercount rates of 1.5% and 0.4%, respectively.

To address undercounting in Australia, the ABS adjusts the numbers it gets from the census. The adjustment is based on another survey called the Post Enumeration Survey (PES), which is carried out by the ABS about three weeks after the census night from 36 000 households. The 36 000 households are selected using sampling methods described in this chapter. The ABS is able to adjust the numbers for the different population groups. Later in this chapter (pages 37–8) we will discuss how the sampling is conducted and how the adjustments are made.

Introduction

In Chapter 1, we introduced the distinction between a population and a sample. Recall that a *population* is the entire set of observations or measurements under study, whereas a *sample* is a subset of observations selected from the entire population.

The three most important factors that determine the appropriate statistical method to use are the problem objective, the type of data and the type of descriptive measurement. In this chapter, our focus will be on the types of data and data collection methods.

In Chapter 1, we also briefly introduced the concept of statistical inference – the process of inferring information about a population from a sample. Because information about populations can usually be described by parameters, the statistical technique used generally deals with drawing inferences about population parameters (e.g. average height of Australian adult males) from sample statistics (e.g. average height of a sample of Australian adult males). Recall that a parameter is a measurement about a population, and a statistic is a measurement about a sample.

In a statistics textbook, we can assume that population parameters are known. In real life, however, calculating parameters is virtually impossible because populations tend to be very large. As a result, most population parameters are not only unknown but also unknowable (e.g. the average length of all fish in the Brisbane River). The problem that motivates the subject of statistical inference is that we often need information about the value of the population parameters to make decisions. For example, to make decisions about whether to expand a line of clothing, we may need to know the average annual expenditure on clothing by Australian adults. Because the size of this population is approximately 25.6 million, determining the population average expenditure is not possible. However, if we are willing

¹ Source: Australian Bureau of Statistics, *Census of Population and Housing – Details of Overcount and Undercount, Australia 2016*, 23 February 2018, cat. no. 2940.0, ABS, Canberra.

estimate

An approximate value of a parameter based on a sample statistic.

to accept less than 100% accuracy, we can use statistical inference to obtain an **estimate**. Rather than investigating the entire population, we select a sample of people, determine the annual expenditure on clothing of this group, and calculate the sample average. Although the probability that the sample average will equal the population average is very small, we would expect the averages to be close. For many decisions we need to know how close these two averages are. We will discuss this further in Chapters 10 and 11.

The discussion above shows that in order to make statistical inference about a population parameter, data collection is crucial. Also, since the selection of statistical techniques depends on the type of data we have, we also need to understand various classifications of data. In Sections 2.2–2.5 of this chapter we will first present the various sources for collecting data and then discuss the basic concepts and techniques of sampling itself.

2.1 Types of data

The objective of statistics is to extract information from data. There are different types of data and information. To help explain this important principle, we need to define some terms.

variable

Any characteristic of a population or sample.

A **variable** is some characteristic of a population or sample. For example, the mark on a statistics exam is a characteristic of statistics exams that certainly is of interest to readers of this book. Not all students achieve the same mark. The marks will vary from student to student, thus the name *variable*. The price of shares is another variable. The prices of most shares vary daily (sometimes every minute). We usually represent the name of the variable using uppercase letters such as X , Y and Z .

The *values* of the variable are the possible observations of the variable. The values of statistics exam marks are the integers between 0 and 100 (assuming the exam is marked out of 100). The values of a stock price are real numbers that are usually measured in dollars and cents (sometimes in fractions of a cent). The observations on stock prices range from 0 to hundreds of dollars. **Data** are the observed values of a variable. For example, in April 2019 the Australian Bureau of Statistics published the rate of unemployment of people 15 years and older for the six states and two territories of Australia: New South Wales, Victoria, Queensland, South Australia, Western Australia, Tasmania, Australian Capital Territory and Northern Territory, respectively, as

4.5, 4.9, 5.9, 6.1, 6.1, 6.8, 3.9, 4.5

These are the data from which we will extract the information we seek. (Incidentally, *data* is plural for *datum*. The rate of unemployment for an individual state is a datum.)

2.1a Three types of data

When most people think of data they think of sets of numbers. However, there are three types of data:²

- 1 numerical (or quantitative or interval) data
- 2 nominal (or categorical or qualitative) data
- 3 ordinal (or ranked) data.

If 75 managers are surveyed and asked to state their age and annual income, the responses they give are real numbers and are called **numerical data**. We also refer to this type of data as *quantitative* or *interval*. All types of statistical calculations are permitted on numerical data.

If the 75 managers surveyed above are also asked to indicate their marital status (single, married, divorced or widowed), their responses are non-numerical, but each response can still be classified as falling into one of four categories. Observations that can be sorted into categories on the basis of qualitative attributes – such as marital status, gender, occupation

² There are actually four types of data, the fourth being ratio data. However, for statistical purposes there is no difference between ratio and numerical data. Consequently, we combine the two types.

or type of dwelling inhabited – constitute **nominal data**. Nominal data are also referred to as *categorical* or *qualitative* data.³ The data in these examples are simply the names of possible classifications.

Knowing the type of data being measured is important, because it is one of the factors that determine which statistical technique should be used. Usually, identifying the data as being either numerical or nominal will be sufficient. But in a few situations it will be necessary to recognise whether or not the non-numerical data under consideration can be ranked. The data are then said to have an **ordinal scale**.

Ordinal data appear to be nominal, but their values are in order. Consequently, ordinal data are also called ranked data. The most important aspect of ordinal data is the order of the values. As a result the only permissible calculations are those involving a ranking process.

If the 75 managers surveyed above were also asked to classify a particular hotel as excellent, good, fair or poor, based on the quality of accommodation provided, their responses here are non-numerical and appear to be nominal. Notice, however, that the responses are ranked in preferential order by the quality of accommodation. The first response (excellent) is the highest rating, the second response (good) is the second-highest rating, and so on. Any numerical representation of the four answers should maintain the ranked ordering of the responses. Such a numerical system would form an ordinal scale. The only constraint upon our choice of numbers for the scale is that they must represent the order of responses; the actual values to be used are arbitrary. For example, we could record the responses as follows:

$$\text{Excellent} = 4; \quad \text{Good} = 3; \quad \text{Fair} = 2; \quad \text{Poor} = 1$$

We could simply reverse this 4-3-2-1 rating system so that Excellent = 1 and Poor = 4, with no effect on the statistical technique to be used.

Another, equally valid representation of the ratings would be the following:

$$\text{Excellent} = 9; \quad \text{Good} = 5; \quad \text{Fair} = 3; \quad \text{Poor} = 1$$

The only information provided by ordinal (ranked) data that is not provided by nominal data is the ranked order of the responses. We still cannot interpret the difference between values for ranked data, because the actual numbers used are arbitrary. For example, the 4-3-2-1 rating implies that the difference between excellent and good ($4 - 3 = 1$) is equal to the difference between fair and poor ($2 - 1 = 1$), whereas the 9-5-3-1 rating implies that the difference between excellent and good ($9 - 5 = 4$) is two times as large as the difference between fair and poor ($3 - 1 = 2$). Since both numbering systems are valid (though arbitrary), no inferences can be drawn about the differences between values of ranked variables. Thus, statistics such as averages are often misleading for ordinal data. To illustrate this point, suppose that of 10 people interviewed, four rated their hotel accommodation as excellent, three as good and three as poor. The average using the 4-3-2-1 system is 2.8, which suggests that the average hotel is between fair and good. Using the 9-5-3-1 system, the average is 5.4, which implies that the average hotel is between good and excellent.

2.1b Performing calculations on the three types of data

Numerical data

All calculations are permitted on numerical data. We often describe a set of numerical data by calculating the average. For example, the average of the 8 unemployment rates listed on page 20 is 5.3%. As you will discover, there are several other important statistics that we will introduce.

nominal data
Observations are categorical or qualitative.

ordinal scale
A scale applied to ranked data.

ordinal data
Ordered nominal data.

³ In recent years, the term *qualitative* has been used by social scientists to represent information gathered in the form of a collection of words, printed text messages and audio/video interviews. Therefore, to avoid confusion, in this book we have avoided using the term *qualitative data*. Instead, we refer to categorical observations as nominal or categorical data.

Nominal data

Because the codes of nominal data are completely arbitrary, we cannot perform any calculations on these codes. To understand why, consider a survey that asks people to report their marital status. Suppose that the first 10 people surveyed gave the following responses:

Single, Married, Married, Married, Widowed,

Single, Married, Married, Single, Divorced

If we use the following numerical codes:

Single = 1, Married = 2, Divorced = 3, Widowed = 4,

we would record these responses as:

1, 2, 2, 2, 4, 1, 2, 2, 1, 3

The average of these numerical codes is 2.0. Does this mean that the average person is married? Now suppose four more persons were interviewed, of whom three are widowed and one is divorced. The data are given here:

1, 2, 2, 2, 4, 1, 2, 2, 1, 3, 4, 4, 4, 3

The average of these 14 codes is 2.5. Does this mean that the average person is married – but halfway to getting divorced? The answer to both questions is an emphatic ‘no’. This example illustrates a fundamental truth about nominal data: calculations based on the codes used to store this type of data are meaningless. All that we are permitted to do with nominal data is count the frequencies or compute the percentages of the occurrences of each category. Thus, we would describe the 14 observations by counting the number of each marital status category and reporting the frequency as shown in the following table.

Category	Code	Frequency	Percentage (%)
Single	1	3	21
Married	2	5	36
Divorced	3	2	14
Widowed	4	4	29

In Chapter 3, we introduce graphical techniques that are used to describe nominal data. Chapter 4 deals with numerical data.

Ordinal data

The most important aspect of ordinal data is the order of the values. As a result, the only permissible calculations are those involving a ranking process. For example, we can place all the data in order and select the code that lies in the middle. As we discuss in Chapter 5, this descriptive measurement is called the median.

It should be understood that the average of ordinal data may provide *some* useful information. However, because of the arbitrary nature of this type of data, the most appropriate statistical techniques are those that put the data in order. (In Chapter 5 we present the *median*, which is calculated by placing the numbers in order and selecting the observation that falls in the middle.) You will find that an ordering process is used throughout this book whenever the data are ordinal.

2.1c Hierarchy of data

The data types can be placed in order of the permissible calculations. At the top of the list we place the numerical data type because virtually *all* calculations are allowed. The nominal data type is at the bottom because *no* calculations other than determining frequencies

are permitted. (We are permitted to perform calculations using the frequencies of codes. However, this differs from performing calculations on the codes themselves.) In between numerical and nominal data lies the ordinal data type. Permissible calculations are those that rank the data.

Higher-level data types may be treated as lower-level ones. For example, in universities and TAFE colleges the marks in a course, which are numerical, are converted to letter grades, which are ordinal. Some graduate courses feature only a pass or fail designation. In this case, the numerical data are converted to nominal data. It is important to point out that when we convert higher-level data to lower-level data we lose information. For example, a mark of 83 on an accounting course exam gives far more information about the performance of that student than does a letter grade of A, which is the letter grade for marks between 75 and 85. As a result, we do not convert data unless it is necessary to do so. We will discuss this later.

It is also important to note that we cannot treat lower-level data types as higher-level data types. The definitions and hierarchy are summarised in the following box.

IN SUMMARY

Types of data

- *Numerical (or quantitative or interval)*

Values are real numbers.

All calculations are valid.

Data may be treated as ordinal or nominal.

- *Ordinal (or ranked)*

Values must represent the ranked order of the data.

Calculations based on an ordering process are valid.

Data may be treated as nominal but not as numerical.

- *Nominal (or categorical or qualitative)*

Values are the arbitrary numbers that represent categories.

Only calculations based on the frequencies of occurrence are valid.

Data may not be treated as ordinal or numerical.

EXAMPLE 2.1

LO1

Census survey responses and their data type

In the 2016 census, an adult member of the household was required to complete a number of questions about every member staying in their household on census night. These included the following types of questions. For each question, determine the data type of possible responses.

- Is the person male or female?
- What is the person's age at last birthday (in whole years)?
- What is the person's marital status (single/married/de facto/divorced/widowed)?
- What is the level of highest educational qualification the person has completed (PhD/Masters/Bachelors//TAFE/High School/Other)?
- What is the total of all weekly wages/salaries, government benefits, pensions, allowances and other incomes the person usually receives (\$2000 or more; \$1500–1999; \$1250–1499; \$1000–1249; \$800–999; \$600–799; \$400–599; \$300–399; \$200–299; \$1–199; nil income; negative income)?



Solution

- a A person's gender (male or female) is a categorical response and the data type is nominal.
- b A person's age at last birthday (in whole years) is numerical.
- c A person's response for marital status is categorical and the type of data is nominal.
- d The level of highest educational qualification is nominal. However, it can be ranked based on the level of education. The responses provided would be ordinal data.
- e As a person's weekly income needs to be selected from a range of income groups, it is nominal. Since the income groups can be ranked, the data are ordinal.

Numerical, ordinal and nominal variables

The variables whose observations constitute our data will be given the same name as the type of data. Thus, for example, numerical data are the observations of a numerical variable.

2.1d Other types of data

The most commonly used other types of data are time series data, cross-sectional data and panel data. Time-series data are collected over time (e.g. income of an employee in a firm over the last 10 years), cross-sectional data are collected over a number of subjects at a single point in time or period of time (for example, income of all the employees in a firm last year) and panel data are data collected over time and subjects (for example, the income of all the employees in a firm over the last 10 years). A data set that is too large or too complex to be dealt with using traditional software packages is called big data.

Time-series data

Data of a numerical variable collected over time are called *time-series data*. The specified time period should be a regular time interval such as yearly, quarterly, monthly, weekly, daily or hourly. For example, if we record the Gross Domestic Product (GDP) of a country over the last 20 years, those data are time-series data.

Cross-sectional data

Data of a variable collected at a particular fixed point in time or period of time but across a number of subjects (such as individuals, households, firms or countries) are called *cross-sectional data*. For example, if we record the 2020 GDP of 30 developed countries in the world, those data are cross-sectional data.

Panel data

Data of a variable collected over time and for a number of subjects are called *panel data*. For example, if we record the individual GDP of 30 developed countries over the last 20 years (2000–19), such data are panel data.

Big data

Big data are any data set that is too large or too complex to handle by standard data-processing techniques or by standard statistical software packages.

For example, imagine that each person out of the 7.7 billion population in this world, on average, sends a text message every minute. Then the number of text messages sent or received in a minute would be around 7.7 billion; over a day it would be around $24 \times 60 \times 7.7 = 11\,088$ billion text messages. According to the Singapore Changi Airport Group report, in 2018 about 65.6 million passengers, 2.1 million tonnes of air freight and 386 000 aircraft went through the Singapore airport. Various information on these 65.6 million passengers, 2.1 million tonnes of air freight and 386 000 aircraft are recorded in the Singapore airport data

base. The medical data base of people in each country has billions of records in total. Bigger supermarkets such as Woolworths and Coles will have a record of every item Australian households (or 25 million Australians) purchase from their supermarkets.

Big data are unstructured, large in volume of information (in billions of billions of data points), generated at a high speed or velocity (billions of billions per second) and consist of different forms (audio, video, text, numbers). People in the business environment use spreadsheet packages such as Excel and statistical software packages such as SPSS, Minitab, SAS and STATA. However, these software packages have limitations when handling big data sets. A different approach is required when handling such data sets. One of the approaches used to analyse big data is machine learning, which is a form of artificial intelligence where computers analyse large data sets and learn patterns that will help them make predictions. However, these sophisticated approaches are beyond the scope of this book.

2.1e Problem objectives/information

As discussed earlier, identifying the type of data is a critical factor in deciding which statistical procedure to use. A second factor is the type of information we need to produce from our data. We will discuss the different types of information in greater detail in Chapter 9 when we introduce problem objectives. However, in this part of the book (Chapters 3, 4 and 5) we will use statistical techniques to describe a set of data and to describe the relationship between two variables.

EXERCISES

- 2.1** Provide two examples each of numerical, ordinal and nominal data.
- 2.2** For each of the following examples of data, determine the data type:
 - a the heights of children attending school in Toowoomba
 - b the starting salaries of MBA graduates in Australia
 - c the final letter grades received by students in a statistics subject
 - d the number of kilometres driven annually by employees in company cars
 - e the age at death of Australians who died due to the COVID-19 pandemic in 2020.
- 2.3** For each of the following examples of data, determine the data type:
 - a the number of daily global deaths due to the COVID-19 pandemic during 2020
 - b the department in which each of a sample of university professors teaches
 - c the weekly closing price of gold throughout 2020
 - d the size of soft drink (large, medium or small) ordered by each of a sample of customers in a restaurant
 - e the number of Toyota cars sold each month in Auckland
- 2.4** Information concerning a magazine's readership is of interest both to the publisher and to the magazine's advertisers. A survey of 500 subscribers included the following questions. For each question, determine the data type of possible responses.
 - a What is your age?
 - b What is your gender?
 - c What is your marital status?
 - d Is your annual income less than \$30 000, between \$30 000 and \$60 000, or over \$60 000?
 - e To how many other magazines do you subscribe?
 - f How do you rate the feature article published in the current issue of the magazine (very good, good or poor)?
- 2.5** A random sample of 100 university academics was taken. Each academic was asked the following questions. Identify the data type for each question.
 - a What is your rank (lecturer, senior lecturer, associate professor, professor)?
 - b What is your annual salary?
 - c In which faculty (Arts, Science, Business, Engineering, etc.) of the university are you employed?
 - d For how many years have you been an academic?
 - e How many different courses have you taught?

- 2.6** Rugby fans are regularly quizzed about their opinions on various aspects of the sport. A sample of 300 rugby fans in Melbourne was asked the following questions. Identify the data type for each question.
- a** How many games do you attend annually?
 - b** Would you rate the entertainment as excellent, good, fair or poor?
 - c** Do you use public transport to arrive at the venue?
 - d** How much money on average do you spend at the cafe at each game?
 - e** Rate the food quality: excellent, satisfactory or very poor.
 - f** Do you take your whole family to the match (yes or no)?
- 2.7** The placement office at a university regularly surveys its graduates one year after graduation and asks for the following information. For each question, determine the type of data.
- a** What is your occupation?
 - b** What is your annual income?
 - c** What degree did you obtain?
 - d** How would you rate the overall quality of instruction at the university?
 - e** What is the amount of your HECS debt?
- 2.8** Residents of a high-rise block of flats in Perth city were recently surveyed and asked a series of questions. Identify the type of data for each question.
- a** What is your age?
 - b** On which floor do you live?
- 2.9** Do you own or rent?
- d** How many bedrooms are there in your flat?
 - e** Does your block have a pool?
 - f** How would you rate your satisfaction with the common facilities (e.g. swimming pool, tennis court) available to the residents: very good, good, poor or very poor?
- 2.10** A sample of shoppers at the Westfield shopping centre at Garden City in Brisbane was asked the following questions. Identify the type of data each question would produce.
- a** What is your age?
 - b** How much did you spend?
 - c** What is your marital status?
 - d** Rate the availability of parking: excellent, good, fair or poor.
 - e** How many stores did you enter during this visit?
- 2.10** At the end of the semester, university students often complete questionnaires about their courses. Suppose that at one university, students were asked to complete the following survey. Determine the type of data each question produces.
- a** Rate the course (highly relevant, relevant, not sure, irrelevant or very irrelevant).
 - b** Rate the lecturer (very effective, effective, not sure, not too effective or not at all effective).
 - c** What was your mid-semester exam grade (HD, D, C, P or F)?
 - d** How many lectures did you attend in this course this semester?

2.2 Methods of collecting data

Most of this book addresses the problem of converting data into information. The question arises: Where do data come from? The answer is that there are a number of ways we can collect data. In some situations, we may have access to raw data already collected by someone or an organisation and readily available for us to use. In other situations, we may need to go into the field and collect raw data ourselves. Before we proceed, however, we will remind you of the definition of data introduced in Section 2.1. Data are the observed values of a variable. That is, we define a variable or variables that are of interest to us and then proceed to collect observations of those variables.

The validity of the results of a statistical analysis clearly depends on the reliability and accuracy of the data used. Whether you are actually involved in collecting the data, performing a statistical analysis on the data or simply reviewing the results of such an analysis, it is important to realise that the reliability and accuracy of the data depend on the method of collection. The four most popular sources of statistical data are:

- 1** published data
- 2** data collected from observational studies
- 3** data collected from experimental studies
- 4** data collected from surveys.

2.2a Published data

The use of published data is often preferred due to its convenience, relatively low cost to obtain and its reliability (assuming that it has been collected by a reputable organisation). There is an enormous amount of published data produced by government agencies and private organisations, and available in printed form, on data tapes and disks, and increasingly on the internet. Data published by the same organisation that collected them are called primary data. An example of primary data would be the data published by the ABS, which collects data on numerous industries as well as conducting the census of the population every five years. The Australian Securities Exchange (ASX) Ltd collects all Australian stock market data every day. These primary sources of information are invaluable to decision makers in both the government and the private sectors.

Secondary data refers to data that are published by an organisation different from the one that originally collected and published the data. A popular source of secondary data is the *Yearbook of National Accounts Statistics* (United Nations, New York), which compiles data from several primary government sources and is updated annually. Another example of a secondary data source is *Compustat*, which sells a variety of financial data tapes that contain data compiled from such primary sources as the New York Stock Exchange.

Care should be taken when using secondary data, as errors may have been introduced as a result of the transcription or due to misinterpretation of the original terminology and definitions employed.

2.2b Direct observation

The simplest method of obtaining data is by direct observation. Data gathered in this way are said to be observational data. For example, suppose that a researcher for a pharmaceutical company wants to determine whether aspirin does reduce the incidence of heart attacks. Observational data may be gathered by selecting a sample of men and women and asking each whether he or she has taken aspirin regularly over the past two years. Each person would be asked whether he or she had suffered a heart attack over the same period. The proportions reporting heart attacks would be compared and a statistical technique that is introduced in Chapter 13 would be used to determine whether aspirin is effective in reducing the likelihood of heart attacks. There are many drawbacks to this method. One of the most critical is that it is difficult to produce useful information in this way. For example, if the statistics practitioner concludes that people who take aspirin suffer fewer heart attacks, can we conclude that aspirin is effective? It may be that people who take aspirin tend to be more health conscious, and health-conscious people tend to have fewer heart attacks. One advantage of direct observation is that it is relatively inexpensive.

2.2c Experiments

A more expensive but better way to produce data is through experiments. Data produced in this manner are called experimental data. In the aspirin illustration, a statistics practitioner can randomly select men and women. The sample would be divided into two groups. One group would take aspirin regularly and the other would not. After two years the statistics practitioner would determine the proportion of people in each group who had suffered a heart attack, and again use statistical methods to determine whether aspirin is effective. If we find that the aspirin group suffered fewer heart attacks, we may more confidently conclude that taking aspirin regularly is a healthy decision.

2.2d Surveys

One of the most familiar methods of collecting primary data is the survey, which solicits information from people concerning such things as income, family size, and opinions on various issues. We are all familiar, for example, with opinion polls that accompany each political election. The Morgan Poll and the Newspoll are two well-known surveys of public opinion whose results are often reported by the Australian media. But the majority of surveys are conducted for private use. Private surveys are used extensively by market researchers to determine the preferences and attitudes of consumers and voters. The results can be used for a variety of purposes, from helping to determine the target market for an advertising campaign to modifying a candidate's platform in an election campaign. It is quite likely that many students reading this book will one day be marketing executives who will 'live and die' by such market research data.

An important aspect of surveys is the response rate. The response rate is the proportion of all people selected who complete the survey. As we discuss in the next section, a low response rate can destroy the validity of any conclusion resulting from the statistical analysis. Statistics practitioners need to ensure that data are reliable.

Personal interview

Many researchers feel that the best way to survey people is by means of a personal interview, which involves an interviewer soliciting information from a respondent by asking prepared questions. A personal interview has the advantage of having a higher expected response rate than other methods of data collection. In addition, there will probably be fewer incorrect responses resulting from respondents misunderstanding some questions, because the interviewer can clarify misunderstandings when asked to. But the interviewer must also be careful not to say too much, for fear of biasing the response. To avoid introducing such biases, as well as to reap the potential benefits of a personal interview, the interviewer must be well trained in proper interviewing techniques and well informed on the purpose of the study. The main disadvantage of personal interviews is that they are expensive, especially when travel is involved.

Telephone interview

A telephone interview is usually less expensive, but it is also less personal and has a lower expected response rate. Unless the issue is of interest, many people will refuse to respond to telephone surveys. This problem is exacerbated by telemarketers trying to sell some products instead of conducting a telephone survey.

Self-administered survey

A third popular method of primary data collection is the self-administered questionnaire, which is usually mailed to a sample of people selected to be surveyed. This is a relatively inexpensive method of conducting a survey and is therefore attractive when the number of people to be surveyed is large. But self-administered questionnaires usually have a low response rate and may have a relatively high number of incorrect responses due to respondents misunderstanding some questions.

Questionnaire design

The instrument used in a survey is called a 'questionnaire'. Whether a questionnaire is self-administered or completed by an interviewer, it must be well designed. Proper questionnaire design takes knowledge, experience, time and money. Some basic points to consider regarding questionnaire design follow.

- 1 The questionnaire should be kept as short as possible to encourage respondents to complete it. Most people are unwilling to spend much time filling out a questionnaire.
- 2 The questions themselves should also be short, as well as simply and clearly worded, to enable respondents to answer quickly, correctly and without ambiguity. Even familiar terms such as ‘unemployed’ and ‘family’ must be defined carefully because several interpretations are possible.
- 3 Questionnaires often begin with simple demographic questions to help respondents get started and become comfortable quickly.
- 4 Dichotomous questions (questions with only two possible responses, such as ‘yes’ and ‘no’ or ‘true’ and ‘false’) and multiple-choice questions are useful and popular because of their simplicity, but they, too, have possible shortcomings. For example, a respondent’s choice of ‘yes’ or ‘no’ to a question may depend on certain assumptions not stated in the question. In the case of a multiple-choice question, a respondent may feel that none of the choices offered is suitable.
- 5 Open-ended questions provide an opportunity for respondents to express opinions more fully, but they are time-consuming and more difficult to tabulate and analyse.
- 6 Avoid using leading questions, such as ‘Wouldn’t you agree that the statistics exam was too difficult?’ These types of questions tend to lead the respondent to a particular answer.
- 7 Time permitting, it is useful to pre-test a questionnaire on a small number of people in order to uncover potential problems, such as ambiguous wording.
- 8 Finally, when preparing the questions, think about how you intend to tabulate and analyse the responses. First determine whether you are soliciting values (i.e. responses) for a numerical variable or a nominal variable. Then consider which type of statistical techniques – descriptive or inferential – you intend to apply to the data to be collected, and note the requirements of the specific techniques to be used. Thinking about these questions will help to assure that the questionnaire is designed to collect the data you need.

Whatever method is used to collect primary data, we need to know something about sampling, the subject of the next section.

EXAMPLE 2.2

LO2 LO3

Identifying the appropriate method of data collection

Discuss the method of data collection you would choose to collect data for the following statistical analyses.

- a A political analyst would like to analyse the voting intentions of New Zealand voters among the political parties: National, Labour, Green, New Zealand First, Māori and Independents/Other.
- b A banking merchant would like to investigate the reasons of the four major banks (NAB, ANZ, CBA and Westpac) for not passing on the full interest rate cuts by the Reserve Bank of Australia to their borrowers.
- c The mayor of a city council in Queensland would like to know the demographic profile of the community living in his city council area.

Solution

- a To survey the voters in New Zealand, a telephone survey would be the most economical and practical way of gathering data.
- b Information on the reasons for the actions of the four major banks on interest rate cuts can only be obtained by face-to-face interviews with bank management so that responses can be obtained without ambiguity.
- c Statistics on the demographic and other characteristics of the community living in the city council area can be obtained from primary sources such as ABS Census, which contains reliable data on these characteristics. Alternatively, a self-administered survey could be done by mailing out a questionnaire to randomly selected individuals in the city council area. However, one should be cautious about the conclusions, as the response rate may be low.

EXERCISES

- 2.11** Briefly describe the difference between primary data and secondary data.
- 2.12** For each of the following data sources, determine the frequency of one of their publications and write down two specific pieces of information contained in the latest published version.
- The Australian Bureau of Statistics
 - Reserve Bank Bulletin
 - CIA Factbook
- 2.13** Describe the difference between observational data and experimental data.
- 2.14** A soft-drink manufacturer has been supplying its cola drink in bottles to grocery stores and in cans to small convenience stores. The company is analysing sales of this cola drink to determine which type of packaging is preferred by consumers.
- a** Is this study observational or experimental?
Explain your answer.
- b** Outline a better method for determining whether a store will be supplied with cola in bottles or in cans, so that future sales data will be more helpful in assessing the preferred type of packaging.
- 2.15** **a** Briefly describe how you might design a study to investigate the relationship between smoking and lung cancer.
b Is your study from part (a) observational or experimental? Explain why.
- 2.16** **a** List three methods of conducting a survey of people.
b Give an important advantage and disadvantage of each of the methods listed in part (a).
- 2.17** List five important points to consider when designing a questionnaire.

2.3 Sampling

The chief motive for examining a sample rather than a population is cost and practicability. Statistical inference permits us to draw conclusions about a population parameter based on a sample that is quite small in comparison to the size of the population. For example, television executives want to know the proportion of television viewers who watch their network's programs. Since six to eight million people may be watching television in Australia on a given evening, determining the actual proportion of the population watching certain programs is impractical and prohibitively expensive. Some weekly magazines provide approximations of the desired information by observing what is watched by a sample of 1000 television viewers. The proportion of households watching a particular program can be calculated for the households in the sample. This sample proportion is then used as an estimate of the proportion of all households (population proportion) that watched the program.

Another illustration of sampling can be taken from the field of quality management. In order to ensure that a production process is operating properly, the operations manager needs to know the proportion of defective units that are being produced. If the quality-control technician must destroy the unit in order to determine whether or not it is defective, there is no alternative to sampling: a complete inspection of the population would destroy the entire output of the production process.

We know that the sample proportion of television viewers or of defective items is probably not exactly equal to the population proportion we want it to estimate. Nonetheless, the sample statistic can come quite close to the parameter it is designed to estimate, if the **target population** (the population about which we want to draw inferences) and the **sampled population** (the population from which we have actually taken a sample) are the same. In practice, these may not be the same.

For example, the magazines' ratings are supposed to provide information about the television shows that all Australians are watching. Hence, the target population is the television viewers of Australia. If the sample of 1000 viewers were drawn exclusively from the state of New South Wales, however, the sampled population would be the television viewers of

target population

The population about which we want to draw inferences.

sampled population

The actual population from which the sample has been drawn.

New South Wales. In this case, the target population and the sampled population are not the same, and no valid inferences about the target population can be drawn. To allow proper estimation of the proportion of all Australian television viewers watching a specific program, the sample should contain men and women from each state and territory of varying ages, incomes, occupations and residences in a pattern similar to that of the target population, which in this case is all Australian television viewers. The importance of sampling from the target population cannot be overestimated, since the consequences of drawing conclusions from improperly selected samples can be costly.

The *Literary Digest* was a popular US magazine of the 1920s and 1930s, which had correctly predicted the outcomes of several US presidential elections. In 1936, the *Digest* predicted that the Republican candidate, Alfred Landon, would defeat the Democrat incumbent, Franklin D. Roosevelt, by a 3 to 2 margin. But in that election, Roosevelt won a landslide victory, garnering the support of 62% of the electorate. The source of this blunder was the sampling procedure: the *Digest* sent out 10 million sample ballots to prospective voters, but only 2.3 million ballots were returned, resulting in a self-selected sample.

Self-selected samples are almost always biased, because the individuals who participate in them are more keenly interested in the issue than the other members of the population. Similar surveys are conducted today, when radio and television stations ask people to call and give their opinions on an issue of interest. Again, only those who are concerned about the topic and have enough patience to get through to the station will be included in the sample. Hence, the sampled population is composed entirely of people who are interested in the issue, whereas the target population is made up of all the people within the listening/watching radius of the radio/television station. As a result, the conclusions drawn from such surveys are frequently wrong. Unfortunately, because the true value of the parameter being estimated is never known (unlike the situation in a political survey, where the election provides the true parametric value), these surveys give the impression of providing useful information. In fact, the results of such surveys are likely to be no more accurate than the results of the 1936 *Literary Digest* poll.

A recent example of the sampling error is the predicted and actual results of the Australian Federal parliamentary election held on 18 May 2019. Until the day of the election, several media outlets and pollsters were predicting an easy win for the opposition Labor party. However, the results turned out to be a clear win for the incumbent government of the Liberal–National coalition. Since the 1980s, telephone polling was the main method of public opinion polls, and sources such as the *White Pages* was the main source of sample selection. However, since mobile phones began to displace household fixed-line phones, and there is no longer a comprehensive list of mobile numbers and fixed-line numbers, selecting a genuinely random sample has become a challenge. This is one of the main reasons for the complete failure of the 2019 Australian election outcome prediction.

In the next section, we discuss a number of different ways in which populations can be surveyed. In all cases, we assume that the surveys are properly performed and that the target population and the sampled population are very similar.

EXERCISES

- 2.18** For each of the following sampling plans, indicate why the target population and the sampled population are not the same.

- a In order to determine the opinions and attitudes of customers who regularly shop at a particular centre, a surveyor stands outside a large department store in the centre and randomly selects people to participate in the survey.
- b A library wishes to estimate the proportion of its books that have been damaged. They decide

to select one book per shelf as a sample, by measuring 30 cm from the left edge of each shelf and selecting the book in that location.

- c A political surveyor visits 200 residences and asks the eligible voters present in the house at the time whom they intend to vote for. The visits take place during the afternoons.

2.4 Sampling plans

The objective in this section is to introduce several different sampling plans, namely, simple random sampling, stratified random sampling and cluster sampling. We begin the presentation with the most basic design.

2.4a Simple random sampling

simple random sample

One in which each element of the population has an equal chance of appearing.

One way to conduct a **simple random sample** is to assign a number to each element in the population, write these numbers on individual slips of paper, toss them into a hat, and draw the required number of slips (the sample size, n) from the hat. This procedure is the kind used in raffles, when all the ticket stubs go into a large rotating drum from which the winners are selected.

Sometimes the elements of the population are already numbered. For example, almost all adults have tax file numbers; all employees of large corporations have employee numbers; many people have driver's licence numbers, Medicare card numbers, student numbers, and so on. In such cases, choosing the procedure to use is simply a matter of deciding how to select from among these numbers.

In other cases, the existing form of numbering has built-in flaws that make it inappropriate as a source of samples. Not everyone has a phone number, for example, so the telephone book does not list all the people in a given area. Many households have two (or more) adults, but only one phone listing. Couples often list the phone number under the man's name, so telephone listings are likely to be disproportionately male. Some people do not have phones, some have unlisted phone numbers, and some have more than one phone; these differences mean that each element of the population does not have an equal probability of being selected.

Once each element of the chosen population has been assigned a unique number, sample numbers can be selected at random. It is usual to employ a computer-generated random-number table, such as **Table 7 in Appendix B**, for this purpose. Alternatively, we can use Excel to perform this function.

EXAMPLE 2.3

L04 L05

Random sample of outstanding credit-account balances

A department-store audit involves checking a random sample from a population of 30 outstanding credit-account balances. The 30 accounts are listed in the following table. Calculate the population average. Use a random-number table such as Table 7 in Appendix B to select five accounts at random. Calculate the average credit-account balance of the sample you have selected.

Account no.	Balance	Account no.	Balance	Account no.	Balance
1	25	11	918	21	159
2	0	12	801	22	489
3	605	13	227	23	115
4	1010	14	0	24	27
5	527	15	47	25	27
6	34	16	0	26	291
7	245	17	102	27	16
8	59	18	215	28	0
9	67	19	429	29	402
10	403	20	197	30	17





Solution

$$\text{Population average } \mu = \frac{25 + 0 + 605 + \dots + 402 + 17}{30} = 248.47$$

Going to our random number table, we select row 1, column 8 as our starting point. We shall go down that column, selecting the first two digits as our random numbers. The random numbers are reproduced here for convenience.

Random number	Random number
22✓	19✓
17✓	51
83	39
57	59
27✓	84
54	20✓

Notice that we had to select more than five numbers, as some of them were greater than 30 (account numbers are labelled from 1 to 30 only). The following five accounts, therefore, are to be audited:

Random number (account number)	Balance (\$)
22	489
17	102
27	16
19	429
20	197

$$\text{Sample average} = \frac{489 + 102 + 16 + 429 + 197}{5} = 246.6$$

Note that, in this example, the sample average and the population average are very close to each other.

EXAMPLE 2.4

L04 L05

Random sample of income tax returns

A government income-tax auditor has been given responsibility for 1000 tax returns. A computer is used to check the arithmetic of each return. However, to determine if the returns have been completed honestly, the auditor must check each entry and confirm its veracity. Because it takes, on average, one hour to completely audit a return and she has only one week to complete the task, the auditor has decided to randomly select 40 returns. The returns are numbered from 1 to 1000. Use a computer random-number generator to select the sample for the auditor.

Solution

There are several software packages that can produce the random numbers we need. Excel is one of these.

Using the computer

Excel Data Analysis

We generated 50 numbers between 1 and 1000 and stored them in column 1. Although we needed only 40 random numbers, we generated 50 numbers because it is likely that some of them will be duplicates. We will use the first 40 unique random numbers to select our sample. Notice that the number 467 is generated twice.

Excel output for Example 2.4: Computer-generated random numbers

	A	B	C	D	E
1	383	246	372	952	75
2	101	46	356	54	199
3	597	33	911	706	65
4	900	165	467	817	359
5	885	220	427	973	488
6	959	18	304	467	512
7	15	286	976	301	374
8	408	344	807	751	986
9	864	554	992	352	41
10	139	358	257	776	231

COMMANDS

- 1 Click **DATA, Data Analysis** (in the Analysis submenu), and select **Random Number Generation** from the **Analysis Tools** drop-down menu. Click **OK**.
- 2 Type the **Number of Variables (1)**.
- 3 Type **Number of Random Numbers (50)**.
- 4 Select the Distribution (**Uniform**).
- 5 Specify the range of the uniform distribution **Parameters (0 and 1)**. Under **Output Options**, select **Output Range:**. Then type the starting cell reference for the **output range (A1)**. Click **OK**. Column A will fill with 50 numbers that range between 0 and 1.
- 6 Make cell B1 active. Multiply cell A1 by 1000 and store in cell B1. Complete column B entries by copying the formula in cell B1 to cells B2 to B50. (**=A1*1000**)
- 7 Make cell C1 active, click **fx**, and select **Math & Trig** from the drop-down menu of categories. In the list of functions that appear below that, select **ROUNDUP** and then click **OK**.
- 8 Specify the first number to be rounded (**B1**).
- 9 Hit **tab** and type the **Number of Digits** (decimal places) (**0**). Click **OK**. Complete column C by copying the formula in cell C1 into cells C2–C50.

The first five steps command Excel to generate 50 uniformly distributed random numbers between 0 and 1 to be stored in column A. Steps 6 to 9 convert these random numbers to integers between 1 and 1000. Each number has the same probability ($1/1000 = 0.001$) of being selected. Thus, each member of the population is equally likely to be included in the sample.

Interpreting the results

The auditor would examine the tax returns selected by the computer. Using the Excel output, the auditor would pick returns numbered 383, 101, 597, ..., 352, 776 and 75 (the first 40 unique numbers). Each of these tax returns would be audited to determine if they were fraudulent. If the objective is to audit these 40 returns, no statistical procedure would be employed. However, if the objective is to estimate the proportion of *all* 1000 returns that were dishonest, the auditor would use one of the inferential techniques that are presented later in this book.

2.4b Stratified random sampling

In making inferences about a population, we attempt to extract as much information as possible from a sample. The basic sampling plan, simple random sampling, often accomplishes this goal at low cost. Other methods, however, can be used to increase the amount of information about the population. One such procedure is stratified random sampling. A **stratified random sample** is obtained by dividing the population into mutually exclusive sets, or strata, and then drawing simple random samples from each stratum.

Examples of criteria for dividing a population into strata (and the strata themselves) are as follows:

- 1** *Gender:* Male
Female
- 2** *Age:* under 20
20–30
31–40
41–50
51–60
over 60
- 3** *Religion:* Christianity
Islam
Buddhism
Hinduism
Other
- 4** *Household income:* under \$30 000
\$30 000–\$59 999
\$60 000–\$79 999
\$80 000 and over

stratified random sample

One in which the population is separated into mutually exclusive layers, or strata, from which simple random samples are drawn.

To illustrate, suppose a public opinion survey is to be conducted in order to determine how many people favour proposed changes to the Medicare scheme. A stratified random sample could be obtained by selecting a random sample of people from each of the four income groups described above. We usually stratify in a way that enables us to obtain particular kinds of information. In this example, we would like to know if people in the different income categories differ in their opinions about the proposed changes to Medicare, since the changes will affect each stratum differently. We avoid stratifying when there is no connection between the survey and the strata. For example, little purpose is served in trying to determine if people within religious strata have divergent opinions about the proposed changes to Medicare.

One advantage of stratification is that, as well as acquiring information about the entire population, we can also make inferences within each stratum or compare strata. For instance, we can estimate what proportion of the lowest income group favours the proposed changes, or we can compare the highest and lowest income groups to determine if they differ in their support of the proposed changes.

Any stratification must be done in such a way that the strata are mutually exclusive: each member of the population must be assigned to exactly one stratum.

After the population has been stratified in this way, we can employ simple random sampling to generate the complete stratified random sample. There are several ways to do this. For example, we can draw random samples from each of the four income groups according to their proportions in the population. Thus, if in the population the relative frequencies of the four groups are as listed below, our sample will be stratified in the same proportions. If a total sample of 1000 is to be drawn, we will randomly select 250 from stratum 1, 400 from stratum 2, 300 from stratum 3, and 50 from stratum 4.

Stratum	Income categories	Population proportions (%)	Stratified random sample of size 1000
1	Under \$30 000	25	250
2	\$30 000–\$59 999	40	400
3	\$60 000–\$89 999	33	330
4	\$90 000 and over	2	20

The problem with this approach, however, is that if we want to make inferences about the last stratum, a sample of 20 may be too small to produce useful information. In such cases, we usually increase the sample size of the smallest stratum (or strata) to ensure that the sample data provide enough information for our purposes. An adjustment must then be made before we attempt to draw inferences about the entire population. This procedure is beyond the level of this book. We recommend that anyone planning such a survey consult an expert statistician or a reference book on the subject ‘Sampling Methods’. Better still, become an expert statistician yourself by taking additional statistics subjects.

2.4c Cluster sampling

cluster sample

A simple random sample of clusters, or groups, of elements.

Cluster sampling is particularly useful when it is difficult or costly to develop a complete list of the population members (making it difficult and costly to generate a simple random sample). It is also useful whenever the population elements are widely dispersed geographically. For example, suppose that we wanted to estimate the average annual household income in a large city. In order to use simple random sampling, we would need a complete list of households in the city from which to sample. To use stratified random sampling, we would again need the list of households, and we would also need to have each household categorised by some other variable (such as age of household head) in order to develop the strata. A less expensive alternative would be to let each block within the city represent a cluster. A sample of clusters could then be randomly selected, and every household within these clusters could be interviewed to determine income. By reducing the distances the surveyor must cover to gather the data, cluster sampling reduces the cost.

But cluster sampling also increases sampling error (see Section 2.5), because households belonging to the same cluster are likely to be similar in many respects, including household income. This can be partially offset by using some of the cost savings to survey a larger sample than would be used for a simple random sample.

2.4d Sample size

Whichever type of sampling plan you select, you still have to decide what size of sample to use. Determining the appropriate sample size will be addressed in detail in Chapters 10 and 11. Until then, we can rely on our intuition, which tells us that the larger the sample size is, the more accurate we can expect the sample estimates to be.

EXAMPLE 2.5

LO5

Describing a suitable sampling plan

Discuss the appropriate sampling plan one would choose to collect data for the following statistical analyses.

- A political analyst would like to analyse the voting intentions of New Zealand voters among the political parties: National, Labour, Green, New Zealand First, Māori and Independents/Other.
- A researcher wants to gather the opinions from adults on legalising marijuana use in Australia.
- A statistician wants to estimate the average age of children in his city. Unfortunately, he does not have a complete list of households.

Solution

- To survey the voters in New Zealand, the random sample could be selected using the simple random sampling method. However, if the intention were also to analyse the voting intentions of male versus female voters, we would use the stratified random sampling method. Other examples in this experiment using stratified random sampling could be to compare the income groups, migrant versus non-migrants, first-time versus other voters, etc.
- Obviously, we would expect, the opinion would differ among different age groups. A stratified random sample of adults from different age groups can be obtained by selecting a simple random sample from the 18–25 age group, another simple random sample from the 25–40 age group and another simple random sample from the over 40 age group. That is, the three age groups 18–25, 25–40 and over 40 represent three strata from which we obtain simple random samples.
- A less expensive way of obtaining sample data would be to divide the city into clusters. A sample of clusters could then be randomly selected, and the age of every child in these clusters could be obtained by interviewing a member of each household.

We now discuss the solution to the opening example in this chapter.

SPOTLIGHT ON STATISTICS**Sampling and the census: Solution**

To adjust for undercounting in Australia, the ABS adjusts the numbers it gets from the census. The adjustment is based on a survey called the Post Enumeration Survey (PES) using a multi-stage cluster sampling method.

The survey is conducted about three weeks after the census and its purpose is to determine how many people were missed in the census and how many were counted more than once.

The 2016 PES was conducted during the period from September to October, after the census fieldwork had been completed in August 2016. In each selected household, a responsible adult member was interviewed and asked about all persons present or usually resident in the household. In addition to obtaining basic demographic information, questions were asked about each person's usual residence, location on census night and any other addresses at which they might have been counted in the census. Using this address information, the corresponding census forms were examined at the processing centre to confirm how many times each person in the PES was counted in the census.

In the 2016 PES, a sample of about 43 100 private dwellings, including 600 dwellings selected from around 33 discrete communities in New South Wales (NSW), Victoria (Vic), Queensland (Qld), South Australia (SA), Western Australia (WA), Tasmania (Tas), Northern Territory (NT) and Australian Capital Territory (ACT) were enumerated. The distribution of the expected number of fully responding households is as follows:

NSW	Vic	Qld	SA	WA	Tas	NT	ACT
9400	8200	7900	4700	5200	2600	3400	1700



Source: Shutterstock.com/Vlada Young



The PES was used to produce counts of the number of people who should have been counted in the census and the number who were actually counted. The ratio of these two numbers represents the net adjustment factor – the amount by which census counts must be adjusted to allow for undercount. The PES adjustment factor was weighted to take into account the chance of people being selected in the PES. The weighted adjustment factor was then applied to the census count to produce an initial estimation of the population. This net adjustment factor takes into account both the people missed by the census and the people counted more than once. It was calculated and applied separately for each state and territory by age, group and gender.

The estimation procedure is illustrated in the equation below:

$$\text{Adjustment factor} = \frac{\text{No. in PES who should have been counted in the census}}{\text{No. in PES who were counted in the census}}$$

$$\text{Estimate} = \text{Census count} \times \text{Weighted adjustment factor}$$

Once initial population estimates have been calculated from the PES, a second stage of estimation takes place using demographic methods. The initial population estimates by age and by gender are then compared with data on the Australian population derived largely from records of births and deaths and overseas arrivals and departures. In 2016, the net undercount for the final population of Australia was 226 407 people. The PES population estimate and the net undercount for different states based on the 2016 census are given in the table.

PES Population estimates, census and undercount, 2016

State/Territory	Census count	Net undercount	PES population estimate
New South Wales	7480220	59194	7 539 414
Victoria	5926593	86028	6 012 621
Queensland	4703210	60550	4 763 760
South Australia	1676633	2974	1 679 607
Western Australia	2474424	9477	2 483 901
Tasmania	509965	296	510 261
Northern Territory	228854	12 020	240 874
Australian Capital Territory	397397	-4 132	393 265
Australia	23 397 296	226 407	23 623 703

Source: Australian Bureau of Statistics, *Census of Population and Housing – Details of Undercount*, Australia 2016, Table 2, cat. no. 2940.0, ABS, Canberra.

EXERCISES

- 2.19** A statistician would like to conduct a survey to ask people their views on a proposed new shopping mall in their community. According to the latest census, there are 800 households in the community. The statistician has numbered each household (from 1 to 800), and would like to randomly select 25 of these households to participate in the study. Use a software package to generate the sample.

- 2.20** A safety expert wants to determine the proportion of cars in his state with worn tyre treads. The state licence plate contains three digits. Use a software package to generate a sample of 20 cars to be examined.

- 2.21** The operations manager of a large plant with four departments wants to estimate the employee working hours lost per month due to accidents.

Describe a sampling plan that would be suitable for estimating the plant-wide loss and for comparing departments.

- 2.22** A large university campus has 60 000 students. The president of the student association wants to conduct a survey of the students to determine their views on an increase in the student activity fee. She would like to acquire information about all the students but would also like to compare the school of business, the faculty of arts and sciences, and the graduate school. Describe a sampling plan that accomplishes these goals.

- 2.23** A telemarketing firm has recorded a list of the households that have purchased one or more of a company's products. These data number in the millions. They would like to conduct a survey of purchasers to obtain information about their attitude concerning the timing of the telephone calls. The president of the company would like to know the views of all purchasers, but would also like to compare the attitudes of people in the west, south, north and east. Describe a suitable sampling plan.

2.5 Sampling and non-sampling errors

Two main types of error can arise when a sample of observations is taken from a population: sampling error and non-sampling error. Managers reviewing the results of sample surveys and studies, as well as researchers who conduct the surveys and studies, should understand the sources of these errors.

2.5a Sampling error

Sampling error refers to differences between the sample and the population that exist only because of the observations that happened to be selected for the sample. Sampling error is an error that we expect to occur when we make a statement about a population that is based only on the observations contained in a sample taken from the population.

To illustrate, consider an example in which we wish to determine the average annual income of Australian blue-collar workers. To do this we would have to ask each Australian blue-collar worker their income, and then calculate the average of all the responses. Because this population consists of several million people, the task is both expensive and impractical. If we are willing to accept less than 100% accuracy, we can use statistical inference to estimate the average income (μ) of the population. If we record the incomes of a sample of the workers and find the average of this sample of incomes (\bar{x}), this sample average is an estimate of the desired population average. But the value of \bar{x} will deviate from the population average (μ) simply by chance, because the value of the sample average depends on which incomes just happened to be selected for the sample. The difference between the true (unknown) value of the population average (μ) and its sample estimate (\bar{x}) is the sampling error. The size of this deviation may be large simply due to bad luck – bad luck that a particularly unrepresentative sample happened to be selected. The only way we can reduce the expected size of this error is to take a larger sample.

Given a fixed sample size, the best we can do is to state the probability that the sampling error is less than a certain amount (as we will discuss in Section 10.5). It is common today for such a statement to accompany the results of an opinion poll. If an opinion poll states that, based on sample results, candidate Kreem has the support of 54% of eligible voters in an upcoming election, that statement may be accompanied by the following explanatory note: This percentage is correct to within three percentage points, 19 times out of 20. This statement means that we estimate that the actual level of support for the candidate is between 51% and 57%, and that, in the long run, this type of procedure is correct 95% of the time.

2.5b Non-sampling error

Non-sampling error is more serious than sampling error, because taking a larger sample won't diminish the size, or the possibility of occurrence, of this error. Even a census can (and probably will) contain non-sampling errors. Non-sampling errors are due to mistakes made in the acquisition of data or due to the sample observations being selected improperly.

Three types of non-sampling errors

- 1** *Errors in data acquisition.* These types of errors arise from the recording of incorrect responses. This may be the result of incorrect measurements being taken because of faulty equipment, mistakes made during transcription from primary sources, inaccurate recording of data due to misinterpretation of terms, or inaccurate responses to questions concerning sensitive issues such as sexual activity or possible tax evasion.
- 2** *Non-response error.* This refers to error (or *bias*) introduced when responses are not obtained from some members of the sample. When this happens, the sample observations that are collected may not be representative of the target population, resulting in biased results (as was discussed in Section 2.2). Non-response can occur for a number of reasons. An interviewer may be unable to contact a person listed in the sample, or the sampled person may refuse to respond for some reason. In either case, responses are not obtained from a sampled person, and bias is introduced. The problem of non-response error is even greater when self-administered questionnaires are used rather than an interviewer, who can attempt to reduce the non-response rate by means of call backs. As noted earlier in this chapter, the *Literary Digest* fiasco in the United States was largely due to a high non-response rate, resulting in a biased, self-selected sample.
- 3** *Selection bias.* This occurs when some members of the target population cannot possibly be selected for inclusion in the sample. Together with non-response error, selection bias played a role in the *Literary Digest* poll being so wrong, as voters without telephones or without a subscription to *Literary Digest* were excluded from possible inclusion in the sample taken.

EXERCISES

- | | |
|---|--|
| 2.24 a Explain the difference between sampling error and non-sampling error.
b Which type of error in part (a) is more serious? Why? | 2.25 Briefly describe three types of non-sampling error.
2.26 Is it possible for a sample to yield better results than a census? Explain. |
|---|--|

Study Tools

CHAPTER SUMMARY

Statistics is involved with dealing with data and techniques used to summarise, extract information and analyse data. To decide which technique to use, it is important to know the type of data. There are three major types of data: numerical, nominal and ordinal.

Because most populations are very large, it is extremely costly and impractical to investigate each member of the population to determine the value of the parameters. As a practical alternative, we *sample* the population and use the sample statistics to draw inferences about the population parameters. Care must be taken to ensure that the *sampled population* is the same as the *target population*.

We can choose from among several different sampling plans, including *simple random sampling*, *stratified random sampling* and *cluster sampling*. Whatever sampling plan is used, it is important to realise that both *sampling error* and *non-sampling error* will occur, and to understand what are the sources of these errors.

The data files for Examples and Exercises are provided on the companion website (accessible through <https://login.cengagebrain.com/>).

PART ONE

Descriptive measures and probability

CHAPTER 3	Graphical descriptive techniques – Nominal data
CHAPTER 4	Graphical descriptive techniques – Numerical data
CHAPTER 5	Numerical descriptive measures
CHAPTER 6	Probability
CHAPTER 7	Random variables and discrete probability distributions
CHAPTER 8	Continuous probability distributions

To help you organise the material that you are about to learn, we have divided the rest of the book into three parts.

Part 1 covers descriptive statistics and probability. These topics constitute the foundation of statistical inference. Chapter 3 introduces the graphical techniques for nominal data and Chapter 4 deals with graphical techniques for numerical data. Chapter 5 presents numerical measures that are used to summarise data. The summary measures introduced in Chapter 5 will be used to make inferences about parameters in later chapters. In Chapters 6 to 8, we present probability and probability distributions that will provide the link between sample statistics and population parameters.

Everything we do in this book is mostly built upon these six chapters. However, Part 1 does much more than just lay the foundation. Both descriptive statistics and probability are subjects that are worth learning for their own intrinsic values.

We all make decisions on a daily basis, most of which are made under uncertainty. Consider an investor who must decide which investment to make, how much money to invest and for how long that investment should be held. There are a large number of events over which the investor has no control. All that the investor can do is attempt to assess the risks and returns associated with each investment. As you will discover, probability plays a central role in this assessment.

We believe that all business and economics graduates will have many opportunities to apply statistical inference techniques and concepts. However, not all of them will do so because of a lack of either knowledge (despite the best efforts of statistics lecturers) or confidence. Descriptive techniques are so common that it is virtually impossible to ignore them. Newspapers, magazines, company annual reports and presentations are filled with applications of descriptive statistics. Knowing how to use and interpret them is a critical skill for all of us.

Graphical descriptive techniques – Nominal data

Learning objectives

This chapter discusses the graphical descriptive methods used to summarise and describe sets of nominal data.

At the completion of this chapter, you should be able to:

- L01** construct charts to summarise nominal data
- L02** use Excel to draw appropriate charts for nominal data
- L03** determine which chart is best for nominal data under a given circumstance
- L04** use charts to describe ordinal data
- L05** use various tabular and graphical techniques to analyse the relationships between two nominal variables.

CHAPTER OUTLINE

Introduction

3.1 Graphical techniques to describe nominal data

3.2 Describing the relationship between two nominal variables

SPOTLIGHT ON STATISTICS

Break bail, go to jail?

XM03-00 An overwhelming majority of Victorian electors (78%) say people charged with a criminal offence who are given bail and then break a bail condition should be immediately sent to jail, according to a special SMS Morgan Poll conducted on the eve of the last Victorian state election.

Victorian electors were asked: 'Many people charged with a criminal offence are given bail. If a person given bail then breaks a bail condition, should that person be immediately sent to jail or not?' This special SMS Morgan Poll was conducted on Thursday 22 November 2018 with a statewide cross-section of 961 Victorian electors aged 18 and over. The responses and party affiliation for a random sample of 200 respondents are stored in file **CH03:XM03-00**. Some of the data are listed below. Determine whether the responses differ on the basis of party affiliation. On pages 71–2 we provide a possible answer.

Source: www.roymorgan.com.au, Finding no: 7813

ID	Party	Response
1	LNP	Yes
2	LNP	Yes
3	LNP	Yes
4	Others	Yes
.		
.		
.	
199	Others	No
200	ALP	Yes



Source: Shutterstock.com/Sentavio

Introduction

In Chapter 1, we pointed out that statistics is divided into two basic areas: (1) descriptive statistics and (2) inferential statistics. In Chapter 2 we presented the various types of data, sampling and survey methods. The purpose of this and the next two chapters is to present the principal methods that fall under the heading of descriptive statistics. In this and the next chapter, we introduce graphical and tabular statistical methods that allow managers to summarise data visually in order to produce useful information – a technique often used in decision making – and discuss ways to use the techniques introduced in an effective and accurate way. In Chapter 5, we introduce another class of descriptive statistical techniques – numerical measures.

Managers frequently have access to large amounts of potentially useful data. But before the data can be used to support a decision, the data must be organised and summarised. Consider, for example, the problems faced by managers who have access to the databases created by the use of credit cards. The database consists of the personal information supplied by the customer when he or she applied for the credit card. This information includes age, gender, place of residence and income of the cardholder. In addition, each time the card is used, the database grows to include a history of the timing, price and brand of each product so purchased. Using the appropriate statistical technique, managers can determine which segments of the consumer market are buying their company's brands. Specialised marketing campaigns, including telemarketing, can be developed. Both *descriptive statistics* and *inferential statistics* would likely be employed in the analysis.

Descriptive statistics involves arranging, summarising and presenting a set of data in such a way that the meaningful essentials of the data can be extracted and grasped easily. Its methods make use of *graphical techniques* (such as pie charts and histograms) and numerical descriptive measures (such as averages) to summarise and present the data in a meaningful way. Although descriptive statistical methods are relatively straightforward, their importance should not be underestimated. Most management, marketing, business and economics students will encounter numerous opportunities to make valuable use of graphical and numerical tools in descriptive statistics when preparing reports and presentations in the workplace. According to a Wharton Business School study in the United States, top managers reach a consensus 25% more quickly when responding to a presentation in which graphics are used.

In Chapter 1, we introduced the distinction between a population and a sample. Recall that a *population* is the entire set of observations or measurements under study, whereas a *sample* is a subset of observations selected from the entire population. The descriptive methods presented in this and the next two chapters apply equally to both a set of data constituting a population and a set of data constituting a sample.

In both the preface and Chapter 1 we pointed out that a critical part of your education as statistics practitioners includes an understanding not only of how to draw graphs and calculate statistics (manually or by computer), but also when to use each technique that we cover and how to interpret the results. The two most important factors that determine the appropriate method to use are the type of data and the information that is needed. In Section 2.1, we discussed the various types of data and how to identify them. In this chapter, we introduce graphical techniques used to describe a set of nominal (categorical) data in Section 3.1, and in Section 3.2 we discuss how to select the most appropriate technique to present a set of nominal data. In Section 3.3 we introduce the presentation of ordinal data. In Section 3.4, we present graphical techniques to describe the relationship between two nominal variables and compare two or more sets of nominal data. Chapter 4 presents some graphical techniques to describe numerical (quantitative) variables.

3.1 Graphical techniques to describe nominal data

frequency distribution

Method of presenting data and their counts in each category or class.

relative frequency distribution

Frequency distribution giving the percentage each category or class represents of the total.

bar chart

A chart in which vertical bars represent data in different categories.

pie chart

A circle subdivided into sectors representing the share of data in different categories.

As discussed in Section 2.1 (page 22), the only allowable calculation on *nominal (categorical) data* is to count the frequency of each value of the variable. We can then summarise the data in a table called a **frequency distribution** table that presents the categories and their counts. A **relative frequency distribution** table lists the categories and the proportion in which each category occurs. We can use graphical techniques to present a picture of the data. There are two graphical techniques we can use: the bar chart and the pie chart.

3.1a Bar and pie charts

Graphical techniques generally catch a reader's eye more quickly than does a table of numbers. The two most popular graphical representations to describe nominal data are the *bar chart* and the *pie chart*. A bar chart is often used to display frequencies, and a pie chart is used to show relative frequencies or proportions. Bar and pie charts are used widely in newspapers and magazines, as well as business and government reports.

A **bar chart** graphically represents the frequency of each category as a bar rising vertically from the horizontal axis; the height of each bar is proportional to the frequency of the corresponding category. Because the bars correspond to categories, the base widths assigned to the bars are arbitrary, although all must be equal. To improve clarity, a space is usually left between bars.

A **pie chart** is a circle that is subdivided into slices. The area of each slice is proportional to the frequency (or relative frequency) of occurrences of each category. Because the entire circle (100%) corresponds to 360° , every 1% of the observations should correspond to $\frac{360}{100} = 3.6^\circ$.

To illustrate the applications of bar charts and pie charts, consider the following example.

EXAMPLE 3.1

L01 L02

Women's magazine readership survey in New Zealand

XM03-01 A magazine readership survey carried out in New Zealand shows that women's magazines are the most popular magazines, having the largest readership and increasing yearly sales. The survey results of 300 readers were recorded and are given below in coded form. The top six magazines considered here are (1) *Australian Women's Weekly (NZ Edition)*, (2) *NZ Woman's Weekly*, (3) *NZ Woman's Day*, (4) *New Idea*, (5) *Next* and (6) *That's Life*. The data, using the codes 1, 2, 3, 4, 5 and 6, are listed below. Create the frequency distribution table and construct a bar chart and a pie chart to summarise the data.

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



Solution

Identifying the technique

The data are nominal because the ‘values’ of the variable, names of magazines, are the six categories.

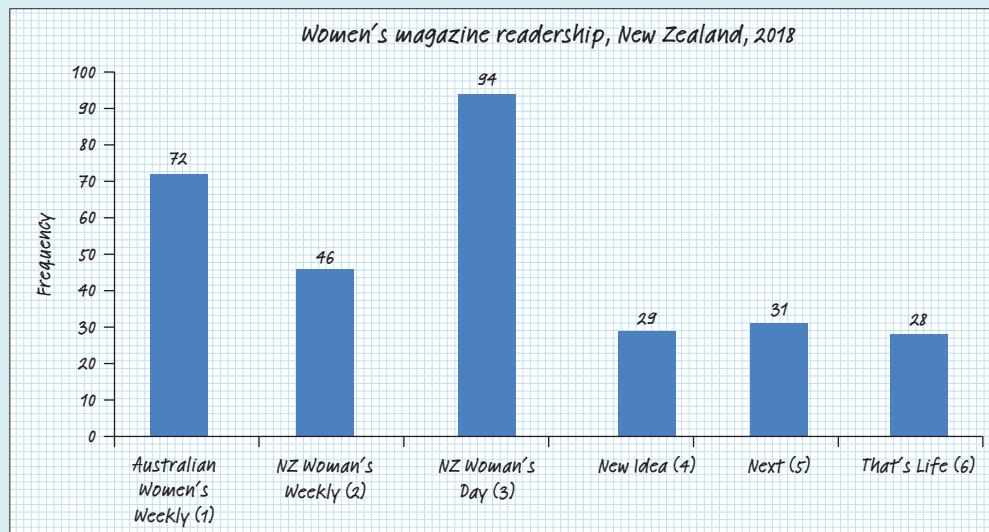
The numbers (1, 2, 3, 4, 5 and 6) used to record the data in the file were assigned completely arbitrarily. The only legitimate statistical technique is to count the number of occurrences (also known as frequencies) of each value and then convert these counts to proportions. The results are shown in **Table 3.1**.

Using the frequency distribution, we first construct a bar chart by drawing a rectangle representing each category. The height of the bar represents either the frequency or relative frequency. **Figure 3.1** depicts the manually drawn bar chart for the magazine readership survey data.

TABLE 3.1 Frequency and relative frequency distributions for the readership of the most popular women’s magazines in New Zealand, 2018

Magazine	Number of readers	Proportion of readers (%)
<i>Australian Women’s Weekly</i> (1)	72	24.0
<i>NZ Woman’s Weekly</i> (2)	46	15.3
<i>NZ Woman’s Day</i> (3)	94	31.3
<i>NZ New Idea</i> (4)	29	9.7
<i>Next</i> (5)	31	10.3
<i>That’s Life</i> (6)	28	9.3
Total	300	100.0

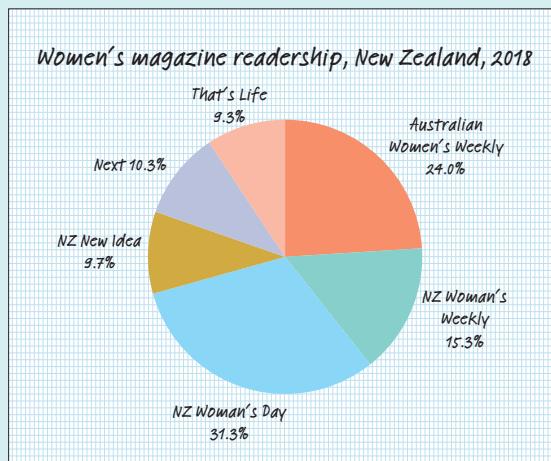
FIGURE 3.1 Bar chart for Example 3.1



The pie chart exhibits the proportion or percentage of readers of each magazine. As the size of each slice of a circle is proportional to the percentage corresponding to that category, the angle between the lines demarcating the *Australian Women’s Weekly* (NZ Edition) readers, for example, is $24.0 \times 3.6 = 86.4^\circ$. The angles of the pie chart for the other five categories are calculated similarly.

Magazine	Proportion of readers (in percentages)	Angle of the slice
Australian Women's Weekly (1)	24.0	$24.0 \times 3.6 = 86.4^\circ$
NZ Woman's Weekly (2)	15.3	$15.3 \times 3.6 = 55.2^\circ$
NZ Woman's Day (3)	31.3	$31.3 \times 3.6 = 112.8^\circ$
New Idea (4)	9.7	$9.7 \times 3.6 = 34.8^\circ$
Next (5)	10.3	$10.3 \times 3.6 = 37.2^\circ$
That's Life (6)	9.3	$9.3 \times 3.6 = 33.6^\circ$
Total	100.0	360°

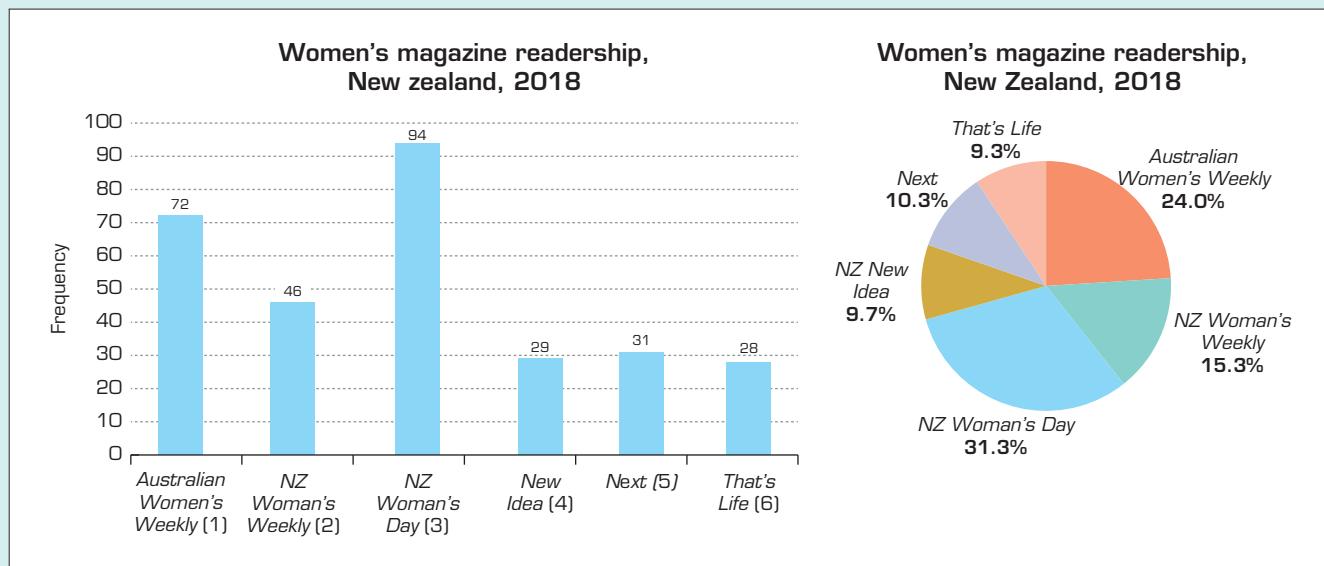
FIGURE 3.2 Pie chart for Example 3.1



Now that you know how to construct bar and pie charts by hand, let's see how we actually draw one in practice using a computer.

Using the computer

Excel bar chart and pie chart for Example 3.1





The following are the Excel commands to draw a bar chart.

COMMANDS

Bar Chart

If you only have access to the raw data (e.g. the data in file **XM03-01**), proceed through the following steps to obtain the frequencies.

- 1 Import the data or Open file **XM03-01**.
- 2 Leave cell B1 as blank. In cells B2–B7, type the names of the six magazines (**Australian Women's Weekly; NZ Woman's Weekly; NZ Woman's Day; New Idea; Next; That's Life**).
- 3 In cell C1, type the title **Frequency**. In cells C2–C7, type '**=COUNTIF(range, criteria)**' to obtain the frequency of each category

$$=COUNTIF(A2:A301,1), =COUNTIF(A2:A301,2), =COUNTIF(A2:A301,3), =COUNTIF(A2:A301,4),$$

$$=COUNTIF(A2:A301,5), =COUNTIF(A2:A301,6).$$

If you already know the number of occurrences of each value, type the magazine names in cells B2–B7 and the frequencies in cells C2–C7, with titles in cells B1 and C1 (as above), and proceed as follows:

- 4 Highlight the category and frequency data columns (**B1:C7**).
- 5 Click **INSERT**. In the **Charts** submenu, select the **Column chart** icon . Then select the first **2D column chart**.
- 6 Click inside the box containing the bar chart. **Chart Tools** will appear on the **Menu bar**. This will allow you to make changes to the chart. Click on the gridlines in the chart and delete. Click on the chart title and change it. If cell C1 is empty, then insert a chart title as follows: Click **DESIGN** under **Chart Tools** and click **Add Chart Element** in the **Chart Layout** submenu. Then click **Chart title** and then select **Above Chart**. To include axis titles, click **Add Chart Element** in the **Chart Layout** submenu, click **Axis titles**, choose **horizontal** or **vertical** and type the axis title. Click **Data Labels**, select **More Data Label Options** and tick only the **Value** check box
(Women's magazine readership, New Zealand, 2018).

To draw a pie chart, use the same instructions with some minor changes as follows.

COMMANDS

Pie chart

Proceed as you did to create a bar chart above, but in step 5, instead of the column chart, select the **Pie chart** icon . In step 6, delete the legends and in **More Data Label Options**, tick the **Category name** and **Percentage** check boxes.

Interpreting the results

From the frequencies presented in **Table 3.1** and also from the bar and pie charts, one can easily see that **NZ Woman's Day** is the most popular women's magazine in New Zealand.

EXAMPLE 3.2

L02

Top 15 beer-consuming countries

XM03-02 The following table presents the per capita beer consumption for the top 15 beer-consuming countries around the world. Use an appropriate graphical technique to depict these numbers.

TABLE 3.2 Per capita beer consumption (in litres), top 15 countries, 2017

	Country	Beer consumption (litres)
1	Australia	72
2	Austria	96
3	Belgium	69
4	Bulgaria	76
5	Croatia	81
6	Czech Republic	137
7	Estonia	71
8	Germany	96
9	Ireland	79
10	Latvia	77
11	Lithuania	92
12	Poland	98
13	Romania	76
14	Slovenia	77
15	United States	75

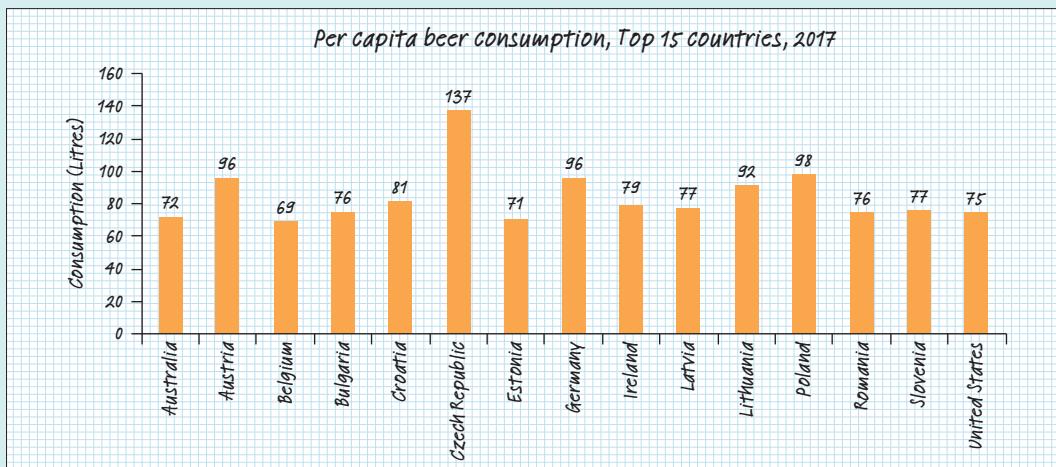
Source: Statista Alcoholic Drinks Report 2018 – Beer CC BY-ND 4.0 <https://creativecommons.org/licenses/by-nd/4.0/>

Solution

Identifying the technique

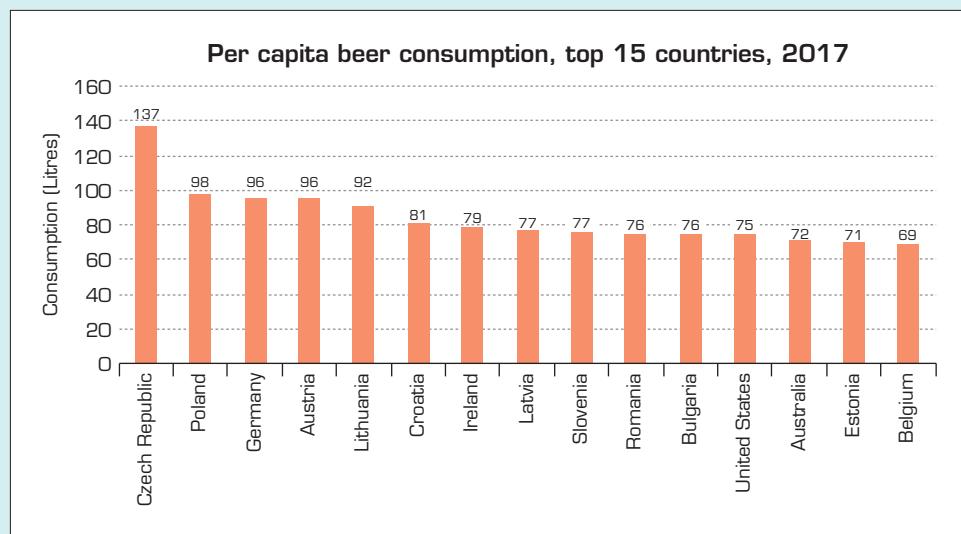
In this example, we are primarily interested in comparing litres of beer consumed per person across countries. There is no use in presenting proportions here. Thus the appropriate technique is the bar chart. **Figure 3.3** depicts the manually drawn bar chart

FIGURE 3.3 Per capita beer consumption: top 15 countries, 2017



An Excel chart can be drawn using the commands given for a bar chart in Example 3.1. To have a clear idea about the ranking of the top 15 countries, we first sort the data using the **Custom Sort...** under **Sort & Filter** in the **Editing** menu tab of the **HOME** screen in Excel. The chart for countries ordered in descending order of beer consumption is presented below.

Excel output for per capita beer consumption for the top 15 countries by ranking



Interpreting the results

Of the beer-consuming countries, the Czech Republic has the highest per capita beer consumption, followed by Poland, Germany and Austria. Australians are ranked thirteenth highest beer consumers in the world. New Zealanders consume well below the Australian level and are not ranked within the top 15.

3.1b Other applications of bar charts and pie charts

As discussed, bar charts and pie charts are frequently used to simply present numbers associated with categories. If the focus is to compare the size or frequency of various categories, a bar chart may be appropriate. Pie charts are effective whenever the objective is to display the components of a whole entity in a manner that indicates their relative sizes. A pie chart allows the reader to more quickly recognise the relative sizes of the categories, as in the breakdown of a budget. Similarly, managers might use pie charts to show the breakdown of a firm's revenues by department, and university students might use pie charts to show the amount of time devoted to daily activities (e.g. eating, 10%; sleeping, 30%; studying, 40%; other activities, 20%).

REAL-LIFE APPLICATIONS

Macroeconomics

Macroeconomics is a major branch of economics that deals with the behaviour of the economy as a whole. Macroeconomists develop mathematical models that predict variables such as gross domestic product, unemployment rates, and inflation. These are used by governments and corporations to help develop

strategies. For example, central banks attempt to control inflation by lowering or raising interest rates. To do this requires that economists determine the effect of a variety of variables, including the supply and demand for energy.

REAL-LIFE APPLICATIONS

Energy economics

One variable that has had a large influence on the economies of virtually every country is energy. The 1973 oil crisis in which the price of oil quadrupled over a short period of time is generally considered to be one of the largest financial shocks to the world's economies. In fact, economists often refer to two different economies: before the 1973 oil crisis and after. Unfortunately, the world will be facing more shocks because of energy for two primary reasons. The first is the depletion of non-renewable sources of

energy and the resulting price increases. The second is the possibility that burning fossil fuels and the creation of more carbon dioxide may be the cause of global warming. One economist predicted that the cost of global warming will be calculated in the trillions of dollars. Statistics can play an important role by determining whether Earth's temperature has been increasing and, if so, whether carbon dioxide is the cause (see Case 4.1).

Consider Examples 3.3 and 3.4. In this chapter, you will also encounter other exercises that involve the issue of energy.

EXAMPLE 3.3

LO1 LO2

Electricity generation in Victoria and NSW

XM03-03 Table 3.3 lists the total electricity generation in the two most populous states of Australia, New South Wales (NSW) and Victoria, from all sources in 2018. Use a graphical technique to display the differences between the sources of electricity for the two states.

TABLE 3.3 Electricity generation (in GWh) by source in NSW and Victoria, 2018

Source	NSW	Victoria
Non-renewable energy sources		
Coal	57 251.2	35 962.1
Gas	2 238.7	3 027.1
Oil	411.5	164.8
Renewable energy sources		
Biomass	1 166.1	666.8
Wind	3 122.0	4 580.0
Hydro	4 946.9	1 114.2
Solar	3 335.8	1 822.6
Total	72 472.2	47 337.6

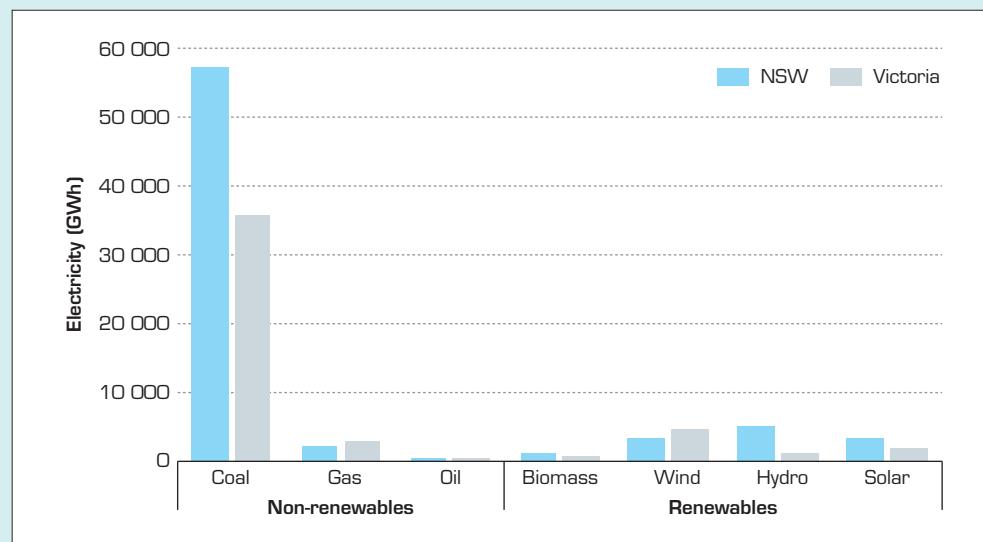
Source: © Department of Agriculture, Water and the Environment 2020.

CC BY 4.0 <https://creativecommons.org/licenses/by/4.0/>

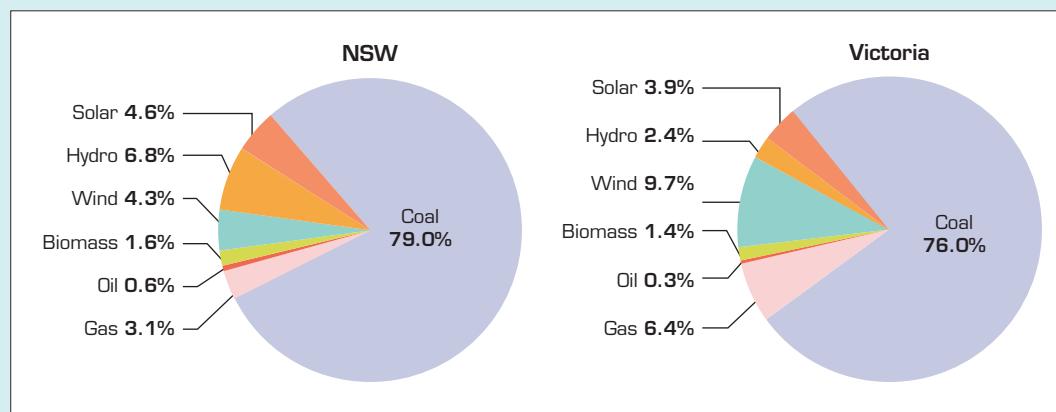
Solution

Identifying the technique

Overall, NSW generates more electricity than Victoria. For the non-renewables, NSW produces more electricity from coal and oil, while Victoria produces more from gas. For the renewables, NSW uses more biomass, hydro and solar, while Victoria uses more wind to produce electricity. The bar chart in **Figure 3.4a** was created using Excel using the commands for a bar chart given in Example 3.1.

FIGURE 3.4A Electricity generation by source, NSW and Victoria, 2018

If we want to make a comparison between the two states based on the shares of the electricity sources, we are interested in describing the proportion of total electricity generation from each source. Thus, the appropriate technique is the pie chart. The next step is to determine the proportions and sizes of the pie slices from which the pie charts are drawn. The bar charts in **Figure 3.4b** were created using Excel using the commands for a pie chart given in Example 3.1.

FIGURE 3.4B Pie charts of electricity generation by source, NSW and Victoria, 2018

Interpreting the results

Coal is the main source of electricity generation in both states. Oil and biomass are the least used electricity sources in both states. In NSW, about 79% of electricity is generated from coal, 6.8% from hydro, 4.6% from solar, 4.3% from wind, 3.1% from gas, 1.6% from biomass and 0.6% from oil, while in Victoria, 76% is generated from coal, 9.7% from wind, 6.4% from gas, 3.9% from solar, 2.4% from hydro, 1.4% from biomass and 0.3% from oil. Both states rely much more on non-renewable than renewable energy sources. Victoria's share of electricity generation from gas and wind is double that of NSW.

EXAMPLE 3.4

LO2

Australian natural gas exports by destination

XM03-04 Australia is one of the major natural gas exporting countries. The data for Australian natural gas exports recorded by country of destination in 2019 are presented in **Table 3.4**. Use an appropriate graphical technique to show the share of total Australian natural gas exports for each destination.

TABLE 3.4 Principal markets for Australian natural gas exports (million tons) by destination, 2019

Destination	Natural gas exports (million tons)	Share of total exports (%)
Japan	35.2	47
China	23.2	31
South Korea	9.0	12
Singapore	3.0	4
India	2.3	3
Rest of the world (ROW)	2.3	3
Total	70.0	100

Source: Department of Industry, Science, Energy and Resources, Commonwealth of Australia Resources and Energy Quarterly March 2020. The Commonwealth of Australia does not necessarily endorse the content of this publication. CC BY 4.0 <https://creativecommons.org/licenses/by/4.0/>

Solution**Identifying the technique**

If our interest is to compare the amount of natural gas exported to various destinations, a bar chart would be appropriate. However, if we are interested in describing the proportion (share) of total natural gas exports by destination, then the appropriate technique is the pie chart. **Figure 3.5** depicts the manually drawn bar chart and pie chart. (Charts drawn using Excel are exactly the same.)

FIGURE 3.5 Australian natural gas exports by destination, 2019

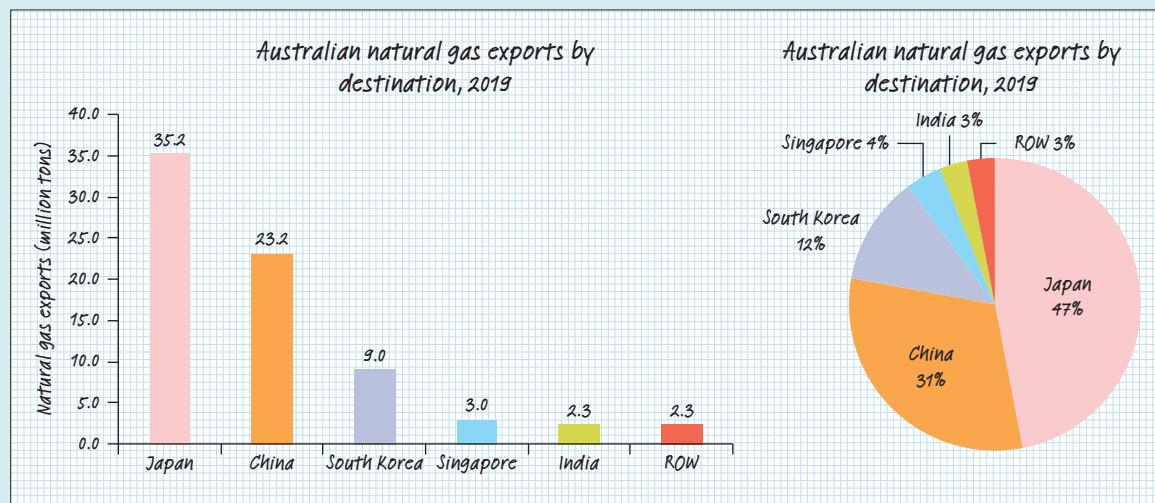
**Interpreting the results**

Figure 3.5 shows that Japan is Australia's major importer of natural gas, accounting for about 47% of the total Australian natural gas exports. The other major importers are China (31%) and South Korea (12%). In total, these three countries import about 90% of Australia's total natural gas exports.

3.1c Selecting the appropriate chart: Which chart is best?

As either a bar chart or a pie chart can be used to represent nominal data graphically, which representation should be used? The answer is that it all depends on what you want to emphasise. If the focus is to compare the size or frequency of various categories, a bar chart may be appropriate.

For example, consider the data in **Table 3.5**, which gives the number of new passenger vehicles sold in Australia for the 10 best-selling car models during 2017 and 2018.

TABLE 3.5 Top 10 best-selling new passenger vehicle sales in Australia, by manufacturer 2017 and 2018

Rank	Model	Volume of sales		Change (%)	Market share	
		2017	2018		2017	2018
1	Toyota	216 566	217 061	0.2	24.2	25.2
2	Mazda	116 349	111 280	-4.4	13.0	12.9
3	Hyundai	97 013	94 187	-2.9	10.8	10.9
4	Mitsubishi	80 654	84 944	5.3	9.0	9.9
5	Ford	78 161	69 081	-11.6	8.7	8.0
6	Holden	90 306	60 751	-32.7	10.1	7.0
7	Kia	54 737	58 815	7.5	6.1	6.8
8	Nissan	56 594	57 699	2.0	6.3	6.7
9	Volkswagen	58 004	56 620	-2.4	6.5	6.6
10	Honda	46 783	51 525	10.1	5.2	6.0
	Total	895 167	861 963		100.0	100.0

Source: Federal Chamber of Automotive Industries, 2019.

If we wish to compare the actual sales in 2017 and 2018 by manufacturer, we can use a bar chart as shown in **Figure 3.6**, using the volume of sales data in **Table 3.5**. As **Figure 3.6** shows, Toyota was the passenger-vehicle market leader in both 2017 and 2018. However, if we wish to emphasise the drop or increase in sales by manufacturers, we can calculate the changes in sales between 2017 and 2018 for each manufacturer, as in the fifth column of **Table 3.5**, to produce the bar chart shown in **Figure 3.7**. As **Figure 3.7** shows, sales have declined for Mazda, Hyundai, Ford, Holden and Volkswagen, while they have increased for Toyota, Mitsubishi, Kia, Nissan and Honda. If, instead, we want to focus on the overall sales profiles for the years 2017 and 2018 separately, we can group the sales by year to produce two separate bar charts, as shown in **Figure 3.8**.

FIGURE 3.6 Bar chart of new passenger vehicle sales, 10 best-selling manufacturers, 2017 and 2018

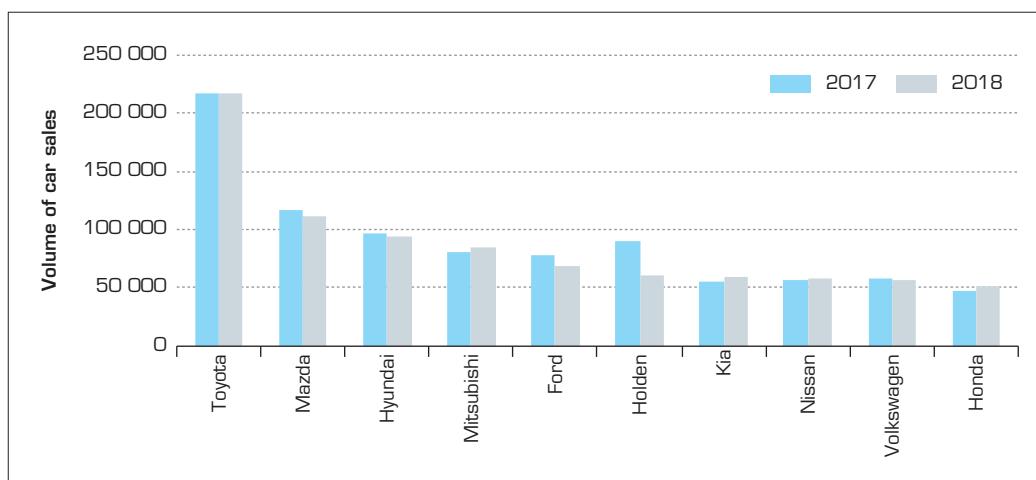
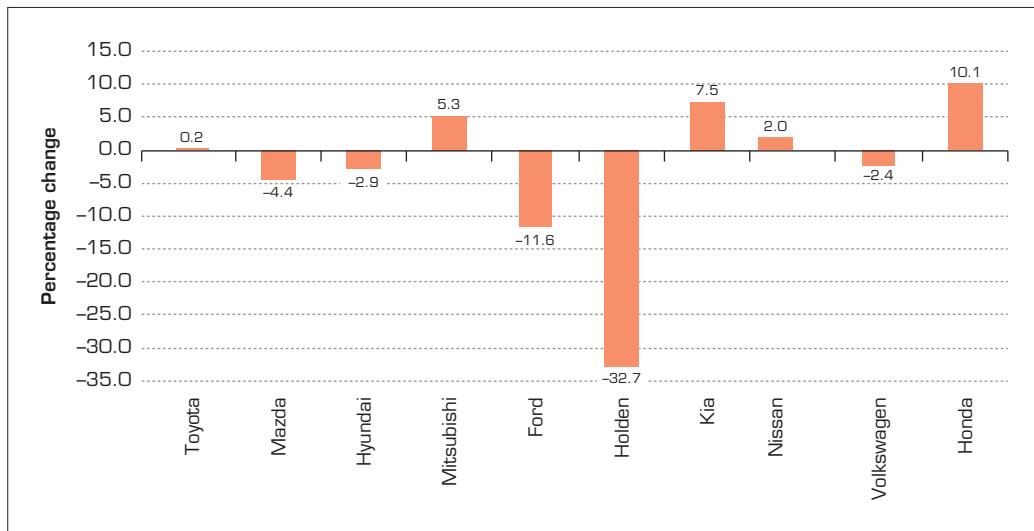
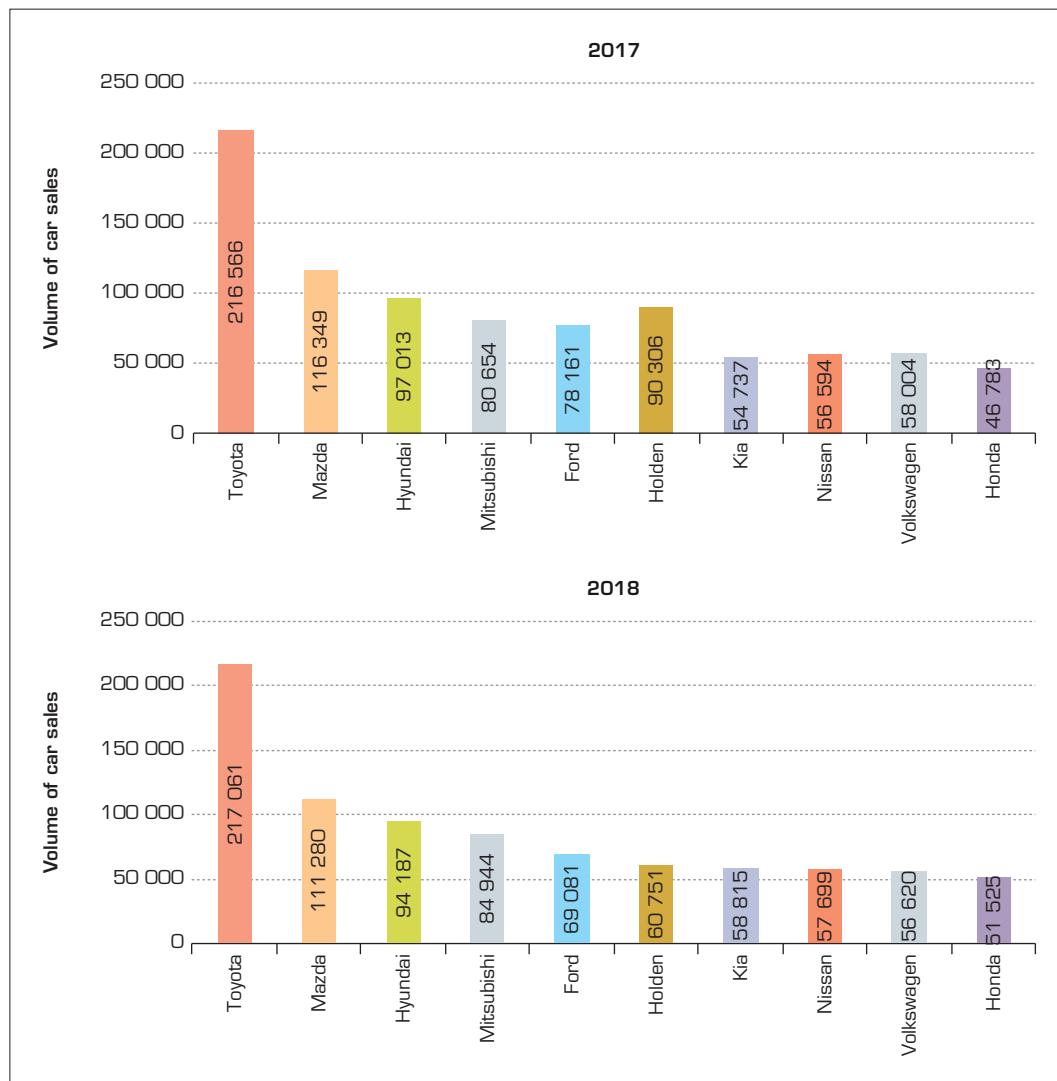


FIGURE 3.7 Bar chart emphasising change in sales by manufacturer between 2017 and 2018**FIGURE 3.8** Bar charts emphasising sales profile by year, 2017 and 2018

Alternatively, if we wish to highlight each manufacturer's changing market share for the top 10 selling cars, constructing two pie charts (one for each year) would be more appropriate, as shown in **Figure 3.9a**. A bar chart comparing the market share of each manufacturer in 2017 and 2018 would also serve our purpose (see **Figure 3.9b**). We can see that from 2017 to 2018, the market shares for Mazda, Hyundai, Volkswagen have remained more or less unchanged, while those for Holden and Ford have decreased. The market shares of Toyota, Mitsubishi, Kia, Nissan and Honda have increased.

FIGURE 3.9A Pie charts emphasising change in market share, 2017 vs 2018

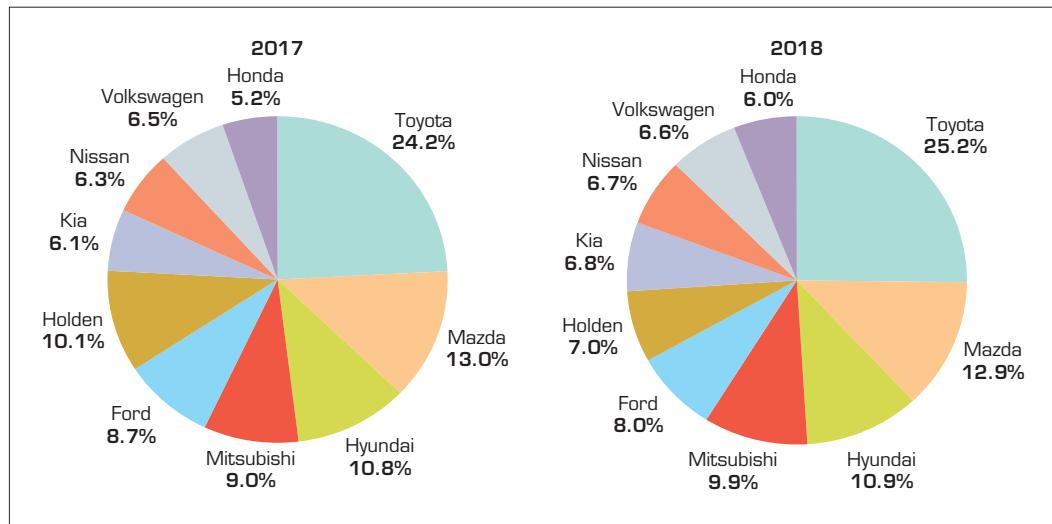
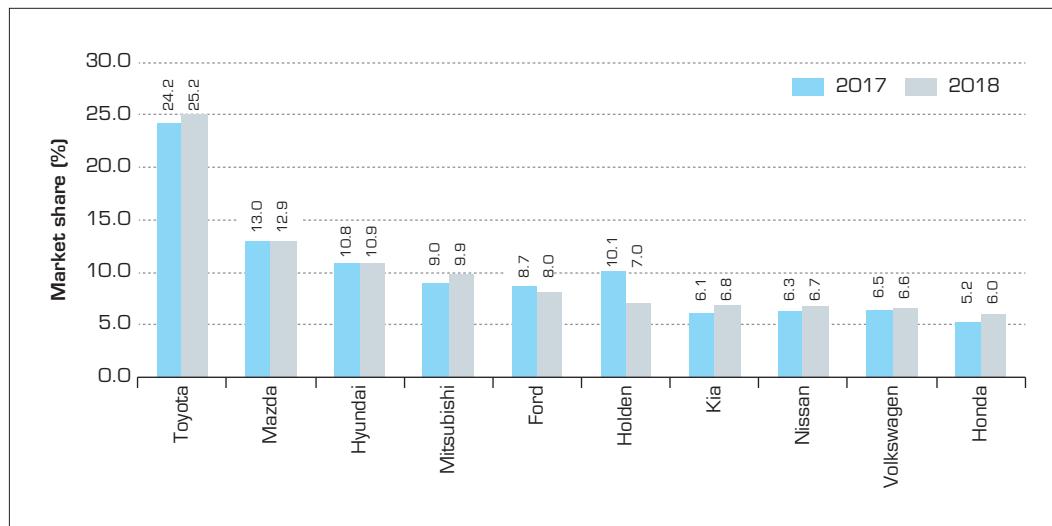


FIGURE 3.9B Bar chart emphasising change in market shares, 2017 vs 2018



EXAMPLE 3.5

LO3

Australian food and fibre exports

XMO3-05 Australian food and fibre exports have been on the increase for decades. Victoria leads the six Australian states in food and fibre exports. The following table lists the value of food and fibre exports (in millions of dollars), by state, in Australia for the years 2014 and 2018.¹ Use an appropriate graphical technique to compare the food exports by state for the two years.

TABLE 3.6 Australian food and fibre exports by state, 2014 and 2018

State	Exports (\$million)	
	2014	2018
Victoria	12 153	14 141
Queensland	7 514	9 120
NSW	7 894	8 621
WA	6 330	6 678
SA	5 055	6 304
Tasmania	774	1 051
Others*	3 982	6 216
Total Australia	43 702	52 131

*Others represent data for exports from the Australian territories, Australian Capital Territory and Northern Territory, and exports for which no state details are provided.

Source: Victorian Food and Fibre Export Performance Report, 2017–18. © State of Victoria (Department of Environment, Land, Water and Planning) CC BY 4.0
<https://creativecommons.org/licenses/by/4.0/>

Solution**Identifying the technique**

If we wish to compare the actual exports for 2014 and 2018, then a bar chart as shown in **Figure 3.10** is appropriate. However, if the emphasis is to highlight each state's share of overall Australian food exports in the years 2014 and 2018, then a pie chart as shown in **Figure 3.11a** or a bar chart as in **Figure 3.11b** is suitable. If we want to show the changes in exports from Australia by state between 2014 and 2018 in percentage form, then **Figure 3.12** would be appropriate.

TABLE 3.7 Australian food and fibre exports by state: Percentage changes and export share, 2014 and 2018

State	Exports (\$million)		Change (%)	Export share (%)	
	2014	2018		2014	2018
Victoria	12 153	14 141	16.4	27.8	27.1
NSW	7 514	9 120	21.4	17.2	17.5
Queensland	7 894	8 621	9.2	18.1	16.5
WA	6 330	6 678	5.5	14.5	12.8
SA	5 055	6 304	24.7	11.6	12.1
Tasmania	774	1 051	35.8	1.8	2.0
Others	3 982	6 216	56.1	9.1	11.9
Total Australia	43 702	52 131	19.3	100.0	100.0

¹ In this example, years 2014 and 2018 refer to financial years 2013–14 and 2017–18 respectively.



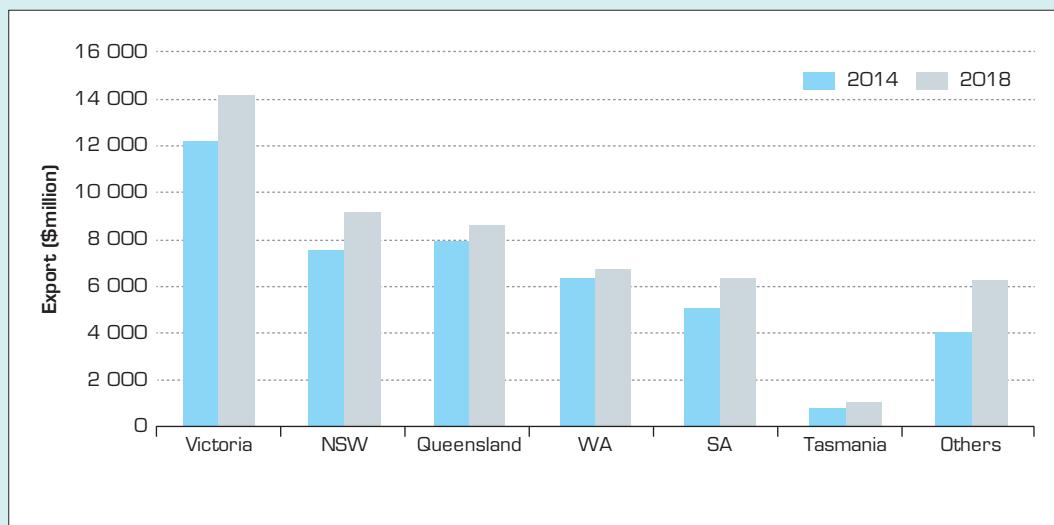
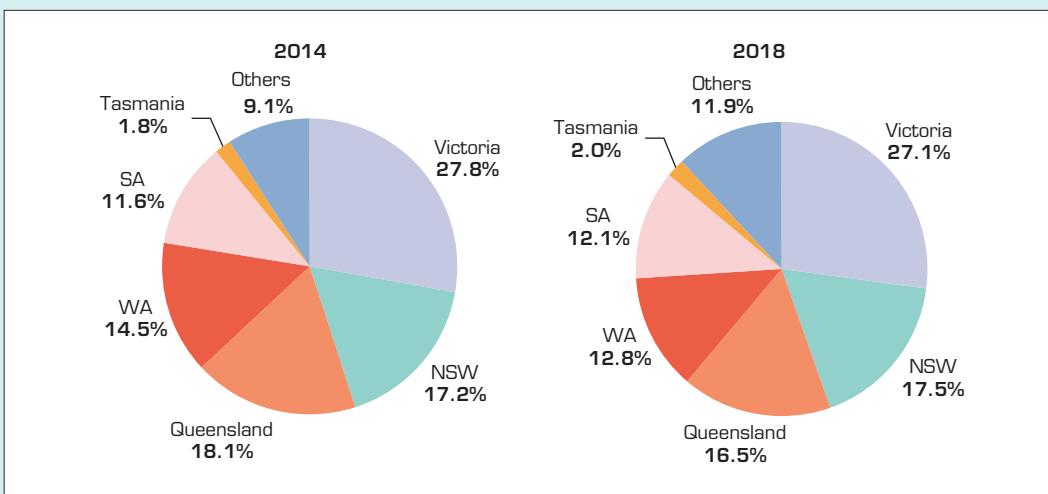
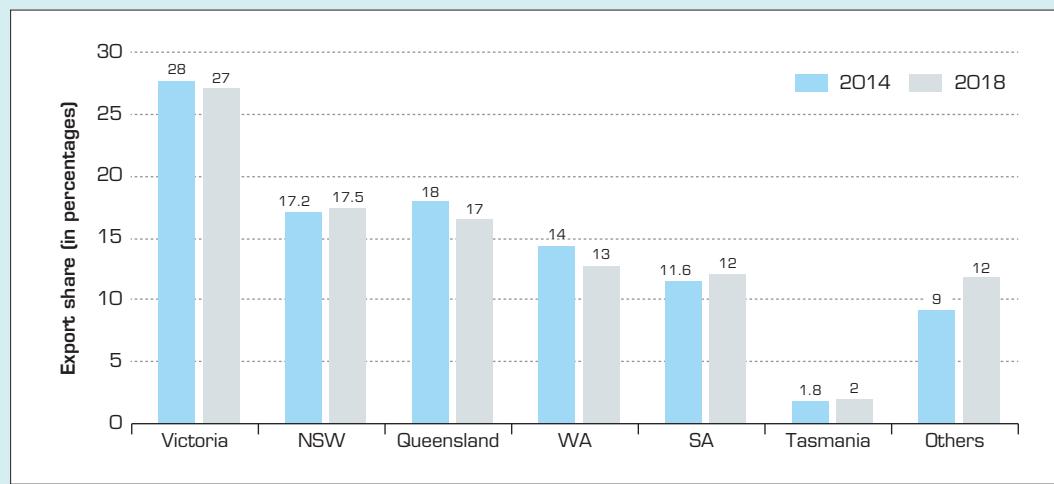
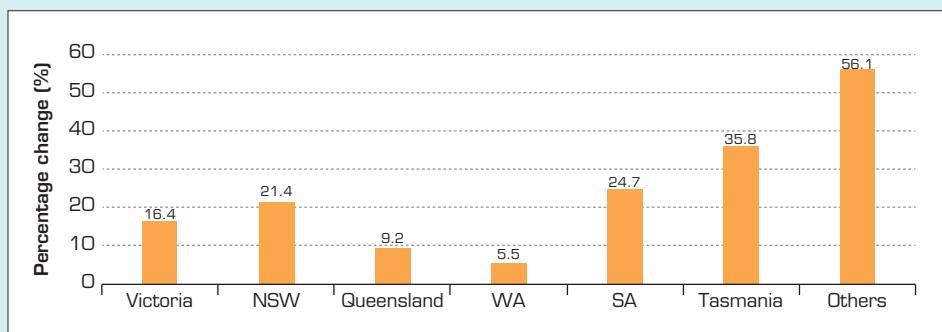
FIGURE 3.10 Bar chart of Australian food and fibre exports by state, 2014 and 2018**FIGURE 3.11A** Pie charts of Australian food and fibre export shares by state, 2014 vs 2018**FIGURE 3.11B** Bar chart of Australian food and fibre export shares by state, 2014 vs 2018

FIGURE 3.12 Bar chart of percentage change in Australian food and fibre exports by state between 2014 and 2018



Interpreting the results

As can be seen from the last row of **Table 3.7**, overall, Australian food and fibre exports increased by 19.3% between 2014 and 2018. The bar chart in **Figure 3.10** shows that Victoria was the major exporter of food and fibre items from Australia during 2014 and 2018, followed by New South Wales, Queensland and Western Australia. All states have had various levels of increase in their exports from 2014 to 2018.

As can be seen from the pie charts and the bar chart in **Figures 3.11a** and **3.11b** respectively, Victoria contributed more than one-fourth of the Australian food exports in 2014 and 2018 and its share decreased slightly from 27.8% to 27.1% during this period. New South Wales' share has increased slightly from 17.2% to 17.5%, and Queensland and Western Australia's shares have decreased from 18.1% to 16.5% and from 14.5% to 12.8%, respectively. South Australia and Tasmania's shares have increased slightly from 11.6% to 12.1% and from 1.8% to 2.0%, respectively, from 2014 to 2018.

Overall, as can be seen from the percentage changes in **Figure 3.12**, food and fibre exports for the individual states have increased in all states between 2014 and 2018. Of the six states, Tasmania had the highest growth (35.8%) followed by South Australia (24.7%), New South Wales (21.4%) and Victoria (16.4%) and Queensland (9.2%); Western Australia had the lowest growth (5.5%) between 2014 and 2018.

3.1d Component bar charts

Rather than using a separate bar for each category, we could use a component bar chart. A component bar chart represents all categories within a single bar. The bar is partitioned into components, with the height of each component proportional to the frequency of the category that it represents. When a comparison of two breakdowns is desired, component bar charts offer a good alternative to using two pie charts, as they usually enable the reader to detect the magnitude of changes in the category sizes more easily.

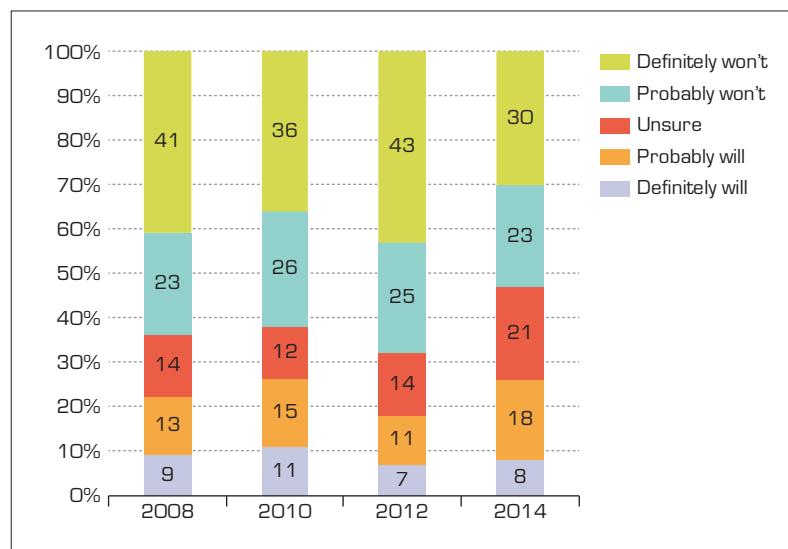
For example, to observe signs of greater optimism among investors between 2008 and 2014, those who were current direct investors at the time were asked about their likelihood of buying shares in the next 12 months. **Table 3.8** below presents the percentage breakdown of the investors and their level of likelihood of buying shares in the next 12 months in 2008, 2010, 2012 and 2014.

TABLE 3.8 Components of likelihood of buying shares in the next 12 months (in percentages) for 2008, 2010, 2012 and 2014

Buying shares	2008	2010	2012	2014
Definitely will	9	11	7	8
Probably will	13	15	11	18
Unsure	13	12	14	21
Probably won't	23	26	25	23
Definitely won't	41	36	43	30

The information in **Table 3.8** can be displayed using a component bar chart as in **Figure 3.13**. The chart shows that the likelihood of buying shares (indicated by ‘probably will’ and ‘definitely will’ responses) has increased from 22% ($13 + 9$) in 2008 to 26% ($18 + 8$) in 2014, whereas the percentage of investors who definitely or probably wouldn’t buy shares has decreased from 64% ($41 + 23$) in 2008 to 53% ($30 + 23$) in 2014. This indicates greater optimism among investors in 2014 than in 2008.

FIGURE 3.13 Likelihood of investors buying shares in the next 12 months, 2008, 2010, 2012 and 2014



3.1e Graphical techniques to describe ordinal data

There are no specific graphical techniques for ordinal data. Consequently, when we wish to describe a set of ordinal data, we will treat the data as if they were nominal and use the techniques described in this section. The only criterion is that the bars in bar charts should be arranged in ascending (or descending) ordinal values; in pie charts, the wedges are typically arranged clockwise in ascending or descending order.

EXAMPLE 3.6

Australian household income by quintile

XMO3-06 Income varies vastly among a country’s people. The table below presents the annual disposable income in 2018 of Australians who fall into the five quintiles.² Present the data in graphical form.

² Quintiles split the population into five equal parts. When you order the values of the variable in ascending order, the 20% of the population with lowest values is referred to as the lower quintile (or quintile 1). Similarly, the 20% of the population with highest values is referred to as the upper quintile (or quintile 5).



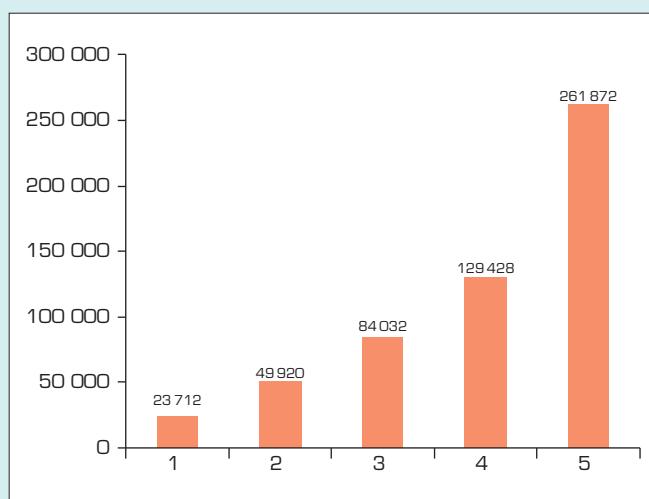
Quintile	Annual disposable income (\$)
1	23 712
2	49 920
3	84 032
4	129 428
5	261 872

Source: Australia's household income and wealth distribution,
<https://mccrindle.com.au/insights/blog/australias-household-income-wealth-distribution/>.

Solution

The disposable income of Australians is categorised into five categories, so these are nominal data. However, the data also have a ranking, as they are grouped based on the level of income (quintiles). Therefore, the data are ordinal (ranked). The ranked nature of the quintiles can be better displayed visually with the use of a bar chart. Therefore, the most suitable way of describing these data is by presenting a bar chart, as in **Figure 3.14**.

FIGURE 3.14 Annual household disposable income by quintiles, Australia, 2018



Interpreting the results

As can be seen from **Figure 3.14**, the annual disposable income of each quintile increases gradually from the first quintile to the fourth quintile. In quintile 5, there is a sharp increase in the annual disposable income, indicating that the income of the top 20% is much higher than the income of those in the next level (quintile 4).

We complete this section by reviewing when bar and pie charts are used to summarise and present data.

IN SUMMARY

Factors that identify when to use frequency and relative frequency tables, and bar and pie charts

- 1 *Objective:* to describe a set of data
- 2 *Data type:* nominal (categorical)

EXERCISES

Learning the techniques

- 3.1 XR03-01** The number of new dwelling units approved in the Australian states and territories during Jan–April 2019 are shown below.

Number of new dwelling units approved in Australia, Jan–April 2019

State/Territory	Approvals
New South Wales	17 599
Victoria	17 985
Queensland	9 393
South Australia	3 338
Western Australia	5 096
Tasmania	1 025
Northern Territory	182
Australian Capital Territory	1 844

Source: Australian Bureau of Statistics, 8731.0—*Building Approvals, Australia*, May 2019.

- a Draw a bar chart to show the number of new dwelling units approved in each state and territory of Australia.
- b Draw a pie chart to display the share of the dwelling units approved in Australia by state and territory.

- 3.2 XR03-02** A car manufacturer is developing a strategy for the distribution of cars to various states in Australia. To do this, the manufacturer is gathering information about the demand for new vehicles in each state. The following table describes the number of new car sales in Australia for each state and territory during December 2017. Use graphical techniques to present:
- a the sales by state or territory
 - b the sales share by state or territory.

Number of new vehicles sales by state/territory, December 2017

State/territory	Sales
NSW	33 014
Victoria	31 336
Queensland	19 543
WA	7 942
SA	6 611
Tasmania	2 166
ACT	1 498
NT	710
Total	102 820

Source: Australian Bureau of Statistics, 9314.0—*Sales of New Motor Vehicles, Australia*, ABS, Canberra, August 2018.

- 3.3 XR03-03** The student administration unit at a university wanted to determine the types of jobs the graduates of the university were performing. A sample of 500 graduates of the university was surveyed five years after graduation and asked to report their occupations. The responses and their frequencies are listed below. Summarise the data with an appropriate graphical method.

Job category	Frequency
1 Unemployed	24
2 Manager	162
3 Blue-collar worker	54
4 Clerical worker	180
5 Other	80

- 3.4 XR03-04** Where do consumers get information about cars? A sample of 993 recent car buyers was asked to identify the most useful source of information about the cars they purchased. The responses and their frequencies are listed below. Graphically depict the responses.

Source	Frequency
1 Consumer guide	516
2 Dealership	279
3 Word of mouth	120
4 Internet	78

- 3.5 XR03-05** The following table presents the population of Fiji by ethnic group in 1996 and 2007. Depict the data graphically, showing whether there was any change in the ethnic composition of the Fijian population between 1996 and 2007.

Distribution of the Fijian population by ethnic group

Ethnic group	Population	
	1996	2007
Chinese	4 939	4 704
Europeans	3 103	2 953
Fijians	393 575	475 739
Indians	338 818	313 798
Total population	775 077	837 271

Source: Census 2007, Fiji Islands Bureau of Statistics, 2008

- 3.6 XR03-06** According to the Fiji Bureau of Statistics, 73 169 overseas visitors arrived into the country in May 2019. The following table presents the

number of arrivals from top seven countries during this period. Depict the data graphically showing the number of arrivals and the country-wise share of the Fiji inbound tourist arrivals.

Number of inbound tourist arrivals by country, Fiji, May 2019

Country	Arrivals
Australia	30 035
New Zealand	16 372
USA	8 585
Pacific Islands	4 179
China	4 028
United Kingdom	1 374
Canada	984
Japan	944
South Korea	813
Others	5 855
Total	73 169

- 3.7 XR03-07** It is claimed that with advances in medicine, Australians are living longer and, due to work and other commitments, are having fewer children. This can be easily verified by looking at the changing age structure of the population. The following table presents the percentage breakdown of the Australian population by age group in 1991, 2011 and 2018. Display the information graphically to verify the claim.

Australian population by age group (%), 1991, 2011 and 2018

Age group	1991	2011	2018
Under 15	21.9	18.8	18.8
15–64	66.8	67.4	65.5
65–84	10.4	11.9	13.6
85 and over	0.9	1.8	2.0

Source: Australian Bureau of Statistics, *Population change, Australian Demographic Statistics*, Dec 2014 and Sept 2018, cat. no. 3101.0, ABS, Canberra.

Applying the techniques

- 3.8 XR03-08 Self-correcting exercise.** The following table shows the value of Australian merchandise trade with various regions of the world for 2016 and 2018.

- a Use a graphical technique to depict the information given in the table.
- b Explain why you selected the particular technique.

Australian merchandise trade by region (in \$m), 2016 and 2018

Region	Exports		Imports	
	2016	2018	2016	2018
Africa	3 967	4 954	3 731	5 734
Americas	29 402	30 276	59 730	60 592
Asia	231 256	307 655	184 665	218 821
Europe	31 995	34 982	77 623	83 663
Oceania	17 230	18 799	18 383	19 622
Others	5 871	6 576	13 364	6 969
Total	319 721	403 242	357 496	395 401

Source: Department of Foreign Affairs and Trade. CC BY 3.0 AU <https://creativecommons.org/licenses/by/3.0/au/legalcode>

- 3.9 XR03-09** Since 2000, consumer prices in Australia have increased more than the increase in consumer prices in a number of other OECD (Organisation for Economic Co-operation and Development) countries. The following table shows the consumer price index of all goods and services in Australia and a number of similar OECD countries for the years 2015–18 with base year 2015 = 100. Display the data using an appropriate graphical technique.

Consumer price index in major OECD countries (base year 2015 = 100, second quarter)

Country	2015	2016	2017	2018
Australia	100.0	101.3	103.3	105.2
Canada	100.0	101.4	103.0	105.4
France	100.0	100.2	101.2	103.1
Germany	100.0	100.5	102.0	103.8
Italy	100.0	99.9	101.1	102.3
Japan	100.0	99.9	100.4	101.3
New Zealand	100.0	100.6	102.5	104.1
Sweden	100.0	101.0	102.8	104.8
United Kingdom	100.0	101.0	103.6	106.0
United States	100.0	101.3	103.4	105.9

Source: OECD Stats Extract, https://stats.oecd.org/Index.aspx?DataSetCode=PRICES_CPI, extracted 15 May 2019

- 3.10 XR03-10** In recent years, Australians have been waiting longer before getting married for the first time and the number of long-term de facto couples is also on the increase. These trends may have some effect on Australian society and government policies. The table below presents the age-specific marriage rate (in percentages) of Australian males and females by age group for the years 1997, 2007 and 2017.

- a Use a graphical technique to compare the marriage rates in 1997, 2007 and 2017, broken down by age, for Australian males.

- b** Use a graphical technique to compare the marriage rates in 1997, 2007 and 2017, broken down by age, for Australian females.
- c** Compare your observations in parts (a) and (b).
- d** Explain why you selected the particular technique.

Age-specific marriage rate (in percentages)

Age group (years)	Male			Female		
	1997	2007	2017	1997	2007	2017
16–19	1.0	0.9	0.6	5.5	3.7	2.0
20–24	27.5	17.8	11.3	45.9	31.6	19.3
25–29	48.8	46.0	35.5	46.5	51.6	41.8
30–34	28.7	36.5	31.6	22.2	31.1	28.1
35–39	15.4	19.7	17.1	11.4	15.4	13.6
40–44	9.5	11.3	10.1	7.5	8.5	8.1
45–49	7.3	8.1	7.6	5.8	6.5	6.1
50 and over	3.5	3.7	3.3	1.9	2.1	1.9

Source: Australian Bureau of Statistics, *Marriages and Divorces, Australia, 2017*, cat. no. 3310.0, ABS, Canberra.

- 3.11 XR03-11** The following table presents the rate of unemployment in New Zealand recorded in the June quarter for 2010, 2012, 2014, 2016 and categorised by ethnic group. Use an appropriate graphical technique to depict the data.

Rate of unemployment in New Zealand by ethnic group: 2010, 2012, 2014 and 2016

Ethnicity	2010	2012	2014	2016
European	4.9	5.2	4.4	3.9
Māori	14.3	12.8	12.1	11.4
Pacific Islander	15.5	14.9	11.8	9.7
Asian	10.4	8.2	6.7	6.0
MELAA*	9.6	11.5	12.7	8.8
Other	3.8	7.5	4.3	5.2

*Middle East/Latin American/African

Source: Stats NZ and licensed by Stats NZ for reuse under the Creative Commons Attribution 4.0 International licence. CC BY 4.0 <https://creativecommons.org/licenses/by/4.0/>

- 3.12 XR03-12** Australian exports and imports by 10 major trading partners (in A\$millions) for 2018 are shown in the following table.

Country	Exports (A\$m)	Imports (A\$m)
China	123 274	71 346
US	21 424	48 752
Japan	51 328	26 267
Singapore	13 164	14 610
Germany	4 170	18 185
Thailand	6 610	18 078
India	21 145	7 971
Korea	23 628	28 674
UK	11 757	16 036
NZ	14 370	13 905
Others	112 330	395 400

Source: Department of Foreign Affairs and Trade. CC BY 3.0 AU <https://creativecommons.org/licenses/by/3.0/au/legalcode>

- a** Show how this information about the top 10 countries with respect to exports and imports can be conveyed using a graphical technique.
- b** Explain why you selected the particular technique in part (a).

- 3.13 XR03-13** The following table displays the electricity and heat production (in gigawatt hours, GWh) patterns of Australia and New Zealand. Use a suitable graph to make a comparison between the electricity and heat production sources in Australia and New Zealand.

Energy source	Australia	New Zealand
Non-renewable energy sources		
Coal	236 640	2 636
Oil	25 503	2 720
Natural gas	41 172	3 547
Nuclear	0	0
Renewable energy sources		
Hydroelectric	1 052	2 083
Biofuel	5 783	1 095
Other (solar etc.)	548	3 134
Total	310 698	15 215

- 3.14 XR03-14** The planet is being threatened by global warming, possibly caused by burning fossil fuels (petroleum, natural gas and coal) that produce carbon dioxide (CO_2). The following table lists the top 15 producers of CO_2 in 2009 and the annual amounts (millions of metric tonnes) of CO_2 from fossil fuels for 2009, 2013 and 2016. Graphically depict these figures. Explain what you have learned and comment on the progress these countries have made in reducing CO_2 emissions.

Country	2009	2013	2016
United States	5 957	5 300	5 012
China	5 323	10 330	10 433
Russia	1 696	1 800	1 662
Japan	1 230	1 360	1 240
India	1 166	2 070	2 534
Germany	844	840	776
Canada	631	550	676
United Kingdom	577	480	368
South Korea	500	630	604
Italy	467	390	358
Iran	451	410	643
South Africa	424	330	391
France	415	370	332
Saudi Arabia	412	490	517
Australia	407	390	415

Source: EDGAR: Trends in global CO₂ emissions: 2014 report, and JRC Science for Policy Report, 2017.

Computer applications

- 3.15 XR03-15** A breakdown of the sales for Woolworths Limited for the years 2015–18 is given below. Use pie charts and a bar chart to describe the sales revenue for these four years.

Business group	Sales revenue (\$m)			
	2015	2016	2017	2018
Australian Food	34 881	34 798	36 371	37 379
Endeavour Drinks	7 251	7 598	7 913	8 271
NZ Supermarkets	5 467	5 592	5 887	5 898
Big W	3 929	3 820	3 598	3 566
Petrol	5 632	4 612	4 682	4 784
Hotels	1 475	1 512	1 553	1 612
Group sales	58 635	57 932	60 004	61 510

Source: © Woolworths Limited.

- 3.16 XR03-16** Within Australia, most state tourist bureaus promote their state heavily in order to attract international tourists, thereby increasing the state's revenue and creating more jobs. The following is the breakdown of international visitors to each state or territory during 2016 and 2019 (based on the departures of overseas visitors by state in which the most time was spent). Use a graphical technique to show the number of international tourists visiting each state or territory for the two years. Also, use an appropriate graphical technique that emphasises the

percentage breakdown of international visitors to the Australian states and territories.

International visitors to Australia: 2016 and 2019

State/territory	Number of tourists ('000)	
	2016	2019
NSW	3119.2	3482.8
Victoria	2040.1	2509.8
Queensland	1826.2	2031.7
SA	217.0	276.2
WA	844.2	884.2
Tasmania	60.6	94.9
NT	83.5	77.1
ACT	78.4	109.1

Source: Australian Bureau of Statistics, 3401.0 – Overseas Arrivals and Departures, Table 11, February 2020, ABS, Canberra.

- 3.17 XR03-17** Food products are recalled by manufacturers for various reasons, including possible contamination, foreign objects and incorrect labelling. The following table presents the number of food recalls for the years 1990 to 2019, as published by Food Standards Australia New Zealand. Use a bar chart to present these statistics.

Food recalls, 1990–2019

Year	Number of recalls
1990	18
1991	15
1992	17
1993	42
.	.
2015	81
2016	72
2017	69
2018	100
2019	87

Source: © Food Standards Australia New Zealand, 2019, CC BY 3.0 <https://creativecommons.org/licenses/by/3.0/au/>

- 3.18 XR03-18** In a taste test, 250 people were asked which of five wines they preferred. The wines were labelled 1, 2, 3, 4 and 5. The 250 responses are stored in the file in coded form.

- a Draw a bar chart to show the preferences for each wine.
- b Draw a pie chart from these data.

- 3.19 XR03-19** The number of Australian female-owned small businesses has been on the increase over the

last few years. But there are large variations in the types of businesses owned by men and women. Suppose that a survey of female-owned and male-owned small businesses was conducted and the type of business each operated was recorded in the following format.

Business	Code
Services	1
Retail/wholesale/trade	2
Finance/insurance/real estate	3
Transportation/communication	4
Construction	5
Manufacturing	6
Agriculture and primary industries	7

The responses of women and men are stored in Columns 1 and 2 respectively of the data file. Use an appropriate graphical technique to summarise and present these data.

- 3.20 XR03-20** When will the world run out of oil? One way to estimate is to determine the oil reserves of the countries around the world. The table below displays the known oil reserves in 2018 of 36 countries in billions of barrels. Graphically describe the figures.

Rank	Country	Crude oil reserves (billion barrels)
1	Venezuela	302.300
2	Saudi Arabia	266.200
3	Canada	170.500
4	Iran	157.200
5	Iraq	148.800
.	.	.
31	Uganda	2.500
32	Argentina	2.162
33	United Kingdom	2.069
34	Gabon	2.000
35	Australia	1.821
36	Colombia	1.665

Source: © Central Intelligence Agency

- 3.21** Refer to Exercise 3.20. The total reserves in the world in 2018 were 1 620 202 billion barrels. Use a graphical technique that emphasises the percentage breakdown of the top 12, the middle 12 and the bottom 12 of the selected 36 countries, plus others.

- 3.22 XR03-22** The table below displays the level of oil production (barrels per day) of 33 selected countries in 2018. Graphically describe the figures.

Rank	Country	Oil production (million barrels/day)
1	Russia	10.580
2	Saudi Arabia	10.130
3	United States	9.352
4	Iran	4.469
5	Iraq	4.454
.	.	.
29	Argentina	0.479
30	Vietnam	0.271
31	Australia	0.263
32	Turkey	0.245
33	Congo	0.244

Source: © Central Intelligence Agency.

- 3.23** Refer to Exercise 3.22. The total oil production in the world in 2018 was 81.0 million barrels per day. The total oil production of the selected 33 countries was 78.38 million barrels per day. Use a graphical technique that emphasises the percentage breakdown of the top 11, middle 11 and the bottom 11 of the selected 33 countries, plus others.

- 3.24 XR03-24** The following table lists the average oil consumption per day (in barrels per day) for 30 selected countries in 2017. Use a graphical technique to present these figures.

Rank	Country	Oil consumption (barrels/day)
1	United States	19 960 000
2	China	12 470 000
3	India	4 521 000
.	.	.
19	Italy	1 236 000
20	Australia	1 175 000
21	Turkey	989 900
.	.	.
27	Argentina	806 000
28	Malaysia	704 000
29	Venezuela	659 000
30	Poland	649 600

Source: © Central Intelligence Agency

3.25 Refer to Exercise 3.24. The total oil consumption in the world in 2017 was 97 161 865 barrels per day. Use a graphical technique that emphasises the percentage breakdown of the top 10, middle 10 and the bottom 10 of the selected 30 countries, plus others.

3.26 XR03-26 Who applies to MBA programs? To help determine the background of the applicants, a sample of 230 applicants to a university's business school was asked to state their undergraduate degree. The degrees were recorded using these codes.

- 1** BA
- 2** BBus
- 3** BEng
- 4** BSc
- 5** Other

- a** Determine the frequency distribution.
- b** Draw a bar chart.
- c** Draw a pie chart.
- d** What do the charts tell you about the sample of MBA applicants?

3.27 XR03-27 Most universities have several different kinds of residences on campus. To help in long-term planning, one university surveyed a sample of graduate students and asked them to report their

marital status or relationship. The possible responses are listed:

- 1** Single
- 2** Married
- 3** Divorced
- 4** Other

The responses for a sample of 250 students were recorded and saved. Draw a graph that summarises the information that you deem necessary.

3.28 XR03-28 A number of business and economics courses require the use of computers. As a result many students buy their own computer. A survey asks 98 students to identify the computer brand they have purchased. The responses are recorded with the following codes.

- 1** Apple
- 2** Dell
- 3** Toshiba
- 4** HP
- 5** Acer
- 6** Other

- a** Use a graphical technique that depicts the frequencies.
- b** Graphically depict the proportions.
- c** What do the charts tell you about the brands of computers used by the students?

3.2 Describing the relationship between two nominal variables

In Sections 3.1–3.3 we presented graphical techniques used to summarise single sets of nominal data; techniques applied to single data sets are called univariate. There are many situations where we wish to depict the relationship between two variables; in such cases **bivariate methods** are required. There are two types of such methods commonly used, namely, tabular methods and graphical techniques.

bivariate methods

Techniques involving or dependent on two variables.

3.2a Tabulating the relationship between two nominal variables

To describe the relationship between two nominal variables, we must remember that we are only permitted to determine the frequency of the values. A variation of the bar chart introduced in Section 3.1 is used to describe the relationship between two nominal (categorical) variables in graphical form. As a first step, we need to produce a **cross-classification table** (also known as contingency table or cross-tabulation table), which lists the frequency of each combination of the values of the two variables. We will illustrate the use of graphs to describe the relationship between two nominal variables using data from the newspaper readership case in Example 3.7.

cross-classification table

A first step in graphing the relationship between two nominal variables.

EXAMPLE 3.7

LO5

Newspaper readership survey

XM03-07 In a major city there are four competing newspapers: N1, N2, N3 and N4. To help design advertising campaigns, the advertising managers of the newspapers need to know which segments of the newspaper market are reading their papers. A survey was conducted to analyse the relationship between newspaper read and occupation. A sample of newspaper readers was asked to name the newspaper they read and to indicate if they were a blue-collar worker (1), a white-collar worker (2), or a professional (3). Some of the data are listed below. Determine whether the two nominal variables are related.

Reader	Occupation	Newspaper
1	2	N2
2	1	N4
3	2	N1
...
...
352	3	N2
353	1	N3
354	2	N3

Solution

There are two ways to determine whether there is a relationship between the occupation of a person and the newspaper they read. One is the tabular method using the joint frequencies, and the other is the graphical method.

Tabular method

By counting the number of times each of the 12 combinations [(1,N1), (1,N2), (1,N3), (1,N4), (2,N1), (2,N2), (2,N3), (2,N4), (3,N1), (3,N2), (3,N3), (3,N4)] occurs, we produced the frequencies in **Table 3.9**.

TABLE 3.9 Cross-classification table of frequencies for Example 3.7

Occupation	Newspaper				Total
	N1	N2	N3	N4	
Blue collar (1)	27	18	38	37	120
White collar (2)	29	43	21	15	108
Professional (3)	33	51	22	20	126
Total	89	112	81	72	354

If occupation and newspaper are related, there will be differences in the newspapers read among the occupations. An easy way to see this is to convert the frequencies in each row (or column) to relative frequencies in each row (or column). That is, compute the row (or column) totals and divide each frequency by its row (or column) total, as shown in **Table 3.10**. (Totals may not equal 1 because of rounding.)

TABLE 3.10 Row relative frequencies for Example 3.7

Occupation	Newspaper				Total
	N1	N2	N3	N4	
Blue collar (1)	0.23	0.15	0.32	0.31	1.00
White collar (2)	0.27	0.40	0.19	0.14	1.00
Professional (3)	0.26	0.40	0.17	0.16	1.00
Total	0.25	0.32	0.23	0.20	1.00





Using the computer

There are several methods by which Excel can produce the cross-classification table. We will use and describe the PivotTable in two ways: first, to create the contingency table featuring the counts and, second, to produce a table showing the row relative frequencies.

	A	B	C	D	E	F
3	Count of Reader	Newspaper				
4	Occupation		1	2	3	4
5		1	27	18	38	37
6		2	29	43	21	15
7		3	33	51	22	20
8	Grand Total		89	112	81	72
						354

	A	B	C	D	E	F
3	Count of Reader	Newspaper				
4	Occupation		1	2	3	4
5		1	0.23	0.15	0.32	0.31
6		2	0.27	0.40	0.19	0.14
7		3	0.26	0.40	0.17	0.16
8	Grand Total		0.25	0.32	0.23	0.20
						1.00

The data must be stored in (at least) three columns, as we've done in file **XM03-07**. Put the cursor somewhere in the data range.

COMMANDS

- 1 Open the data file (**XM03-07**). Highlight the data including the column titles (**A1:C355**).
- 2 Click **INSERT** and **PivotTable** under the **Charts** submenu. Then select **Pivot Chart & Pivot Table**.
- 3 Make sure that the **Table/Range** is correct and click **OK**.
- 4 Drag the **Occupation** button from the menu that appears on the right of the screen to the **Drop Row Fields Here** section of the box. Drag the **Newspaper** button to the **Drop Column Fields Here** section. Drag the **Reader** button to the **Drop Value Fields Here** section. Right-click any number in the table, click **Summarize Values By**, and check **Count**. This will produce a contingency table for the counts (frequencies).
- 5 To convert to row percentages, right-click any number in the table, click **Summarize Values By, More options...,** and **Show values as**. Select **% of rows** from the drop-down menu and then click **OK**. Format the data into decimals by highlighting all cells with percentages and right-clicking. Select **Number Format...,** select **Number** under **Category** and select the number of **Decimal places**. Click **OK**. This will produce the contingency table for the row relative frequencies.

Interpreting the results

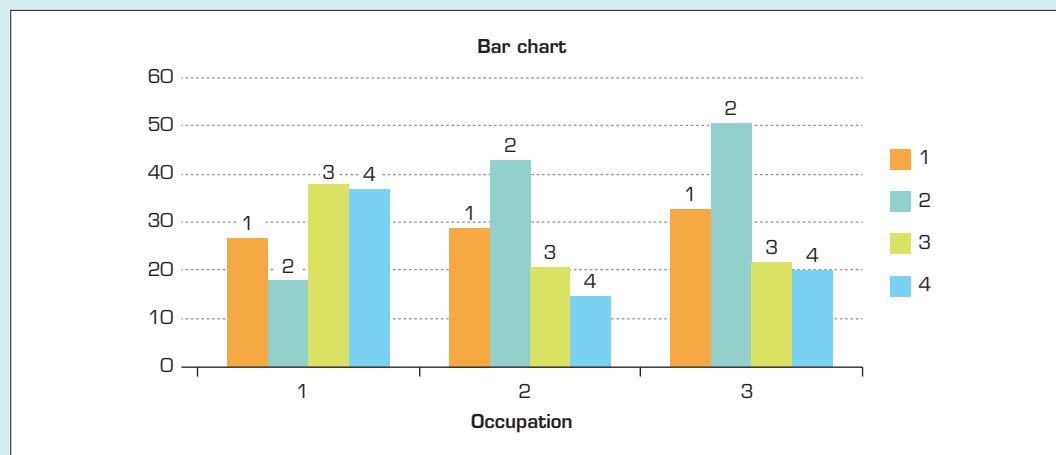
Notice that the relative frequencies in the second and third rows are similar and that there are large differences between row 1 and rows 2 and 3. This tells us that blue-collar workers tend to read different newspapers from both white-collar workers and professionals and that white-collar workers and professionals have quite similar newspaper choices.

Graphical method

To determine whether there is a relationship between the two variables, we can also choose to draw three bar charts – one for each occupation depicting the four newspapers. We will use Excel for this purpose. The manually drawn charts are identical.

There are several ways to graphically display the relationship between two nominal variables. We have chosen two-dimensional bar charts for each of the three occupations. The charts can be created from the output of the PivotTable (either counts as we have done or row proportions) as shown in **Figure 3.15**.



FIGURE 3.15 Excel bar chart showing readership of the four newspapers by occupation

COMMANDS

After clicking on a cell in the cross-classification table, click **INSERT** and in the **Charts** submenu, select the **Column chart** icon . Then select the first **2-D column** chart. You can do the same from any completed cross-classification table.

Interpreting the results

If the two variables are unrelated, the patterns exhibited in the bar charts should be approximately the same. If some relationship exists, then some bar charts will differ from others.

The graphs tell us the same story as the table. The shapes of the bar charts for occupation types 2 and 3 (white collar and professional) are very similar. Both differ considerably from the bar chart for occupation type 1 (blue collar).

3.2b Comparing two or more sets of nominal data

We can interpret the results of the cross-classification table of the bar charts in a different way. In Example 3.7, we can consider the three occupations as defining three different populations. If differences exist between the columns of the frequency distributions (or between the bar charts), then we can conclude that differences exist among the three populations. Alternatively, we can consider the readership of the four newspapers as four different populations. If differences exist among the frequencies or the bar charts, then we conclude that there are differences between the four populations.

We are now in a position to address the question posed in the introduction to this chapter.

SPOTLIGHT ON STATISTICS

Break bail, go to jail?: Solution

Using the technique introduced above, we produced the following bar chart.

Interpreting the results

As you can see, there are substantial differences between the bars in terms of responses. We can conclude that responses and party affiliation are related.

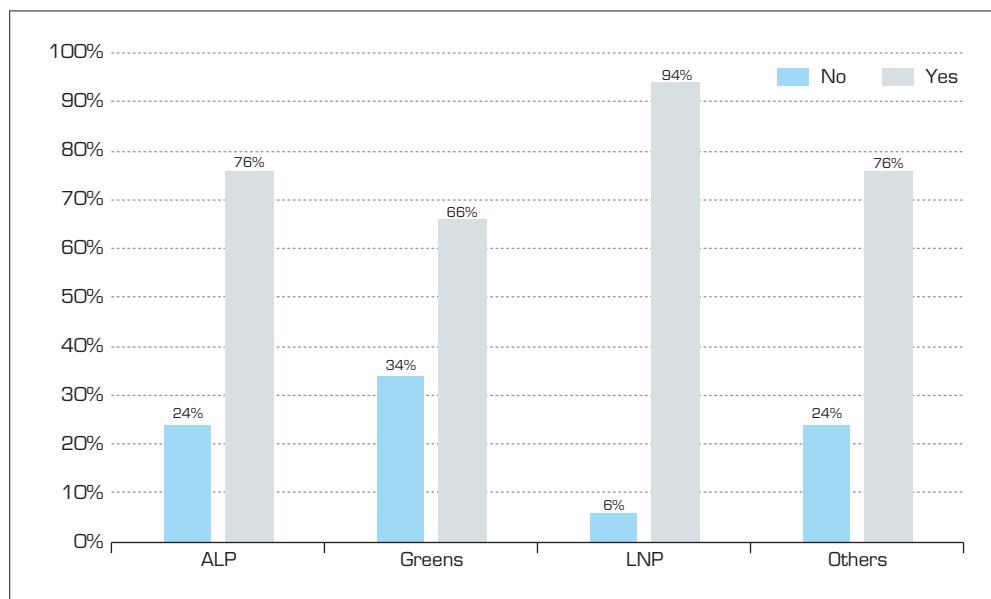


Source: Shutterstock.com/Sentavio



However, we can also conclude that differences in responses exist based on the party affiliation of the Victorian voters; specifically, more LNP voters tend to identify themselves with the 'yes' response compared to the green voters who are more inclined with a 'no' response. The share of 'yes' and 'no' responses of ALP and Other voters were not different.

FIGURE 3.16 Bar chart showing share of response by party affiliation



3.2c Data formats

There are several ways to store the data to be used in this section to produce a table or a bar or pie chart.

- 1 The data are in two columns. The first column represents the categories of the first nominal variable, and the second column stores the categories for the second variable. Each row represents one observation of the two variables. The number of observations in each column must be the same. Excel can produce a cross-classification table from these data. (To use Excel's PivotTable, there also must be a third variable representing the observation number.) This is the way the data for Example 3.7 were stored.
- 2 The data are stored in two or more columns, with each column representing the same variable in a different sample or population. For example, the variable may be the type of undergraduate degree of applicants to an MBA program, and there may be five universities we wish to compare. To produce a cross-classification table, we would have to count the number of observations of each category (undergraduate degree) in each column.
- 3 The table representing counts in a cross-classification table may have already been created.

IN SUMMARY

Factors that identify when to use a cross-classification table

- 1 *Objective:* to describe the relationship between two variables and compare two or more sets of data
- 2 *Data type:* nominal (categorical)

EXERCISES

Learning the techniques

- 3.29 XR03-29** The following table summarises the data from a survey on the ownership of iPads for families with different levels of income ($C_1 < C_2 < C_3$). Determine whether the two nominal variables are related.

Ownership	Level of income		
	C_1	C_2	C_3
No	40	32	48
Yes	30	48	52

Applying the techniques

- 3.30 XR03-30 Self-correcting exercise.** The trustee of a company's superannuation scheme has solicited the opinions of a sample of the company's employees regarding a proposed revision of the scheme. A breakdown of the responses is shown in the following table. Use an appropriate graphical presentation to determine whether the responses differ among the three groups of employees.

Responses	Blue-collar workers	White-collar workers	Managers
For	67	32	11
Against	63	18	9

Computer applications

- 3.31 XR03-31** The associate dean of a business school was looking for ways to improve the quality of applicants to its MBA program. In particular she wanted to know whether the undergraduate degrees of applicants to her school and the three nearby universities with MBA programs differed. She sampled 100 applicants of her program and an equal number from each of the other universities.

She recorded their undergraduate degree (1 = BA, 2 = BEng, 3 = BBA, 4 = other) as well as the university (codes 1, 2, 3 and 4). Use a graphical technique to determine whether the undergraduate degree and the university each person applied to appear to be related.

- 3.32 XR03-32** Is there brand loyalty among car owners in their purchase of petrol? To help answer the question a random sample of car owners was asked to record the brand of petrol in their last two purchases (1 = BP, 2 = Shell, 3 = Caltex, 4 = Other). Use a graphical technique to formulate your answer.

- 3.33 XR03-33** The cost of smoking for individuals, the companies for whom they work, and society in general is in the many billions of dollars. In an effort to reduce smoking, campaigns about the dangers of smoking have been undertaken by various government and non-government organisations. Most of these have been directed at young people. This raises the question, 'Are you more likely to smoke if your parents smoke?' To shed light on this issue, a sample of 20- to 40-year-old people was asked whether they smoked and whether their parents smoked. The results were recorded in the following way:

Column 1:

1, 2, ..., 225 Parent Identification (ID)

Column 2:

1 = do not smoke

2 = smoke

Column 3:

1 = neither parent smoked

2 = father smoked

3 = mother smoked

4 = both parents smoked

Use a graphical technique to produce the information you need.

Study Tools

CHAPTER SUMMARY

Descriptive statistics is concerned with methods of summarising and presenting the essential information contained in a set of data, whether the set is a population or a sample taken from a population. In this chapter, we presented graphical techniques for nominal data.

Bar charts, pie charts and frequency distributions are used to summarise single sets of *nominal (categorical) data*. Because of the restrictions applied to this type of data, all that we can show is the frequency and proportion of each category. The type of chart to use in a particular situation depends on the particular information the user wants to emphasise.

To describe the relationship between two nominal (categorical) variables, we construct *cross-classification tables* and bar charts.

The data files for Examples, Exercises and Cases are provided on the companion website (accessible through <https://login.cengagebrain.com/>).

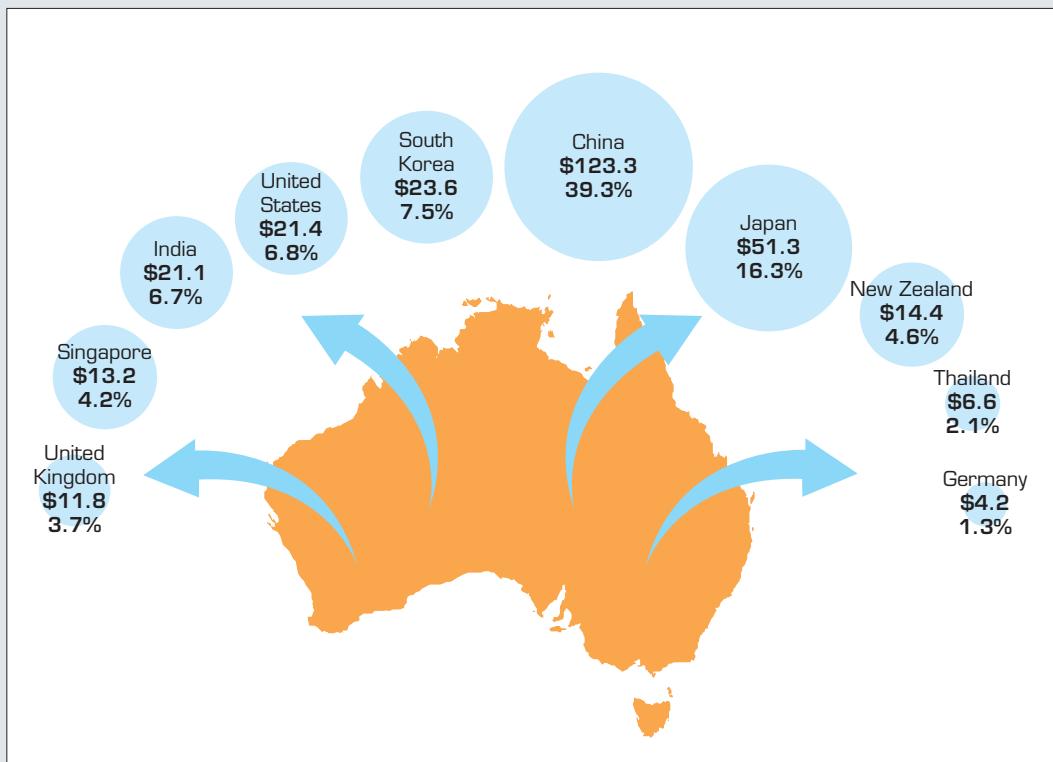
SUPPLEMENTARY EXERCISES

3.34 XR03-34 An increasing number of statistics courses use a computer and software rather than manual calculations. A survey of statistics instructors asked them to report the software used in their courses. The responses are:

- 1 Excel
- 2 Minitab

- 3 SAS
 - 4 SPSS
 - 5 Other
- a Using the recorded data produce a frequency distribution.
 - b Graphically summarise the data so that the proportions are depicted.

Australia's top 10 export markets, 2017–18 (billions of dollars)



Source: Department of Foreign Affairs and Trade, Two way trading partners, <https://dfat.gov.au/trade/resources/trade-at-a-glance/pages/html/two-way-trading-partners.aspx>



- c What do the charts tell you about the software choices?

3.35 XR03-35 Australian exports to the world reached A\$403.2 billion in 2017–18. The top 10 Australian export markets in 2017–18, the value of exports (in billions of dollars) and their share of the total export market are given in the previous diagram.

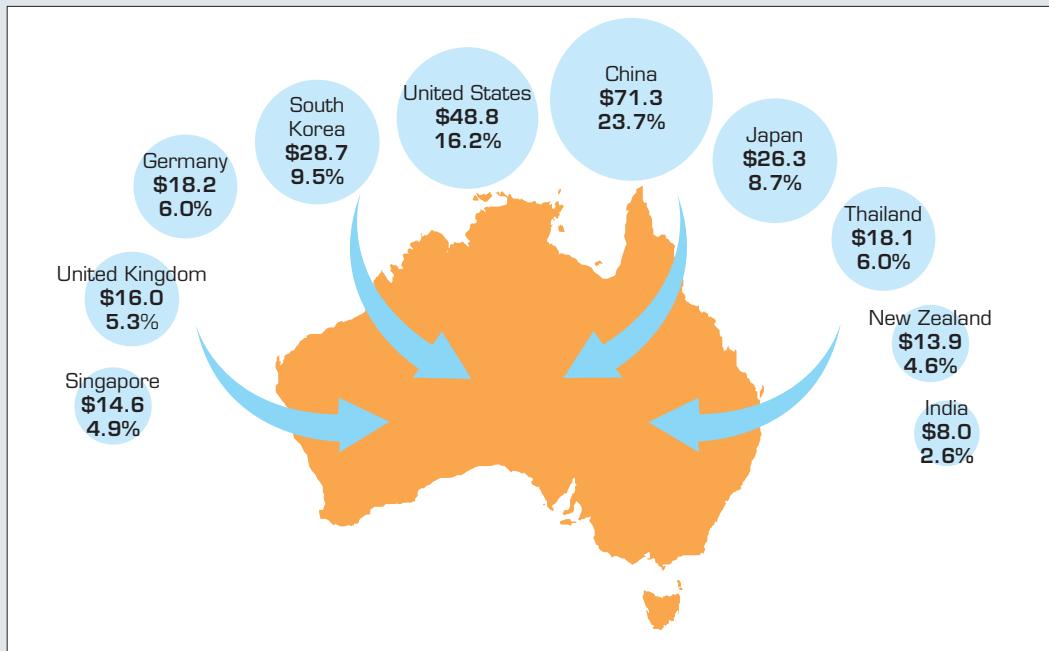
- a Graphically depict the data using a bar chart.
b Graphically depict the data using a pie chart.

- c What do the charts tell you about Australia's top 10 export markets?

3.36 XR03-36 Australia's total imports reached A\$395.4 billion in 2017–18. Australia's top 10 import markets in 2017–18, the value of imports (in billions of dollars) and their share of the total import market are given in the following diagram.

- a Graphically depict the data using a bar chart.
b Graphically depict the data using a pie chart.
c What do the charts tell you about Australia's top 10 import markets?

Australia's top 10 import markets, 2017–18 (billions of dollars)



Source: Department of Foreign Affairs and Trade, Two way trading partners, <https://dfat.gov.au/trade/resources/trade-at-a-glance/pages/html/two-way-trading-partners.aspx>

3.37 XR03-37 The export statistics published by the Australian government each year contain detailed information about the amount of produce exported overseas from each state and territory. The following table presents food and fibre exports for 2014–18. Develop an interesting and informative graphical descriptive method to display the data.

Australian food and fibre exports by state of origin (\$m), 2014–18

State	2014	2015	2016	2017	2018
Victoria	12 153	12 252	12 114	12 767	14 141
NSW	7 514	7 968	8 350	9 736	9 120
Queensland	7 894	9 177	9 078	9 699	8 621
WA	6 330	6 564	6 528	7 088	6 678
SA	5 055	5 444	5 548	5 882	6 304
Tasmania	774	759	844	814	1 051
Others*	3 982	4 578	4 708	5 034	6 216
Total Australia	43 704	46 742	47 170	51 020	52 132

*Australian territories

Source: Victorian Food and Fibre Export Performance Report, 2017–18 © State of Victoria (Department of Environment, Land, Water and Planning) CC BY 4.0 <https://creativecommons.org/licenses/by/4.0/>



3.38 XR03-38 The following table shows the energy consumption (in petajoules or PJ) in Australia by industry in 2010/11, 2012/13 and 2016/17. Display the data graphically and comment on the energy consumption patterns of various sectors during the three years.

Total final energy consumption in Australia (PJ), by sector

Industry	2010/11	2012/13	2016/17
Agriculture	92	102	123
Mining	395	426	478
Manufacturing & construction	1290	1353	951
Transport	594	625	652
Commercial	420	445	333
Residential	1034	1046	1270

Source: Australian Bureau of Statistics, Feb 2018, *Energy Account Australia, 2016–2017*, cat. no. 4604.0, ABS, Canberra.

3.39 XR03-39 The following table lists the top 10 major exporting destinations of Victorian food and fibre, and the total value (in millions of dollars) of exports to these countries for the years 2014–18. Use appropriate graphical techniques to compare the food and fibre exports by destination over the five years, 2014–18.

Top 10 Victorian food and fibre exports by destination, 2014–18

Destination	A\$ million				
	2014	2015	2016	2017	2018
China	2947	2932	3218	3613	4582
Japan	831	966	990	918	1088
United States	646	1139	961	826	973
New Zealand	677	677	736	708	758
Indonesia	632	573	556	560	597
Hong Kong	390	420	498	502	513
Malaysia	489	468	404	419	498
India	312	353	308	629	380
South Korea	369	402	400	385	366
Singapore	468	438	348	354	345

Source: Victorian Food and Fibre Export Performance Report, 2017–18
 © State of Victoria (Department of Environment, Land, Water and Planning)
 CC BY 4.0 <https://creativecommons.org/licenses/by/4.0/>

3.40 XR03-40 The director of the Department of Tourism is preparing a presentation on international inbound tourism to Australia. Following is one of the tables

she intends to show to her audience, which details the number of international inbound tourist arrivals to Australia from its top 20 markets during (the financial year) 2018.

- a Use appropriate graphical techniques to depict the information given in the table.
- b In what way is your graphical display more useful than the tabular presentation?

Overseas arrivals by country of residence, top 20 countries, Australia, 2018

	Country	Number ('000)		Country	Number ('000)
1	China	1432.1	11	Indonesia	208.8
2	NZ	1384.9	12	Germany	207.3
3	US	789.1	13	Taiwan	202.8
4	UK	733.4	14	Canada	182.0
5	Japan	469.2	15	Philippines	143.7
6	Singapore	447.8	16	France	142.6
7	Malaysia	401.1	17	Vietnam	110.9
8	India	357.7	18	Thailand	99.4
9	Hong Kong	308.7	19	Sri Lanka	42.2
10	South Korea	288.0	20	Pakistan	19.6

Source: Australian Bureau of Statistics, August 2019 *Overseas Arrivals and Departures*, cat. no. 3401.0, ABS, Canberra.

3.41 XR03-41 Opinions about the economy are important measures because they can become self-fulfilling prophecies. Annual surveys are conducted to determine the level of confidence in the future prospects of the economy. A sample of 1000 adults was asked the question, 'Compared with last year, do you think this coming year will be: 1, better; 2, the same; or 3, worse?' Use a suitable graphical technique to summarise these data. Describe what you have learnt.

3.42 XR03-42 There are several ways to teach applied statistics. The most popular approaches are:

- 1 emphasise manual calculations
- 2 use a computer combined with manual calculations
- 3 use a computer exclusively with no manual calculations.

A survey of 100 statistics instructors asked each respondent to report his or her approach. Use a graphical method to extract the most useful information about the teaching approaches.

Case Studies

CASE 3.1 Analysing the COVID-19 deaths in Australia by gender and age group

C03-01 The first case of the coronavirus (COVID-19) pandemic in Australia was identified on 23 January 2020. Since then, as of 30 June 2020, there have been 7881 confirmed cases and 103 deaths due to the virus. (* A 36-year-old Australian male diagnosed with COVID-19 who died in Iceland has not been included here.) Due to early lockdown measures, self-isolation of known cases and close contacts, and quarantine controls of international arrivals, Australia has managed to control the virus until now. Data on the number of deaths by states, age group and gender are recorded (source: <https://www.covid19data.com.au/states-and-territories>).

- a Analyse the number of confirmed cases and deaths by states.
- b Analyse the number of deaths by states and age groups.
- c Analyse the data by age groups and gender.

Number of COVID-19 cases and number of deaths, Australia, 30 June 2020		
State	COVID-19 cases	Deaths
New South Wales (NSW)	3211	48
Victoria (VIC)	2195	20
Queensland (QLD)	1079	6
Western Australia (WA)	594	9
South Australia (SA)	436	4
Tasmania (TAS)	225	13
Australian Capital Territory (ACT)	111	3
Norther Territory (NT)	30	0
Australia	7881	103

Number of COVID-19 deaths by age group, state/territory and gender, Australia, 30 June 2020											
Age group	State/Territory								Gender		
	NSW	VIC	QLD	WA	SA	TAS	ACT	NT	Australia	Male	Female
40–49 yrs				1					1	1	
50–59 yrs	1	1							2	1	1
60–69 yrs	4	3	1		1			1	10	5	5
70–79 yrs	11	6	3	6	3	5			34	22	12
80–89 yrs	16	9	2	2		5	2		36	19	17
90–99 yrs	16	1				3			20	8	12
Total	48	20	6	9	4	13	3	0	103	56	47

Source: covid19data.com.au. CC BY 4.0 <https://creativecommons.org/licenses/by/4.0/>

CASE 3.2 Corporate tax rates around the world

C03-02 Many countries are lowering taxes on corporations in an effort to make their countries more attractive for investment. In the following table, we list the marginal effective corporate tax rates among Organisation for Economic Co-operation and Development (OECD) countries. Develop a graph that depicts these figures. Discuss your observations from the graph.

Country	Manufacturers	Services	Aggregate
Australia	27.7	26.6	26.7
Austria	21.6	19.5	19.9
Belgium	-6.0	-4.1	0.5
Canada	20.0	29.2	25.2
Czech Republic	1.0	7.8	8.4
Denmark	16.5	12.7	13.4
Finland	22.4	22.9	22.8
France	33.0	31.7	31.9
Germany	30.8	29.4	29.7
Greece	18.0	13.2	13.8
Hungary	12.9	12.0	12.2
Iceland	19.5	17.6	17.9
Ireland	12.7	11.7	12.0
Italy	24.6	28.6	27.8
Japan	35.2	30.4	31.3
Korea	32.8	31.0	31.5
Luxembourg	24.1	20.3	20.6
Mexico	17.1	12.1	13.1
Netherlands	18.3	15.0	15.5
New Zealand	27.1	25.4	25.7
Norway	25.8	23.2	23.5
Poland	14.4	15.0	14.9
Portugal	14.8	16.1	15.9
Slovakia	13.3	11.7	12.0
Spain	27.2	25.2	25.5
Sweden	19.3	17.5	17.8
Switzerland	14.8	15.0	14.9
Turkey	22.7	20.2	20.8
UK	22.7	27.8	26.9
US	32.7	39.9	36.9

CASE 3.3 Trends in CO₂ emissions

C03-03 The carbon dioxide (CO₂) emissions for the top 20 CO₂-producing countries and the CO₂ emissions per capita are recorded. Using appropriate graphical techniques, present the data. Discuss your observations from the graphs.

CO₂ emissions and CO₂ emissions per capita, top 20 countries, 2018

	Country	CO ₂ emissions (1000 megaton)	CO ₂ emissions per capita (ton)
1	China	11.2	7.8
2	United States	5.25	15.9
3	India	2.59	1.9
4	Russia	1.73	11.6
5	Japan	1.20	9.5
6	Germany	0.75	9.0
7	Iran	0.71	8.6
8	South Korea	0.69	13.5
9	Saudi Arabia	0.60	17.8
10	Canada	0.59	15.6
11	Indonesia	0.54	2.0
12	Mexico	0.49	3.9
13	Brazil	0.49	2.2
14	South	0.48	7.6
15	Australia	0.41	3.9
16	Turkey	0.41	5.0
17	United Kingdom	0.37	5.5
18	Italy	0.34	5.6
19	Poland	0.33	8.8
20	France	0.32	5.0

Source: Olivier J. G. J. and Peters J. A. H. W. (2019), *Trends in global CO₂ and total greenhouse gas emissions: 2019 report*. PBL Netherlands Environmental Assessment Agency, The Hague.

CASE 3.4 Where is the divorce rate heading?

C03-04 The number of people in Australia getting divorced is increasing at an alarming rate. Divorce does not only affect the couple involved: researchers have found that it has a chain effect on their children and other family members. The following table shows the age-specific divorce rates per married population among men over selected years: 1981, 1991, 2001, 2011 and 2016. Use a number of appropriate graphical presentations to display the information.

Age-specific divorce rates per married male population, Australia

Age group	1981	1991	2001	2011	2016
24 and under	14	11	13	18	10
25–29	22	22	22	18	16
30–34	19	20	24	17	14
35–39	16	18	21	17	14
40–44	14	15	19	17	15
45–49	11	13	17	16	15
50–54	8	9	13	15	14
55–59	5	6	9	10	10
60–64	4	4	6	7	7
65 and over	2	2	2	2	3

Source: Australian Bureau of Statistics, *Marriages and Divorces, Australia*, 2017, cat. no. 3310.0, ABS, Canberra.

CASE 3.5 Geographic location of share ownership in Australia

C03-05 In the past, surveys indicated a disparity in the level of share ownership between Australia's states and territories. Today, this disparity has disappeared, with nearly half the population in most states having some form of share ownership. The following table indicates the geographic location of share ownership for selected years in five Australian states. Use any appropriate graphical technique to present the information.

Percentage of total share ownership in Australia in five selected states for selected years

State/Territory	Percentage of share ownership				
	1991	1997	2004	2010	2014
NSW/ACT	14	32	46	43	36
Victoria	18	36	43	39	35
Queensland	15	37	42	36	30
South Australia	16	33	41	32	27
Western Australia	18	34	48	41	33

CASE 3.6 Differing average weekly earnings of men and women in Australia

C03-06 Although a lot has been achieved in Australia to reduce the difference between men and women in a number of social status indicators, wage differences are still a matter of concern. The following table presents the average weekly cash earnings of male and female adults for each Australian state and territory and for Australia as a whole. Present the information using appropriate graphical techniques.

Average weekly (all full-time adult employees total) cash earnings (A\$), 2018

State/Territory	Males	Females
New South Wales	1807.80	1502.00
Victoria	1696.20	1490.30
Queensland	1777.60	1409.50
South Australia	1585.40	1393.30
Western Australia	2040.00	1495.30
Tasmania	1565.50	1320.00
Northern Territory	1877.90	1552.60
ACT	1977.90	1671.60
Australia	1823.70	1534.10

Source: Australian Bureau of Statistics, *Average Weekly Earnings*, February 2019, cat. no. 6302.0, ABS, Canberra.

CASE 3.7 The demography of Australia

C03-07 There are six states and two territories in Australia, each with a different population size and land area. In terms of land area, Western Australia is the biggest state, while New South Wales ranks fifth. However, in terms of population, New South Wales has the largest population, while Western Australia ranks fourth. Another interesting observation is that Aboriginal and Torres Strait Islander peoples make up only 3% of the total Australian population. The following table presents information about the land size and population, including Aboriginal and Torres Strait Islander populations, of each Australian state and territory.

Australian population size (including the Aboriginal and Torres Strait Islander population) and land area by state and territory, December 2018

State/Territory	Area ('000 km ²)	Australian population ('000s)	Aboriginal and Torres Strait Islander population ('000s)
New South Wales	802	8 046	266
Victoria	228	6 526	58
Queensland	1 727	5 053	221
South Australia	984	1 743	42
Western Australia	2 526	2 606	101
Tasmania	68	532	29
Northern Territory	1 346	246	75
ACT	2	424	8
Total (Australia)	7 683	25 176	800

Note: The latest population statistics available for Aboriginal and Torres Strait Islander peoples is for June 2016.

Source: Australian Bureau of Statistics, Dec 2018, *Australian Demographic Statistics*, cat. no. 101.0, ABS, Canberra.

Use appropriate graphical techniques to show:

- a the area of each state and territory as a proportion of the total area of Australia
- b the distribution of the Australian population across the different states and territories
- c the distribution of the Aboriginal and Torres Strait Islander population in each state and territory as a proportion of the total Aboriginal and Torres Strait Islander population
- d the distribution of the Aboriginal and Torres Strait Islander population in each state and territory as a proportion of the overall Australian population in the corresponding state and territory.

CASE 3.8 Survey of graduates

C03-08 A survey of business school graduates undertaken by a university placement officer asked, among other questions, in which area each person was employed. The areas of employment are:

- accounting
- finance
- general management
- marketing/sales
- other

Additional questions were asked and the responses were recorded in the following way:

Column	Variable
1	Identification number
2	Area
3	Gender (1 = female, 2 = male)
4	Job satisfaction (4 = very, 3 = quite, 2 = little, 1 = none)
5	Number of weeks job searching
6	Salary (\$'000)

The placement officer wants to know the following:

- a Do female and male graduates differ in their areas of employment? If so, how?
- b Are area of employment and job satisfaction related?
- c Are salary and number of weeks needed to obtain the job related?

CASE 3.9 Analysing the health effect of the coronavirus pandemic

C03-09 The coronavirus (COVID-19) pandemic commenced in Wuhan, China, in December 2019 and had spread to more than 200 countries worldwide within 3 months. The devastating effect of the virus can be seen from the number of deaths in a number of countries. The number of confirmed cases and number of deaths for 208 countries and the total were recorded as at 5 May 2020. Present appropriate individual charts to display the number of confirmed cases and number of deaths for the top 15 countries. Also present the share contributed by each of the top 15 countries in the confirmed cases and deaths of the global COVID-19 pandemic.

Source: Coronavirus Source Data, <https://ourworldindata.org/coronavirus-source-data>, 5 May 2020.

CASE 3.10 Australian domestic and overseas student market by states and territories

C03-10 The contribution of the overseas student market to most Australian educational institutions cannot be underestimated. To see how the overseas student numbers have increased over time, we collected data for commencing domestic and overseas students in the years 1989, 1999, 2009 and 2018. The data are given below. Using appropriate graphical techniques, analyse the data in relation to the individual states and territories, and compare the change in the domestic and overseas student market share between 1989 and 2018.

Commencing domestic and overseas student numbers by state/territory, Australia, 1989, 1999, 2009 and 2018

State or Territory	Domestic students				International students			
	1989	1999	2009	2018	1989	1999	2009	2018
New South Wales	56 117	89 382	150 513	196 832	2 874	13 548	43 309	68 801
Victoria	52 246	71 434	114 981	179 062	2 935	14 276	46 737	82 685
Queensland	26 731	51 051	88 855	102 692	1 242	7 341	28 622	30 149
Western Australia	18 507	26 401	50 045	55 501	1 565	5 546	18 040	15 033
South Australia	15 401	19 084	30 467	44 613	554	2 616	10 233	12 893
Tasmania	3 916	5 190	9 440	16 838	192	434	2 072	3 181
Northern Territory	1 588	2 553	3 895	4 703	60	99	252	676
Australian Capital Territory	6 586	7 355	13 196	17 828	470	1 031	3 521	6 692
Multi-State		3 954	9 145	13 785		121	1 783	2 374
Australia (Total)	181 092	276 404	470 537	631 854	9 892	45 012	154 569	222 484

Source: © Commonwealth of Australia. CC BY 4.0 International <https://creativecommons.org/licenses/by/4.0/>

CASE 3.11 Road fatalities in Australia

C03-11 Road fatalities are a major concern to the Australian government. Road accidents not only cause driver fatalities, but also the death of many road users such as passengers in vehicles, pedestrians, motorcyclists and cyclists. The following tables present data for the number of deaths due to road accidents categorised by road user groups, age group and gender from 2016–2020.

- a Analyse the data by all road user groups
- b Analyse the data by gender
- c Analyse the data by day of the week
- d Analyse the data by time of day.

Road fatalities by road user group for Australia, 2016–2020 (year ending March)

Year	Number of road deaths by road user and gender, 2016–2020 (year ending March)							
	Number by type of road user					Total	Number by gender	
	Driver	Passenger	Pedestrian	Motorcyclist	Pedal cyclist		Males	Females
2016	582	242	163	223	35	1 247	900	347
2017	600	210	172	226	21	1 234	909	325
2018	583	235	181	212	43	1 263	919	344
2019	529	211	170	209	37	1 161	874	287
2020	570	197	159	184	41	1 154	886	268

Year ^c	Number of road deaths by day of the week and time of day			
	Day of week ^a		Time of day ^b	
	Weekday	Weekend	Daytime	Night
2016	721	423	715	429
2017	693	457	682	468
2018	720	440	739	421
2019	653	422	663	412
2020	672	396	651	417

a 'Weekday' refers to 6 am Monday through to 5:59 pm Friday

b 'Day' refers to 6 am through to 5:59 pm

c Some accidents did not record the day of the week, or time of day are excluded.

Source: © Commonwealth of Australia. CC BY 4.0 International <https://creativecommons.org/licenses/by/4.0/>

CASE 3.12 Drinking behaviour of Australians

C03-12 Drinking alcohol is very much part of the Australian culture. The legal drinking age in Australia is 18. To identify whether there are differences in the level of alcohol consumption by various age groups, we recorded data from the National Drug Strategy Household Survey 2016, an online survey of 23 722 respondents, for the number of standard drinks usually consumed by Australians on a single occasion by age group. Use appropriate graphical techniques to present the following:

- a** age distribution by number of drinks
- b** distribution of level of drinking by age group.

Number of standard drinks consumed in a single occasion by age group, Australia

Age group	Abstained	<2 drinks	3–4 drinks	5–6 drinks	7–10 drinks	11–12 drinks	>13 drinks
14–17 years	75	12	4	3	4	0.6	1
18–24 years	19	26	18	12	16	3	6
25–29 years	20	35	20	10	10	1	3
30–39 years	17	42	22	8	7	1	3
40–49 years	17	44	22	9	6	1	2
50–59 years	18	47	20	7	6	0.7	1
60–69 years	22	50	17	7	3	0.6	0.4
70+ years	31	54	11	2	2	0	0

Graphical descriptive techniques – Numerical data

Learning objectives

This chapter discusses the graphical descriptive methods used to summarise and describe sets of numerical data. It also discusses graphical excellence and deception.

At the completion of this chapter, you should be able to:

- L01** tabulate and draw charts and graphs to summarise numerical data
- L02** use graphs to analyse time-series data
- L03** use various graphical techniques to analyse the relationships between two or more numerical variables
- L04** understand deception in graphical presentation
- L05** understand how to present statistics in written reports and oral presentations.

CHAPTER OUTLINE

Introduction

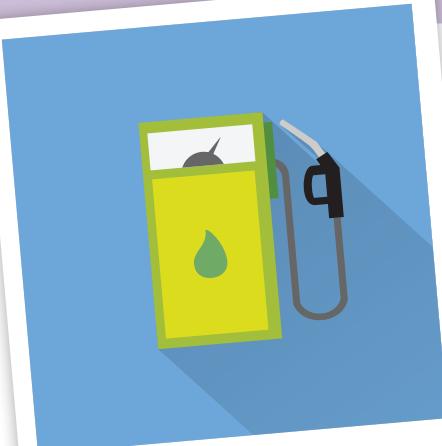
- 4.1** Graphical techniques to describe numerical data
- 4.2** Describing time-series data
- 4.3** Describing the relationship between two or more numerical variables
- 4.4** Graphical excellence and deception

SPOTLIGHT ON STATISTICS

Were oil companies gouging Australian customers?

The price of oil has been increasing for several reasons. First, oil is a finite resource; the world will eventually run out. In early 2019, the world was consuming more than 100 million barrels of oil per day – more than twice the oil consumed 50 years ago. The total proven world reserves of oil are about 1.665 trillion barrels. At today's consumption levels the proven reserves will be exhausted in 50 years. [It should be noted, however, that in 2010, the proven reserves of oil amounted to 1.474 trillion barrels, indicating that new oil discoveries are offsetting increasing usage.] Second, industries in China and India are rapidly expanding and require ever-increasing amounts of oil. Third, over the last 15 years hurricanes have threatened the oil rigs in the Gulf of Mexico.

The price increases in crude oil are reflected in the retail price of petrol. In 2002, the average retail price of petrol per litre in Australia overall was 87.3 cents. For the three major states of Australia, the price was 88.4 cents for New



Source: iStock.com/Dicraftsman

► South Wales (NSW), 87.6 cents for Victoria and 80.9 cents for Queensland, and the price of oil (Dubai Fetch crude) was A\$45.77 per barrel. (A barrel is equal to 159.18 litres). Since then, the prices of both crude oil and petrol have increased substantially. By December 2019, the average retail price of petrol in Australia overall had increased to 142.0 cents, and in NSW to 141.1 cents, Victoria to 141.1 cents and Queensland to 142.8 cents per litre, and the price of crude oil had increased to A\$85.70 per barrel.

Many drivers complained that the oil companies were guilty of price gouging. That is, they believed that when the price of crude oil increased, the price of petrol also increased, but when the price of crude oil decreased, the decrease in the price of petrol seemed to lag behind. To determine whether this perception is accurate we collected the annual price data for both commodities. These data for crude oil price and petrol prices in Australia and the three Australian states of NSW, Victoria and Queensland are stored in file **CH04\XM04-00**. Were crude oil and petrol prices related? On pages 116–17 we provide a possible answer.

Source: <https://fleetautonews.com.au/historical-pump-prices-in-australia/>

Introduction

Chapter 3 introduced graphical techniques used to summarise and present nominal data. In this chapter, we do the same for *numerical data*. Section 4.1 presents techniques to describe a set of numerical data, Section 4.2 introduces time series and the methods used to present time-series data, and Section 4.3 describes the techniques we use to describe the relationship between two numerical variables. We complete this chapter with a discussion of how to properly use graphical techniques in Section 4.4.

4.1 Graphical techniques to describe numerical data

histogram

Graphical presentation of a frequency distribution of numerical data.

In this section, we introduce several tabular and graphical techniques that are used when the data are numerical. The most important of these graphical techniques is the **histogram**, which is developed from the tabular representation known as the *frequency distribution*. As you will see, the histogram is not only a powerful graphical technique used to summarise numerical data, but it is also used to help explain an important aspect of probability (see Chapter 8).

First, however, we will present an example through which we will introduce the tabular method, the frequency distribution, and then the graphical method, the histogram.

REAL-LIFE APPLICATIONS

Pricing

Following the deregulation of the electricity market in Queensland, several new companies entered the market to compete in the business of providing electricity for individual households. In almost all cases, these providers competed on price, since the service each offered was similar. Pricing a service or product in the face of stiff competition is very difficult.

Factors to be considered include supply, demand, price elasticity and the actions of competitors. Determining the appropriate rate structure is facilitated by acquiring information about the behaviour of customers and, in particular, the size of the electricity bills.

EXAMPLE 4.1

L01

An analysis of electricity bills

XM04-01 In the past few years, a number of companies have been created to compete in the electricity retail market. Suppose, as part of a much larger study, one such electricity provider in Brisbane wanted to get some information concerning the monthly bills of new customers in the first month after signing on with the company. A survey of 200 new Brisbane residential customers was undertaken, and their first month's bills were recorded. These data appear in **Table 4.1**. As a first step, the marketing manager of the company wanted to summarise the data in preparation for a presentation to the board of directors of the company. What information can be extracted from the data?

Solution**Identifying the technique and calculating manually****TABLE 4.1** Monthly electricity bills (in dollars) for 200 new Brisbane customers

196.65	258.50	232.60	257.00	125.60	119.50	166.10	180.50	160.00	175.80
468.75	310.20	262.92	299.20	110.25	325.40	234.50	212.00	228.55	73.05
320.50	244.40	85.75	70.50	77.35	102.20	220.75	154.75	248.70	460.50
300.50	194.50	248.00	166.50	161.30	200.00	212.85	100.50	208.70	263.60
213.05	210.20	204.75	100.40	313.50	361.00	436.50	452.60	466.05	59.50
140.60	360.00	310.70	240.00	222.00	430.00	403.00	436.35	219.00	198.00
290.00	456.50	213.10	240.00	388.10	240.00	460.50	225.00	216.00	416.50
216.95	237.40	320.50	224.05	110.50	250.50	220.00	124.30	425.50	315.50
360.50	235.00	194.00	247.00	160.00	470.00	103.50	170.00	390.00	155.00
317.95	203.25	220.00	184.00	210.00	157.75	222.15	127.35	176.85	190.00
195.55	109.20	186.75	153.50	310.30	98.40	170.50	107.90	240.50	158.50
220.50	240.15	97.80	219.50	380.10	236.50	224.15	140.00	226.00	225.00
255.60	260.50	340.50	214.15	281.00	230.85	460.00	195.00	108.70	266.70
289.00	275.50	88.50	155.20	105.35	317.65	260.40	315.10	160.50	153.60
194.55	101.55	209.50	140.40	280.15	200.70	200.50	241.05	470.50	238.00
374.25	455.50	234.04	108.50	188.80	165.00	311.40	168.00	225.00	297.60
382.05	246.25	333.00	410.00	272.50	350.50	260.00	120.50	440.00	201.75
185.55	291.55	291.10	125.50	103.40	319.15	251.55	223.95	265.00	240.50
219.10	262.00	108.50	220.30	213.50	275.88	100.60	237.05	162.80	270.90
215.60	378.65	245.00	160.00	280.50	203.05	212.20	285.45	260.50	197.30

Very little information about the monthly bills can be acquired by casually reading through the 200 observations in **Table 4.1**. The manager can probably see that most of the bills are for greater than \$50, but that is likely to be the extent of the information garnered by browsing through the data. If you examine the data more carefully, you may discover that the smallest bill is \$59.50 and the largest is \$470.50. To gain useful information, we need to know how the bills are distributed between \$59.50 and \$470.50. Are there many small bills and some large bills, or are most of the bills in the centre of the range with very few extreme values? What is the amount of the 'typical bill'? Are the bills similar or do they vary considerably? To help answer these questions and others like them, we will construct a frequency distribution from which a histogram can be drawn. In the previous chapter, a frequency distribution was created by counting the number of times each category of the nominal variable occurred. We create a frequency distribution for numerical data by counting the number of observations that fall into each of a series of intervals, called classes that cover the complete range of observations. We discuss how to decide the number of classes and the upper and lower limits of the intervals later. We have



chosen nine classes defined in such a way that each observation falls into one and only one class. These classes are defined as follows:

- Amounts that are more than 50 but less than or equal to 100 (50 up to 100)
- Amounts that are more than 100 but less than or equal to 150 (100 up to 150)
- Amounts that are more than 150 but less than or equal to 200 (150 up to 200)
- Amounts that are more than 200 but less than or equal to 250 (200 up to 250)
- Amounts that are more than 250 but less than or equal to 300 (250 up to 300)
- Amounts that are more than 300 but less than or equal to 350 (300 up to 350)
- Amounts that are more than 350 but less than or equal to 400 (350 up to 400)
- Amounts that are more than 400 but less than or equal to 450 (400 up to 450)
- Amounts that are more than 450 but less than or equal to 500 (450 up to 500)

Notice that the intervals do not overlap, so there is no uncertainty about which interval to assign to any observation. Moreover, because the smallest number is 59.5 and the largest is 470.5, every observation will be assigned to a class. Finally, the intervals are equally wide. Although this is not essential, it makes the task of reading and interpreting the graph easier. To create the frequency distribution manually, we count the number of observations that fall into each interval. **Table 4.2** presents the frequency distributions.

TABLE 4.2 Frequency distribution of electricity bills for 200 new Brisbane customers

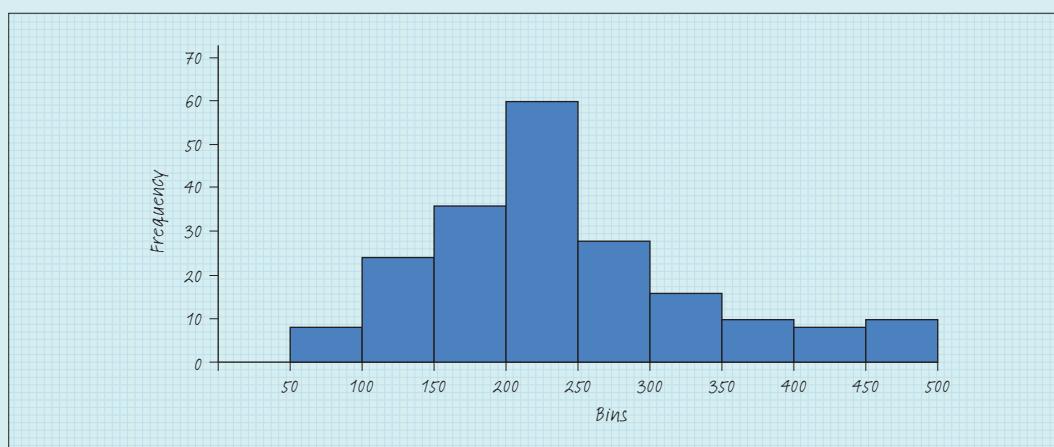
Class limit*	Tally	Frequency
50 up to 100		8
100 up to 150		24
150 up to 200		36
200 up to 250		60
250 up to 300		28
300 up to 350		16
350 up to 400		10
400 up to 450		8
450 up to 500		10
Total		200

* Classes contain observations greater than their lower limits (except for the first class) and less than or equal to their upper limits.

Although the frequency distribution provides information about how the numbers are distributed, the information is more easily understood and imparted by drawing a picture or graph. The graph is called a *histogram*. A histogram is created by drawing rectangles whose bases are the intervals and whose heights are the frequencies.

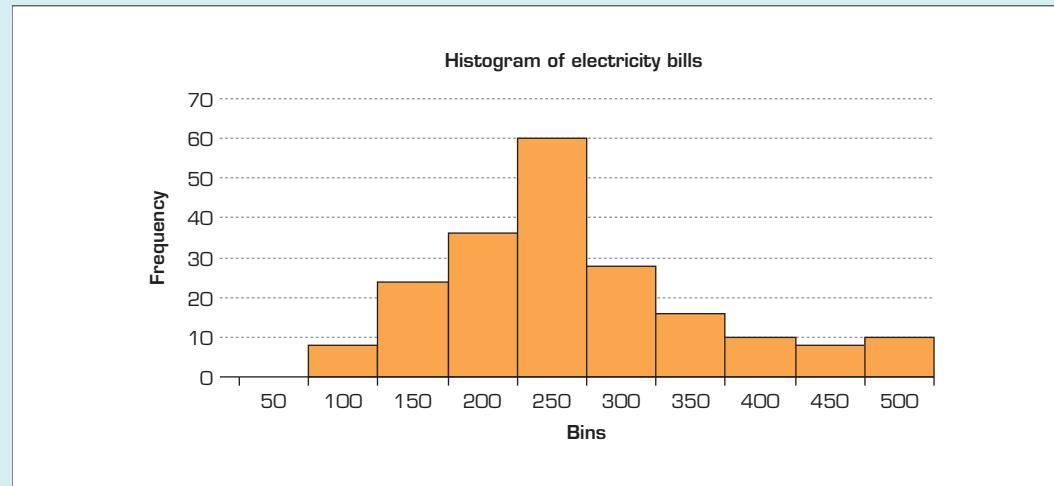
Figure 4.1 depicts the manually drawn histogram constructed from the frequency distribution in **Table 4.2**. We call this diagram a frequency histogram because the numbers on the vertical axis represent the frequencies.

Figure 4.1 was drawn by hand. We now show how histograms are actually drawn in practice, using Excel to do the work for us.

**FIGURE 4.1** Histogram of electricity bills of 200 new Brisbane customers

Using the computer

Excel output for Example 4.1



Note: In an Excel output, the numbers that appear along the horizontal axis represent the upper limits of the class intervals even though they appear in the centre of the classes on the histogram.

COMMANDS

- 1 Type the data in column A or open the data file (**XM04-01**).
- 2 Type **Bins** in cell B1.
- 3 In B2, type the upper limit of the first class interval. In B3, type the upper limit of the second interval, and so on, in order to complete the listing of bins (**50 100 150 ... 500**).
- 4 Click **DATA, Data Analysis** and **Histogram**. Click **OK**.
- 5 For **Input Range**, type the location of the data (including cell containing name, if necessary) (**A1:A201**).
- 6 Click the box for **Bin Range** and type in the location of the bins (**B1:B11**).
- 7 If the name of the variable and bins have been included in the first row of the data ranges, click the check box for **Labels**.
- 8 Output will normally be stored in a new sheet of the workbook. To store the output on the same sheet (Sheet 1), in the **Output Options** select **Output Range** and type the starting cell reference in the space provided (**D1**). Also click the checkbox for **Chart Output** and then click **OK**. This will produce a histogram.

- 9 You will notice gaps between each bar on the histogram. To remove the gaps, first right-click the mouse on one of the rectangular bars. You will notice little circles appearing in the corners of each bar, together with a menu.
- 10 From this menu select **Format Data Series...** A menu will appear on the right-hand side of the screen.
- 11 Under the **Series Options** tab, move the pointer to **Gap Width** and use the slider to change the number from 150 to 0. Click the cross at the top left corner of the menu to close it.
Note that except for the first class, Excel counts the number of observations in each class that are greater than the lower limit and less than or equal to the upper limit.
You may wish to improve the appearance of the histogram by typing a new caption to replace **Bin**. You can do so by first clicking on the word **Bin** on the histogram to select it. To edit the word, click again and type a new caption (e.g. **Bills** for Example 4.1).
Note: If you don't list the bins as in steps 2 and 3, Excel will automatically construct its own set of bins from which the histogram will be constructed.

Interpreting the results

The histogram gives us a clear view of the way the bills are distributed. Most of the bills are in the low to middle range (\$100–350) and a relatively small number of electricity bills are at the high end of the range. The company needs to know more about the electricity consumption patterns of the low and high users. With the additional information, the marketing manager may suggest an alteration of its pricing for the different levels of electricity consumption.

4.1a Frequency distribution and classes

classes
Non-overlapping intervals of numerical data.

A *frequency distribution* is an arrangement or table that groups data into non-overlapping intervals called **classes**, which cover the complete range of observations, and record the number of observations in each class.

We will now discuss how to decide on the approximate number of classes we used to build a frequency distribution such as the one presented in **Table 4.2**.

4.1b Determining the number of classes

As **Table 4.2** illustrates, a set of data presented in the form of a frequency distribution is more manageable than the original set of raw data, although some of the detailed information is lost. For example, the information in the frequency distribution does not allow us to say anything about the actual values within each class.

When constructing a frequency distribution, we must first decide upon the appropriate number and size of classes to use. The number of class intervals we select depends entirely on the number of observations in the data set. The more observations we have, the larger the number of class intervals we need to use to draw a useful histogram. **Table 4.3** provides guidelines on choosing the number of classes. In Example 4.1, we had 200 observations. The table tells us to use 7, 8, 9 or 10 classes.

TABLE 4.3 Approximate number of classes in frequency distributions

Number of observations (n)	Number of classes (K)
Less than 50	5–7
50–200	7–9
200–500	9–10
500–1000	10–11
1000–5000	11–13
5000–50000	13–17
More than 50000	17–20

Although the appropriate number of classes to use is ultimately a subjective decision, an alternative to the guidelines listed in **Table 4.3** is given by Sturge's formula, which recommends that the approximate number of classes K required to accommodate n observations:

$$K = 1 + 3.3 \log_{10} n$$

In Example 4.1, $n = 200$ and $K = 1 + 3.3 \log_{10} 200 = 1 + 3.3(2.3) = 8.6$. Therefore, we can use either 8 or 9 classes. We chose to use 9 classes.

Class interval width

Once the number of classes to be used has been chosen, the class width is calculated by taking the difference between the largest and smallest observations and dividing it by the number of classes. Thus:

$$\text{Class width} = \frac{\text{Largest value} - \text{Smallest value}}{\text{Number of classes}}$$

In Example 4.1, we calculate

$$\text{Class width} = \frac{470.50 - 59.50}{9} = 45.67$$

For convenience, we round this number to 50. This is an acceptable action because there is no fixed rule about the number of class intervals, which ultimately determines the class width. The classes are the intervals 'more than 50 but equal to or less than 100', 'more than 100 but equal to or less than 150', and so on. Notice that the class intervals were created so that there was no overlap. For example, the value 100 was included in the first class but is not included in the second. We then count the number of observations that fall into each class interval (or, more precisely, we let the computer count the observations). The counts, or frequencies, of observations are then listed next to their respective classes. This tabular representation of the data, as shown in **Table 4.2**, is called the *frequency distribution*. The next step in the information-mining process is to draw a picture of the data by constructing a histogram.

Exceptions to the guidelines

In some situations, the guidelines we have just provided may not yield useful results. One such example occurs when there is a wide range of values with very large numbers of observations in some class intervals. This may result in some empty or nearly empty classes in the middle of the histogram. One solution is to create unequal class intervals. Unfortunately, this produces histograms that are more difficult to interpret. Another solution is to allow the last interval to be 'more than the upper limit of the previous interval'.

4.1c Histograms

Although the frequency distribution provides information about the way the numbers are distributed, the information is more easily understood and imparted by drawing a picture or graph. The graph is called a *histogram*. A histogram is created by drawing rectangles whose bases correspond to the class intervals; the area of each rectangle equals the number of observations in that class. When the class widths are equal (which is usually the case), the height of each rectangle will be proportional to the number of observations in that class.

4.1d Relative frequency histograms

class relative frequency

Percentage of data in each class.

relative frequency histogram

A histogram with the vertical axis representing relative frequencies.

Instead of showing the absolute frequency of observations in each class, it is often preferable to show the proportion (or percentage)¹ of observations falling into the various classes. To do this, we replace the class frequency by the **class relative frequency**, which is calculated as follows:

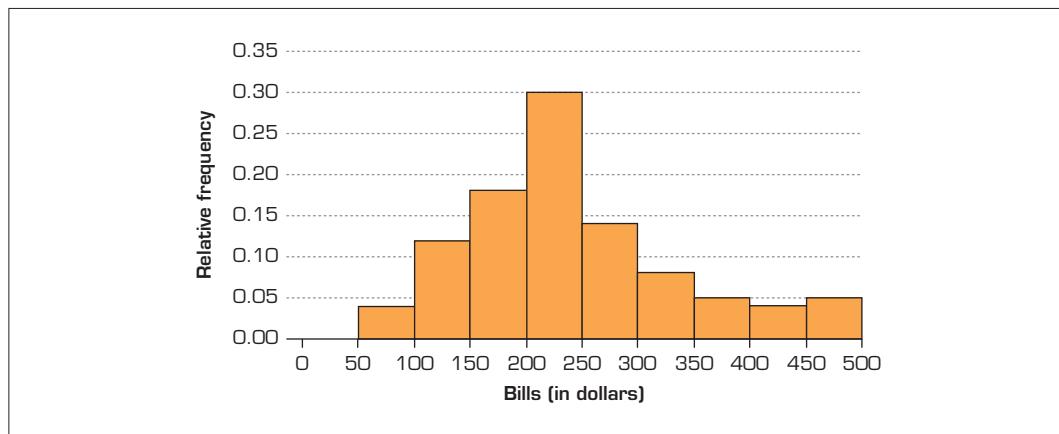
$$\text{Class relative frequency} = \frac{\text{Class frequency}}{\text{Total number of observations}}$$

Returning to Example 4.1, we can then talk about a *relative frequency distribution* (see **Table 4.4**) and a **relative frequency histogram** (see **Figure 4.2**). Notice that, in **Figure 4.2**, the area of any rectangle is proportional to the relative frequency, or proportion, of observations falling into that class. In any relative frequency distribution, the sum of all the relative frequencies is always equal to 1.

TABLE 4.4 Relative frequency distribution of electricity bills for 200 new customers

Class limits	Relative frequency
50 up to 100	8/200 = 0.04
100 up to 150	24/200 = 0.12
150 up to 200	36/200 = 0.18
200 up to 250	60/200 = 0.30
250 up to 300	28/200 = 0.14
300 up to 350	16/200 = 0.08
350 up to 400	10/200 = 0.05
400 up to 450	8/200 = 0.04
450 up to 500	10/200 = 0.05
	Total = 1.00

FIGURE 4.2 Relative frequency histogram of electricity bills with equal class widths



These relative frequencies are useful when you are dealing with a sample of data, because they provide insights into the corresponding relative frequencies for the population from which the sample was taken. Furthermore, relative frequencies should be used when you are comparing histograms or other graphical descriptions of two or more data sets.

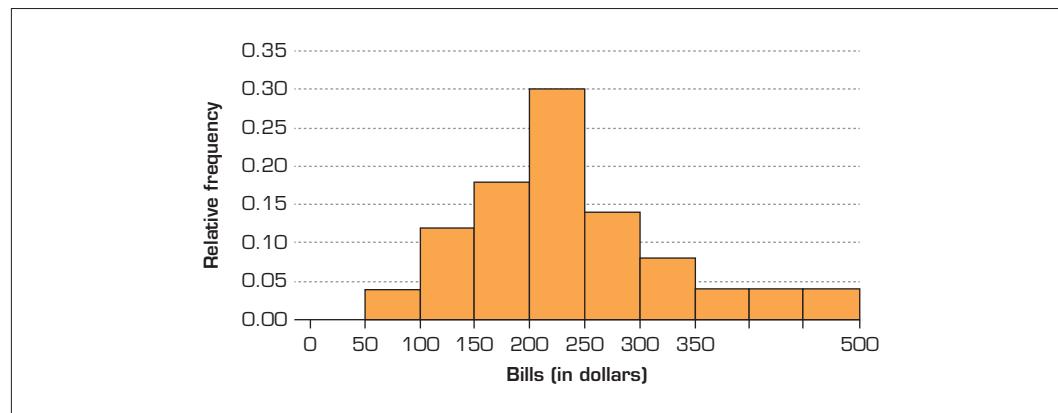
¹ Over the course of this book, we express relative frequencies (and later probabilities) variously as decimals, fractions and percentages.

Relative frequencies permit a meaningful comparison of data sets even when the total numbers of observations in the data sets differ.

4.1e Histograms with unequal class widths

To facilitate interpretation of a histogram, it is generally best to use equal class widths whenever possible. In some cases, however, *unequal class widths* are called for to avoid having to represent several classes with very low relative frequencies. For example, recalling the electricity bills in Example 4.1, suppose that, instead of having 5% of the electricity bills falling between \$450 and \$500, we have 14% of the bills sparsely scattered between \$350 and \$500. This new situation might best be represented by the relative frequency histogram shown in **Figure 4.3**, where the uppermost three classes have been combined. It is important, however, that the height of the corresponding rectangle be adjusted (from 0.14 to 0.14/3) so that the area of the rectangle remains proportional to the relative frequency of all observations falling between \$350 and \$500.

FIGURE 4.3 Relative frequency histogram of electricity bills with unequal class widths



The observations in the preceding example all fall within a fairly compact range. In some cases, however, observations may be sparsely scattered over a large range of values at either end of the distribution. If this situation arises, it may be necessary to use an open-ended class to account for the observations, as shown in the relative frequency distribution in **Table 4.5**. Because weekly individual incomes ranging from \$5000 to hundreds of thousands of dollars are scattered fairly sparsely over a wide range of values, we use a single class with no specified upper limit to capture these incomes. Notice also that **Table 4.5** makes use of unequal class widths, which were discussed in the preceding paragraph.

TABLE 4.5 Percentage distribution of Australian weekly household income, 2016*

Weekly Income range (\$)	Relative frequency (%)
Under \$400	4.5
\$400 up to \$800	19.6
\$800 up to \$1200	13.7
\$1200 up to \$1600	11.5
\$1600 up to \$2000	10.4
\$2000 up to \$3000	18.7
\$3000 up to \$5000	15.7
Over \$5000	5.7

* up to means 'does not include the upper limit of the range'; 2016 refers to the financial year 2015–16.

Source: Australian Bureau of Statistics, 2015–16, *Household Income and Income Distribution, Australia, 2015–2016*, cat. no. 6523.0, ABS, Canberra.

4.1f Stem-and-leaf display

One of the drawbacks of the histogram is that we lose potentially useful information by classifying the observations into groups. In Example 4.1, we learnt that there are 156 observations that fall between 50 and 300, while 44 observations fall between 300 and 500. By classifying the observations we did acquire useful information. However, the histogram focuses our attention on the frequency of each class and by doing so sacrifices whatever information was contained in the actual observations.

stem-and-leaf display

Display of data in which the stem consists of the digits to the left of a given point and the leaf the digits to the right.

The statistician John Tukey introduced a method of organising numerical data called the **stem-and-leaf display**, which is a method that, to some extent, overcomes this loss. This display, which may be viewed as an alternative to the histogram, is most useful in preliminary analysis. In particular, it provides a useful first step in constructing a frequency distribution and histogram. The histogram remains the best display to use in formal presentations.

Suppose that we wish to organise the data shown in **Table 4.6** into a more usable form. These data represent a sample of 20 house prices in the suburbs of South and West Auckland, New Zealand. A *stem-and-leaf display* for these data is shown in **Table 4.7**.

TABLE 4.6 Sample of 20 house prices in the suburbs of South and West Auckland (\$'00 000)

6.8	8.7	6.2	6.8	5.7	6.2	5.9	7.4	7.4	5.9
6.6	5.6	5.1	7.3	8.2	6.6	5.4	6.3	5.3	7.1

The first step in developing the display is to decide how to split each observation (house price) into two parts: a stem and a leaf. In this example, we have defined the stem as the digits to the left of the decimal point and the leaf as the digit to the right of the decimal point. The first two prices in **Table 4.6** are, therefore, split into stems and leaves as follows:

House price	Stem	Leaf
6.8	6	8
8.7	8	7

The remaining observations are split into their stem and leaf components in a similar manner.

Having determined what constitutes the stem and the leaf of an observation, we next list the stems in a column from smallest to largest, as shown in **Table 4.7**. Once this has been done, we consider each observation in turn and place its leaf in the same row as its stem, to the right of the vertical line. The resulting stem-and-leaf display, as in **Table 4.7**, presents the original 20 observations in a more organised fashion. The first line in the table, describing stem 5, has seven leaves: 7, 9, 9, 6, 1, 4 and 3. The seven observations represented in the first row are therefore 5.7, 5.9, 5.9, 5.6, 5.1, 5.4 and 5.3. Similarly, seven observations are represented in the second row as well. It seems that most house prices are in the lower end of the values.

TABLE 4.7 Stem-and-leaf display for 20 house prices in South and West Auckland suburbs in New Zealand (\$'00 000)

Stem	Leaf	Ordered leaf
5	7 9 9 6 1 4 3	1 3 4 6 7 9 9
6	8 2 8 2 6 6 3	2 2 3 6 6 8 8
7	4 4 3 1	1 3 4 4
8	7 2	2 7

Whether to arrange the leaves in each row from smallest to largest (as is done in the last column of **Table 4.7**), or to keep them in order of occurrence is largely a matter of personal

preference. The advantage of having the leaves arranged in order of magnitude is that, for example, you can then determine more easily the number of observations less than 7.3 (that is, house prices less than \$730 000).

From the (ordered) stem-and-leaf display in **Table 4.7** we can quickly determine that the house prices range from \$510 000 to \$870 000, that most prices fall between \$510 000 and \$680 000, and that the shape of the distribution is not symmetrical. A stem-and-leaf display is similar to a histogram turned on its side, but the display holds the advantage of retaining the original observations. Moreover, because the stems are listed in order of size, the middle observation(s) can be determined fairly easily. In this example, the two middle prices are \$630 000 and \$660 000; splitting the difference, we can assert that half the prices are below \$645 000 and half are above it. On the other hand, a histogram can readily accommodate a large number of observations, can display relative frequencies and can be adapted more easily to changes in the classes used.

Besides serving an end in itself, the advantages of a stem-and-leaf display are that (1) it shows the shape of the distribution; (2) it displays the specific data values; and (3) it makes it easier to identify any strong outliers. A major disadvantage of a stem-and-leaf display is that it is impractical to construct for large data sets.

Excel does not produce stem-and-leaf displays.

4.1g Shapes of histograms

The shape of a histogram would indicate how the observations are spread. We often describe the shapes of histograms on the basis of the following characteristics.

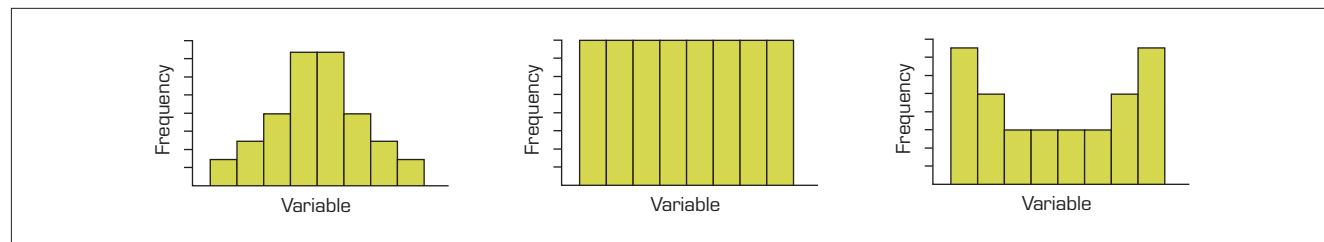
Symmetry

A histogram is said to be symmetric if, when we draw a line down the centre of the histogram, the two sides have identical shape and size. **Figure 4.4** depicts three **symmetric histograms**.

symmetric histograms

Histograms in which, if lines were drawn down the middle, the two halves would be mirror images.

FIGURE 4.4 Three symmetric histograms



Skewness

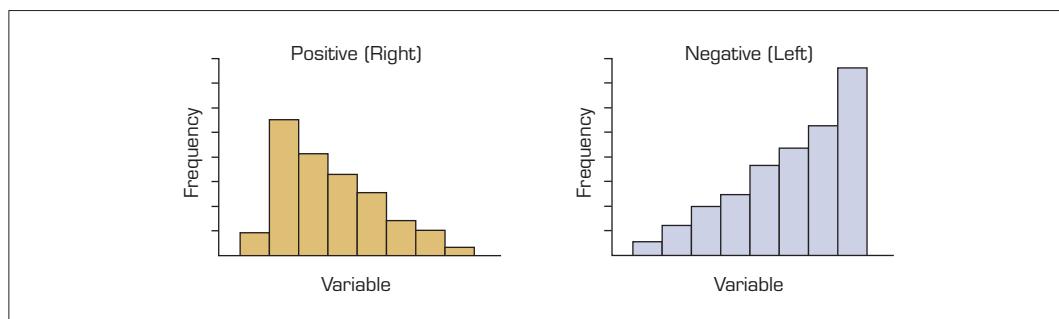
A skewed histogram is one that features a long tail extending either to the right or to the left. The former is called a **positively (right) skewed histogram**, and the latter is called a **negatively (left) skewed histogram**. **Figure 4.5** depicts examples of both. Incomes of individuals working for large companies are usually positively skewed, since there is a large number of relatively low-paid workers and a small number of well-paid executives. The time taken by students to write exams is frequently negatively skewed; a few students turn in their papers early, while most prefer to re-read their papers until the formal end of the test period.

positively skewed histogram

A histogram with a long tail to the right.

negatively skewed histogram

A histogram with a long tail to the left.

FIGURE 4.5 Skewed histograms

Number of modal classes

modal class

The class with the largest number of observations.

unimodal histogram

A histogram with only one mode.

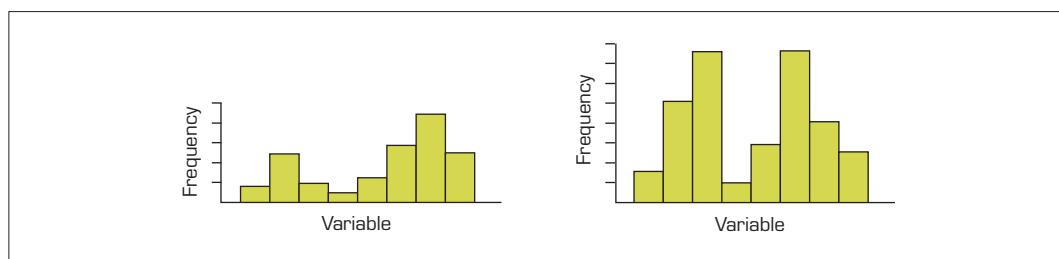
bimodal histogram

A histogram with two modes.

multimodal histogram

A histogram with two or more peaks.

As we will discuss in Chapter 5, a mode is the observation that occurs with the greatest frequency. A **modal class** is the one with the largest number of observations. Thus, a **unimodal histogram** is one with a single peak. The histograms in **Figure 4.5** are unimodal. A **bimodal histogram** is one with two peaks, which are not necessarily equal in height. The final marks in the authors' manual statistics subject appear to be bimodal. **Figure 4.6** depicts such histograms. We leave it to you to interpret the implications of this information. A **multimodal histogram** is one with two or more peaks.

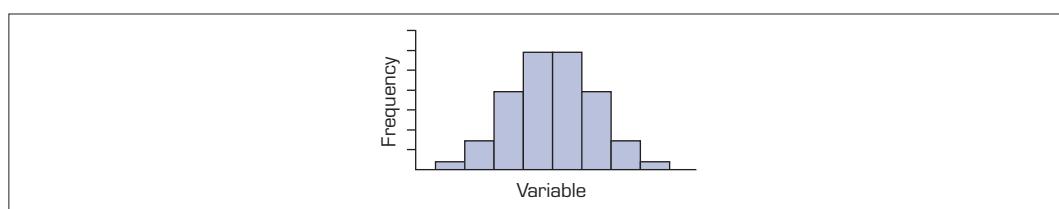
FIGURE 4.6 Bimodal histograms

Bell-shaped distribution

bell-shaped

Symmetric in the shape of a bell (or mound-shaped).

A special type of symmetric unimodal histogram is one that is **bell-shaped**. You will discover later in this book the importance of the normal distribution, which appears bell-shaped when drawn. **Figure 4.7** illustrates a bell-shaped histogram. Many statistical techniques require that the population be bell-shaped, and we often draw the histogram to check that this requirement is satisfied.

FIGURE 4.7 Bell-shaped histogram

REAL-LIFE APPLICATIONS

Stock and bond valuation

A basic understanding of how financial assets, such as stocks and bonds, are valued is critical to good financial management. Understanding the basics of valuation is necessary for capital budgeting and capital structure decisions. Moreover, understanding the basics of valuing investments such as stocks and bonds is at the heart of the huge and growing discipline known as investment management.

A financial manager must be familiar with the main characteristics of the capital markets in which

long-term financial assets such as stocks and bonds are traded. A well-functioning capital market provides managers with useful information concerning the appropriate prices and rates of return that are required for a variety of financial securities with differing levels of risk. Statistical methods can be used to analyse capital markets and summarise their characteristics, such as the shape of the distribution of stock or bond returns.

REAL-LIFE APPLICATIONS

Return on investment

The return on an investment is calculated by dividing the gain (or loss) by the value of the investment. For example, a \$100 investment that is worth \$106 after one year has a 6% rate of return. A \$100 investment that loses \$20 has a -20% rate of return. For many investments, including individual shares and share portfolios (combinations of various shares), the rate of return is a *variable*. That is, the investor does not know in advance what the rate of return will be.

Investors are torn between two goals. The first is to maximise the rate of return on investment; the second goal is to reduce risk. If we draw a histogram of the returns for a certain investment, the location of the centre of the histogram gives us some information about the return one might expect from that investment. The spread or variation of the histogram provides us with guidance about the risk. If there is little variation, an investor can be quite confident in



Source: Shutterstock.com/SergeyP

predicting his or her rate of return. If there is a great deal of variation, the return becomes much less predictable and thus the investment riskier. Minimising the risk becomes an important goal for investors and financial analysts.

EXAMPLE 4.2

L01

Comparing returns on two investments

XM04-02 Suppose that you are facing a decision on where to invest that small fortune that remains after you have deducted the anticipated expenses for the next year from the earnings from your summer job. A friend has suggested two types of investment, and to help make the decision you acquire some annual rates of return for each type. You would like to know what you can expect by way of the return on your investment, as well as other types of information, such as whether the rates are spread out over a wide range (making the investment risky) or are grouped tightly together (indicating relatively low risk). Do the data indicate that it is possible that you can do extremely well with little likelihood of a large loss? Is it likely that you could lose money (negative rate of return)?



The returns for the two types of investments are listed below. Draw histograms for each set of returns and report on your findings. Which investment would you choose and why?

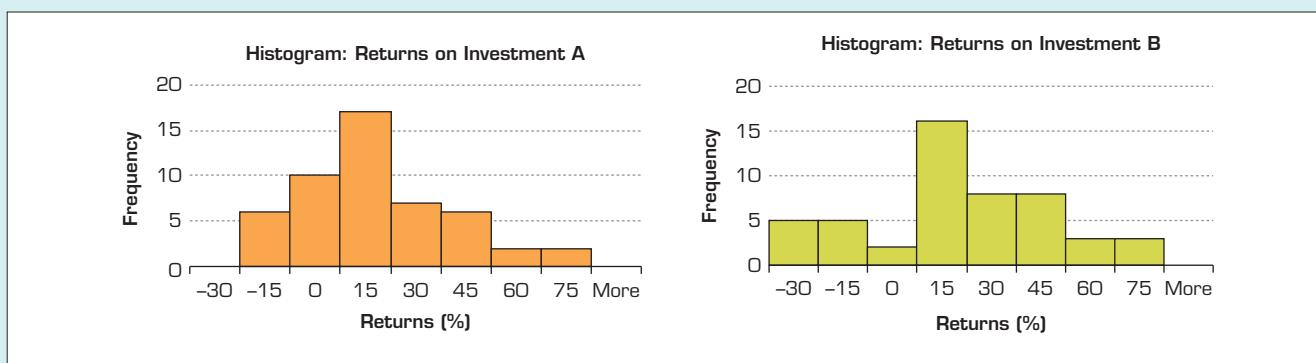
Returns on investment A (%)				Returns on investment B (%)			
30.00	6.93	13.77	-8.55	30.33	-34.75	30.31	24.30
-2.13	-13.24	22.42	-5.29	-30.37	54.19	6.06	-10.01
4.30	-18.95	34.40	-7.04	-5.61	44.00	14.73	35.2
25.00	9.43	49.87	-12.11	29.00	-20.23	36.13	40.70
12.89	1.21	22.92	12.89	-26.01	4.16	1.53	22.18
-20.24	31.76	20.95	63.00	0.46	10.03	17.61	3.24
1.20	11.07	43.71	-19.27	2.07	10.51	1.20	25.1
-2.59	8.47	-12.83	-9.22	29.44	39.04	9.94	-24.24
33.00	36.08	0.52	-17.00	11.00	24.76	-33.39	-38.47
14.26	-21.95	61.00	17.30	-25.93	15.28	58.67	13.44
-15.83	10.33	-11.96	52.00	8.29	34.21	0.2	68.00
0.63	12.68	1.94		61.00	52.00	5.23	
38.00	13.09	28.45		-20.44	-32.17	66.00	

Solution

We draw the histograms of the returns on the two investments by using Excel.

Using the computer

Excel output for Example 4.2



Interpreting the results

As the two data sets have an equal number of observations, we could compare the two histograms based on absolute frequencies. (If the numbers of observations are not equal, then histograms based on relative frequencies should be compared.) By comparing the two histograms, we can extract the following information:

- 1 The centre of the histogram of the returns of investment A is slightly lower than that for investment B.
- 2 The spread of returns for investment A is considerably less than that for investment B.
- 3 Both histograms are slightly positively skewed.

These findings suggest that investment A is superior. Although the returns on A are slightly less than those for B, the wider spread for B makes it unappealing to most investors. Both investments allow for the possibility of a relatively large return.

The interpretation of the histograms is somewhat subjective. Other viewers may not concur with our conclusion. In such cases, numerical techniques provide the detail and precision lacking in most graphs. In Chapter 5, we introduce a number of summary statistics to complement the graphical description of the data.

EXAMPLE 4.3

LO1

Business statistics marks: Manual versus computer calculations

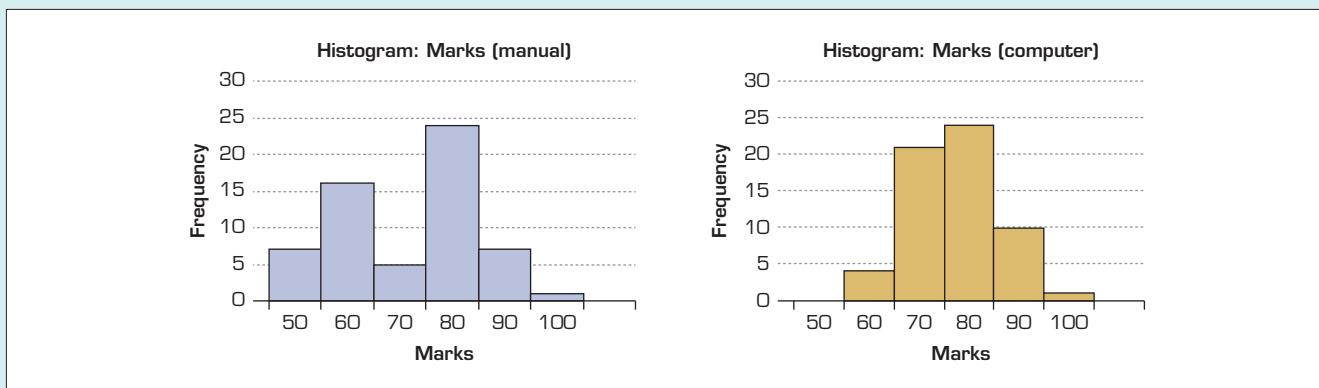
XM04-03 The final marks in a statistics course that emphasised mathematical proofs, derivations and manual (hand) calculations, both during the class and on exams, are listed below. The marks obtained by students in the same course after the emphasis was changed to applications with most of the calculations performed using a computer, are also listed below. Draw histograms for both groups and interpret the results.

Marks (manual course)				Marks (computer course)			
77	67	53	54	65	81	72	59
74	82	75	44	71	53	85	66
75	55	76	54	66	70	72	71
75	73	59	60	79	76	77	68
67	92	82	50	65	73	64	72
72	75	82	52	82	73	77	75
81	75	70	47	80	85	89	74
76	52	71	46	86	83	87	77
79	72	75	50	67	80	78	69
73	78	74	51	64	67	79	60
59	83	53	44	62	78	59	92
83	81	49	52	74	68	63	69
77	73	56	53	67	67	84	69
74	72	61	56	72	62	74	73
78	71	61	53	68	83	74	65

Solution**Using the computer**

The Excel commands are the same as those used in Example 4.1.

Excel output for Example 4.3



Note that Excel puts the upper limit of the class interval at the midpoint of the class. For example, 80 in the histogram refers to the class interval $70 < x \leq 80$.

Interpreting the results

As the number of observations is the same for each group of marks, the absolute frequency histograms can be used for comparison. The histogram of the marks in the 'manual' statistics course is bimodal. The larger modal class consists of the marks in the 70s. The smaller modal class contains the marks in the 50s. There appear to be



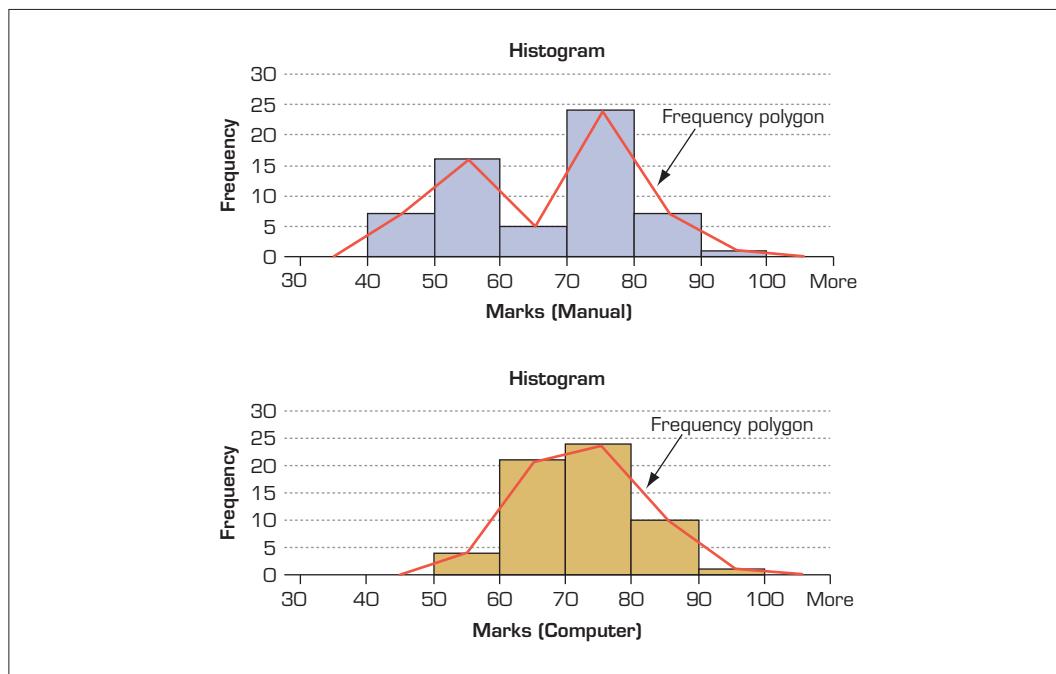
few marks in the 60s. This histogram suggests that there are two groups of students. Because of the emphasis on mathematical manipulation in the course, one may conclude that those who performed poorly in the course are weaker mathematically than those who performed well.

Now we compare the shape of the histograms for the ‘manual’ and ‘computer’ statistics courses. The first histogram, for the manual statistics course, is bimodal and appears to be spread out. The second histogram, for the computer statistics course, is unimodal and bell-shaped, and it appears that its spread is less than that of the first histogram. One possible interpretation is that this type of course allows students who are not particularly mathematical to learn statistics and to perform as well as the mathematically inclined students.

4.1h Frequency polygons

Another common way of presenting a frequency distribution graphically is to use a frequency polygon (see [Figure 4.8](#)). A frequency polygon is obtained by plotting the frequency of each class above the midpoint of that class and then joining the points with straight lines. The polygon is usually closed by considering one additional class (with zero frequency) at each end of the distribution and then extending a straight line to the midpoint of each of these classes. Frequency polygons are useful for obtaining a general idea of the shape of the distribution. As with histograms, we can plot relative frequencies rather than frequencies, thereby obtaining a relative frequency polygon.

FIGURE 4.8 Frequency polygon of business exam marks: manual and computer calculation



4.1i Ogives

Given a set of observations that have been grouped into classes, we have seen that the relative frequency distribution identifies the proportion of observations falling into each class. In some instances, however, our needs are better served by a **cumulative relative frequency** distribution. The cumulative relative frequency of a particular class is the proportion of observations that are less than or equal to the upper limit of that class. That is, to obtain the cumulative frequency of a particular class, we add the frequency of that class to the frequencies of all the previous classes. [Table 4.8](#) displays the cumulative relative frequency distribution of the electricity bills in Example 4.1.

cumulative relative frequency

Percentage of observations less than or equal to the upper limit of a class.

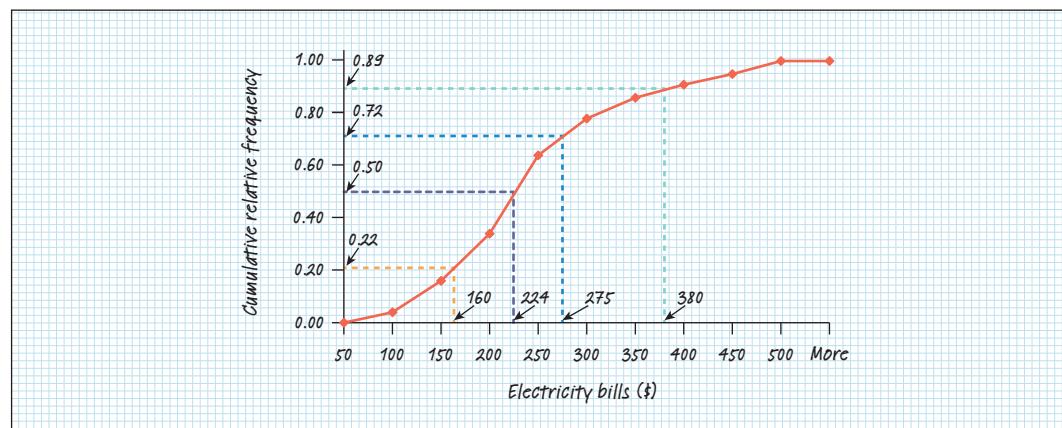
TABLE 4.8 Cumulative relative frequencies for Example 4.1

Classes	Frequency	Cumulative frequency	Relative frequency	Cumulative relative frequency
50 up to 100	8	8	8/200 = 0.04	8/200 = 0.04
100 up to 150	24	8 + 24 = 32	24/200 = 0.12	32/200 = 0.16
150 up to 200	36	8 + 24 + 36 = 68	36/200 = 0.18	68/200 = 0.34
200 up to 250	60	128	60/200 = 0.30	128/200 = 0.64
250 up to 300	28	156	28/200 = 0.14	156/200 = 0.78
300 up to 350	16	172	16/200 = 0.08	172/200 = 0.86
350 up to 400	10	182	10/200 = 0.05	182/200 = 0.91
400 up to 450	8	190	8/200 = 0.04	190/200 = 0.95
450 up to 500	10	200	10/200 = 0.05	200/200 = 1.00

From the cumulative relative frequency distribution we can state, for example, that 34% of the bills were less than \$200 and 78% were less than \$300. Another way of presenting this information is to use an **ogive**, which is a graphical representation of the cumulative relative frequency distribution, as drawn manually in **Figure 4.9**.

ogive

Line graph of cumulative relative frequency.

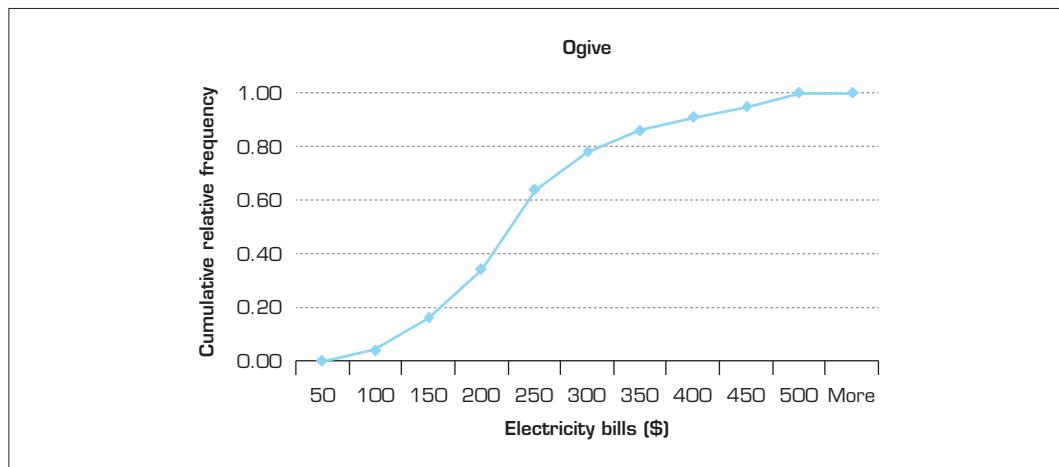
FIGURE 4.9 Ogive for electricity bills

The cumulative relative frequency of each class is plotted above the *upper limit* of the corresponding class, and the points representing the cumulative relative frequencies are then joined by straight lines. The ogive is closed at the lower end by extending a straight line to the lower limit of the first class. Once an ogive like the one shown in **Figure 4.9** has been drawn, the approximate proportion of observations that are less than any given value on the horizontal axis can be read from the graph. Thus, for example, we can estimate manually from **Figure 4.9** that the proportion of electricity bills that are less than \$380 is approximately 89% and the proportion greater than \$380 is 11%. The proportion of bills less than \$275 is about 72%. We can also estimate that 22% of the bills are less than \$160, and 50% of the bills are less than \$224.

Excel can be used to produce a cumulative frequency distribution and ogive as follows.

Using the computer

Excel output of the ogive for Example 4.1



COMMANDS

Proceed through the first seven steps to create a histogram (see page 89).

- 8 Output will normally be stored in a new sheet of the workbook. To store the output on the same sheet (Sheet 1), in the **Output Options** select **Output Range** and type the starting cell reference in the space provided. Also click the checkbox for **Chart Output** and **Cumulative Percentage**, then click **OK**. This will produce a cumulative frequency distribution and ogive as shown above.
- 9 Remove the 'More' category row from the frequency table.
- 10 Right click on any of the rectangles on the histogram and click **Delete**.
- 11 You can change the scale by right-clicking on an axis within the graph, selecting **Format axis...** A menu will appear on the right-hand side of the screen and changes can be made to the values under **AXIS OPTIONS**. Make any changes (e.g. make the **Maximum** value of Y equal to 1.0) and then click the cross at the top right hand corner of the menu to close it.

REAL-LIFE APPLICATIONS

Credit scorecards

Credit scorecards are used by banks and financial institutions to determine whether applicants will receive loans. The scorecard is the product of a statistical technique that converts questions about income, residence and other variables into a score. The higher the score, the higher the probability that the applicant will repay. The scorecard is a formula

produced by a statistical technique called logistic regression. For example, a scorecard may score age categories in the following way:

Less than 25	20 points
25 to 39	24 points
40 to 55	30 points
Over 55	38 points

► Other variables would be scored similarly. The sum for all variables would be the applicant's score. A cut-off score would be used to predict those who will repay and those who will default. Because no scorecard is perfect, it is possible to make two types of error: granting credit to those who will default and not lending money to those who would have repaid. This application has some relevance to Exercises 4.27 and 4.28.



Source: © iStockphoto/AIMSTOCK

We complete this section with a review of when to use a histogram, an ogive or a stem-and-leaf display. Note that we use the term *objective* to identify the type of information produced by the statistical technique.

IN SUMMARY

Factors that identify when to use a histogram or ogive

- 1** *Objective*: to describe a set of data
- 2** *Data type*: numerical – cross-sectional data

Factors that identify when to use a stem-and-leaf display

- 1** *Objective*: to describe a small set of data
- 2** *Data type*: numerical (and cross-sectional)

EXERCISES

Learning the techniques

- 4.1** How many classes should a histogram contain if the number of observations is 125?
- 4.2** Determine the number of classes of a histogram for 1500 observations.
- 4.3** A data set consists of 300 observations that range between 147 and 239.
 - a** What is an appropriate number of classes to have in the histogram?
 - b** What class intervals would you suggest?
- 4.4** A statistics practitioner would like to draw a histogram of 40 observations that range from 5.2 to 6.1.
 - a** What is an appropriate number of classes to have in the histogram?
 - b** Define the upper limits of the classes.

- 4.5 XR04-05** The numbers of weekly sales calls by a sample of 30 telemarketers are listed below.

14	8	6	12	21	4	9	3	25	17
9	5	8	18	16	3	17	19	10	15
5	20	17	14	19	7	10	15	10	8

- a** Construct a histogram.
- b** Construct an ogive.
- c** Describe the shape of the histogram.

- 4.6 XR04-06** The mid-semester exam marks (marked out of 30) in a finance course are as follows:

17	11	6	21	25	21	8	17	22	19	18	19	14
6	21	20	25	19	9	12	16	16	10	29	24	

- a** Develop a frequency distribution table for the data and draw a frequency histogram. Use five class intervals, with the lower boundary of the first class being five marks.

- b** Draw a relative frequency histogram for the data.
- c** What is the relationship between the areas under the histogram you have drawn and the relative frequencies of observations?
- 4.7 XR04-07** An Uber driver kept track of the number of calls he received over a 28-day period. Draw a frequency histogram of these data and describe it.
- | | | | | | | |
|----|----|---|----|----|----|----|
| 10 | 10 | 7 | 7 | 3 | 8 | 11 |
| 8 | 10 | 7 | 7 | 7 | 5 | 4 |
| 9 | 7 | 8 | 4 | 17 | 13 | 9 |
| 7 | 12 | 8 | 10 | 4 | 7 | 5 |
- 4.8 XR04-08** The times taken (in minutes) by a class of 30 students to complete a statistics exam are as follows:
- | | | | | | | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 61 | 86 | 61 | 58 | 70 | 75 | 66 | 77 | 66 | 64 | 73 | 91 | 65 | 59 | 86 |
| 82 | 48 | 67 | 55 | 77 | 80 | 58 | 94 | 78 | 62 | 79 | 83 | 54 | 52 | 45 |
- a** Draw a stem-and-leaf display for the data.
- b** Develop a frequency distribution table for the data using six class intervals and draw a frequency histogram.
- c** Draw a relative frequency histogram for the data.
- d** Briefly describe what the histogram and the stem-and-leaf display tell you about the data.
- e** Draw a cumulative relative frequency histogram and the ogive for the data.
- f** What proportion of the students took (i) less than or equal to 70 minutes to complete the exam? (ii) more than 70 minutes to complete the exam?
- ### Applying the techniques
- 4.9 XR04-09 Self-correcting exercise.** A large investment firm in Sydney wants to review the distribution of the ages of its stockbrokers. The firm feels that this information would be useful in developing plans relating to recruitment and retirement options. The ages of a sample of 25 brokers are as follows:
- | | | | | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 50 | 64 | 32 | 55 | 41 | 44 | 24 | 46 | 58 | 47 | 36 | 52 | 54 |
| 44 | 66 | 47 | 59 | 51 | 61 | 57 | 49 | 28 | 42 | 38 | 45 | |
- a** Draw a stem-and-leaf display for the ages.
- b** Develop a frequency distribution table for the data, using five class intervals and the value 20 as the lower limit of the first class, then draw a frequency histogram.
- c** Draw a relative frequency histogram for the data, using five class intervals and the value 20 as the lower limit of the first class.
- d** Draw a frequency polygon.
- e** Draw an ogive for the data.
- f** What proportion of the total area under the histogram from part (c) falls between 20 and 40?
- 4.10 XR04-10** The number of cars exceeding the 40 km/h speed limit in a school zone at Sunnybank in the past 25 (school) weeks are shown below.
- | | | | | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 41 | 48 | 38 | 27 | 29 | 37 | 33 | 31 | 40 | 50 | 33 | 28 | 46 |
| 31 | 38 | 24 | 56 | 43 | 35 | 37 | 27 | 29 | 44 | 34 | 28 | |
- Draw the following graphs:
- a** A histogram with five classes
- b** A histogram with 10 classes
- c** A stem-and-leaf display
- d** A frequency polygon for parts (a) and (b)
- 4.11 XR04-11** The amount of time (in seconds) needed for assembly-line workers to complete a weld at a car assembly plant in Adelaide was recorded for 40 workers.
- | | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|
| 69 | 60 | 75 | 74 | 68 | 66 | 73 | 76 | 63 | 67 |
| 69 | 73 | 65 | 61 | 73 | 72 | 72 | 65 | 69 | 70 |
| 64 | 61 | 74 | 76 | 72 | 74 | 65 | 63 | 69 | 73 |
| 75 | 70 | 60 | 62 | 68 | 74 | 71 | 73 | 68 | 67 |
- Draw the following graphs:
- a** A relative frequency histogram with six classes
- b** A relative frequency histogram with 12 classes
- c** An ogive for parts (a) and (b)
- 4.12** The percentage distribution of weekly Australian household incomes for 2016 is shown in **Table 4.5** on page 93.
- a** Use this table to develop a cumulative relative frequency distribution.
- b** Graph the cumulative relative frequency distribution and the ogive.
- c** What percentage of the annual incomes were less than \$62 400 in 2016?
- d** Use the graph in part (b) to estimate the weekly income below which 50% of the Australian households fell.
- 4.13 XR04-13** A survey of 60 individuals leaving a Westfield shopping centre asked how many shops they entered during this visit. The responses are listed below.
- | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 5 | 5 | 1 | 4 | 3 | 4 | 4 | 1 | 0 | 4 | 6 |
| 4 | 3 | 1 | 6 | 9 | 1 | 3 | 2 | 4 | 3 | 3 | 9 |
| 2 | 4 | 3 | 6 | 2 | 2 | 8 | 7 | 6 | 4 | 5 | 1 |
| 6 | 2 | 2 | 5 | 3 | 8 | 0 | 2 | 5 | 4 | 4 | 4 |
| 3 | 4 | 1 | 1 | 4 | 8 | 5 | 2 | 3 | 1 | 1 | 7 |
- a** Draw a histogram to summarise these data and describe the shape of the histogram.
- b** Draw an ogive.
- c** Describe your findings.

Computer applications

4.14 XR04-14 Users of previous editions of this book could download an Excel add-in called Data Analysis Plus from our website. We recorded the number of daily downloads during a 78-day period.

- a Draw a histogram.
- b Describe its shape.

4.15 XR04-15 To help determine the need for more golf courses, a survey was undertaken. A sample of 75 self-declared golfers was asked how many rounds of golf they played last year and their responses were recorded.

- a Draw a histogram.
- b Draw an ogive.
- c Describe what you have learnt.

4.16 XR04-16 The annual income in thousands of dollars for a sample of 200 junior trainees in a fast-food chain was recorded.

- a Draw a histogram.
- b Use the graph drawn in part (a) to describe the shape of the distribution of annual incomes of junior trainees. Discuss what you have learnt.
- c Draw an ogive.
- d Estimate the proportion of junior trainees who earn:
 - i less than \$20 000
 - ii more than \$35 000
 - iii between \$25 000 and \$40 000.

4.17 XR04-17 The final marks on a statistics exam are recorded.

- a Draw a histogram.
- b Briefly describe what the histogram tells you about the shape of the distribution.
- c Draw a cumulative relative frequency distribution, and draw the ogive for the marks.
- d Estimate the proportion of marks that are less than 70.
- e Estimate the proportion of marks that are less than 75.

4.18 XR04-18 A real estate agency in a suburb in Adelaide wishes to investigate the distribution of prices of houses sold during the past year. The prices are recorded in thousands of dollars (\$'000).

- a What do the histogram and stem-and-leaf display tell you about the prices?
- b Draw an ogive for the house prices.
- c Estimate the proportion of prices that are less than \$400 000.
- d Estimate the proportion of prices that are less than \$550 000.

4.19 XR04-19 The president of a local consumer association is concerned about reports that similar generic medicines are being sold at widely differing prices at local chemists. A survey of 100 chemists recorded the selling price of one popular generic medicine.

- a Draw a histogram.
- b What does the histogram tell you about the price?

4.20 XR04-20 The number of customers entering a bank between 10 a.m. and 3 p.m. for each of the past 100 working days was recorded and stored in the following way:

- Column 1: Number of arrivals between 10 a.m. and 11 a.m.
- Column 2: Number of arrivals between 11 a.m. and 12 p.m.
- Column 3: Number of arrivals between 12 p.m. and 1 p.m.
- Column 4: Number of arrivals between 1 p.m. and 2 p.m.
- Column 5: Number of arrivals between 2 p.m. and 3 p.m.
- a Use appropriate graphical techniques to describe each set (column) of data.
- b Describe the shape of the number of arrivals for each time period.
- c Discuss similarities and differences between time periods.
- d What are the implications of your findings?

4.21 XR04-21 The lengths of time (in minutes) taken to serve each of 420 customers at a local restaurant were recorded.

- a How many class intervals should a histogram of these data contain?
- b Draw a histogram using the number of classes specified in part (a).
- c Is the histogram symmetric or skewed?
- d How many modes are there?
- e Is the histogram bell-shaped?

4.22 XR04-22 The number of copies made by an office copier was recorded for each of the past 75 days. Graph the data using a suitable technique. Describe what the graph tells you.

4.23 XR04-23 The volume of water used by each of a sample of 350 households was measured (in litres) and recorded. Use a suitable graphical statistical method to summarise the data. What does the graph tell you?

- 4.24 XR04-24** The number of books shipped out daily by Amazon.com was recorded for 100 days. Draw a histogram and describe your findings.
- 4.25 XR04-25** The lengths (in centimetres) of 150 newborn babies were recorded. Use whichever graphical technique you judge suitable to describe these data. What can you observe from the graph?
- 4.26 XR04-26** The weights (in grams) of a sample of 240 tomatoes grown with a new type of fertiliser were recorded. Draw a histogram and describe your findings.
- 4.27 XR04-27** A small bank that had not yet used a credit scorecard wanted to determine whether a scorecard would be advantageous. The bank manager took a random sample of 300 loans that were granted and scored each on a scorecard borrowed from a similar bank. This scorecard is based on the responses supplied by the applicants to questions such as age, marital status and household income. The cut-off is 650, which means that those scoring below this figure are predicted to default and those scoring above are predicted to repay. Two hundred and twenty of the loans were repaid, the rest were not.

The scores of those who repaid and the scores of those who defaulted were recorded.

- a** Use a graphical technique to present the scores of those who repaid.
- b** Use a graphical technique to present the scores of those who defaulted.
- c** What have you learnt about the scorecard?

- 4.28 XR04-28** Refer to Exercise 4.27. The bank manager decided to try another scorecard, this one based not on the responses of the applicants but on credit bureau reports, which list problems such as late payments and previous defaults. The scores using the new scorecard of those who repaid and the scores of those who did not repay were recorded. The cut-off score is 650.
- a** Use a graphical technique to present the scores of those who repaid.
 - b** Use a graphical technique to present the scores of those who defaulted.
 - c** What have you learnt about the scorecard?
 - d** Compare the results of this exercise with those of Exercise 4.27. Which scorecard appears to be better?

4.2 Describing time-series data

As well as classifying data by type, we can classify them according to whether the observations are measured at the same time or whether they represent measurements at successive points in time. The former are called **cross-sectional data**, the latter are **time-series data**.

The techniques described in Section 4.1 are applied to cross-sectional data. All the data for Examples 4.1–4.3 were probably determined at the same point in time. To give another example, consider a real estate consultant who feels that the selling price of a house is a function of its size, age and lot size. To estimate the specific form of the function, she samples 100 recently sold homes and records the price, size, age and lot size for each home. These data are cross-sectional: they are all observations taken at the same point in time. The real estate consultant is also working on a separate project to forecast the monthly housing starts in southern Queensland during the next year. To do so, she collects the monthly housing starts in this region for each of the past five years. These 60 values (housing starts) represent time-series data because they are observations taken over time. Note that the original data may be numerical or nominal. All the illustrations above deal with numerical data. A time series can also list the frequencies and relative frequencies of a nominal variable over a number of time periods. For example, a brand-preference survey asks consumers to identify their favourite brand. These data are nominal. If we repeat the survey once a month for several years, the proportion of consumers who prefer a certain company's product each month would constitute a time series.

cross-sectional data

Data measured across a population (or a sample) at one point in time.

time-series data

Data measured on the same variable at different points of time.

line chart

A chart showing the movement of a variable over time.

time-series chart

Line chart in which the categories are points in time.

4.2a Line charts

Time-series data are often graphically depicted on a **line chart**, which is a plot of the variable over time. It is created by plotting the value of the variable on the vertical axis and the time periods on the horizontal axis. Since a line chart is normally used when the categories are points in time, it is known alternatively as a **time-series chart**.

EXAMPLE 4.4

LO2

Queensland exports and imports, 1989–2018

XM04-04 The total values of Queensland's exports and imports (in millions of dollars) from 1989 to 2018 are presented below. Use a graphical technique to show the trend in the total exports and imports.

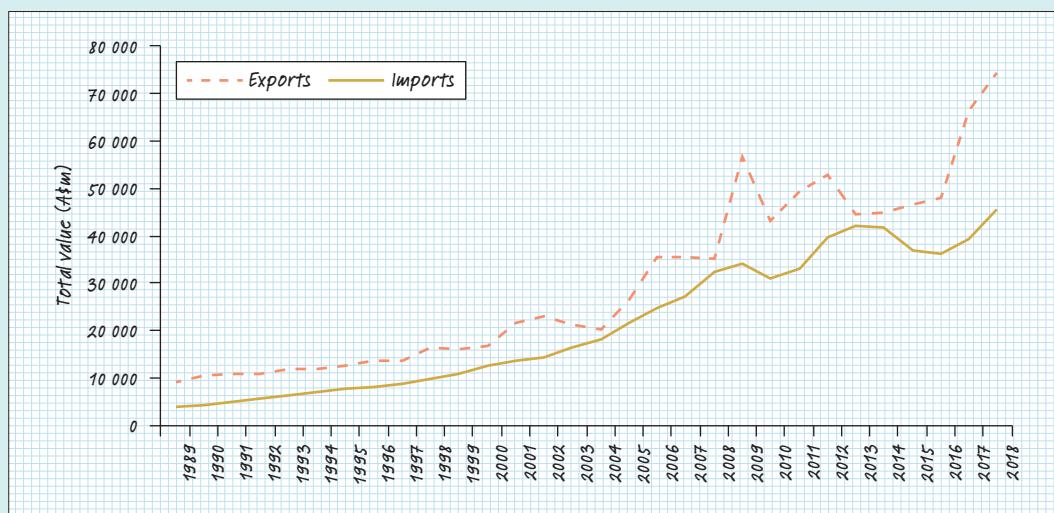
Year	Exports	Imports	Year	Exports	Imports
1989	9033.8	3787.8	2004	20128.5	18078.5
1990	10465.7	4258.1	2005	26368.5	21685.0
1991	10727.5	4903.5	2006	35384.7	24862.6
1992	10865.1	5626.7	2007	35439.2	27197.7
1993	11798.2	6334.2	2008	35317.8	32360.0
1994	11984.2	6869.2	2009	56552.9	33972.4
1995	12509.6	7771.0	2010	43265.6	31035.2
1996	13624.6	8052.2	2011	49353.4	33024.2
1997	13566.6	8636.8	2012	52867.7	39787.1
1998	16297.1	9751.0	2013	44433.0	42173.6
1999	15907.4	10809.9	2014	44811.7	41860.1
2000	16644.1	12749.0	2015	46488.6	36803.0
2001	21471.9	13781.8	2016	47866.5	36161.8
2002	23151.1	14217.6	2017	66485.9	39193.9
2003	21381.0	16357.9	2018	74262.0	45688.2

Source: Australian Bureau of Statistics, Foreign Trade; <http://www.qgso.qld.gov.au/subjects/economy/trade/tables/os-exports-imports-goods-qld-aus/index.php>.

Solution

A bar chart is useful if the focus is on the comparison of the value of exports and imports in different years. But if the objective of the graph is to focus on the trend in the value of exports and imports over the years – rather than to compare the total amounts in different years – then a line chart is appropriate, as drawn manually in **Figure 4.10**.

FIGURE 4.10 Line chart of Queensland's overseas exports and imports, 1989–2018





Excel can be used to create a line chart as follows.

Using the computer

COMMANDS

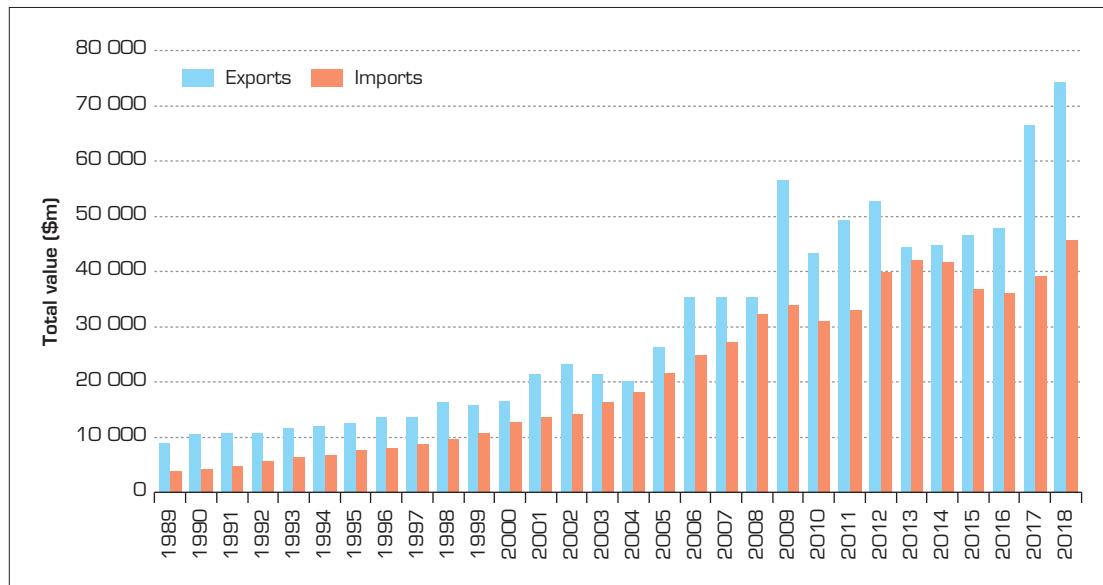
- 1 Type the data in three columns or open the data file. (**XR04-04**)
- 2 Delete the first column (X variable) title 'Year' in cell A2. Highlight the year column and the two Y-variable columns with title and click **INSERT**. (**A2:C32**).
- 3 In the **Charts** submenu, select the **Line chart** icon. Then click on one of the **2-D line** charts.
- 4 Click on **Chart title** on the graph and type the relevant chart title. Click on **DESIGN** or **FORMAT** under **CHART TOOLS** to make whatever changes you wish to make (e.g. insert axis title, change the solid line to dash type etc.).

To draw a line chart for only one variable, highlight the year column and the column of data for that variable. To draw more than two line charts (for three or more variables), highlight all columns of data you wish to graph. Note that the year column should not have a title, whereas the other columns should have titles.

4.2b Which chart is better?

We can use either a line chart or a bar chart to present time-series data. In **Figure 4.10**, we displayed the total value of Queensland exports and imports for 1989–2018 using a line chart. **Figure 4.11** displays the same data using a bar chart. Which chart is better? The answer depends on two factors: the objective of the graph and the number of periods. If the objective of the graph is to focus on the *trend* in exports and imports over the years, a line chart is superior. If the goal is to compare the values of exports and imports in different years, a bar chart is recommended.

FIGURE 4.11 Bar chart of Queensland's exports and imports, 1989–2018



If there is a large number of periods (e.g. the total value of Queensland's exports and imports for 1989–2018), a line chart looks less cluttered and makes a clearer impression.

Pie charts, bar charts and line charts are used extensively in reports compiled by businesses, governments and the media. Variations on these and other pictorial representations of data abound, and their possibilities are limited only by their creators' imaginations. The objective of all such charts is to present a summary of the data clearly and in a form that allows the reader to grasp the relevant comparisons or trends quickly.

REAL-LIFE APPLICATIONS

Measuring inflation: Consumer Price Index

Inflation is the increase in the prices for goods and services. In most countries, inflation is measured using the Consumer Price Index (CPI). The CPI works with a basket of some 300 goods and services in the United States (and a similar number in other countries), including such diverse items as food, housing, clothing, transportation, health and recreation. The basket is defined for the 'typical' or 'average' middle-income family, and the set of items and their weights are revised periodically (every 5 years in Australia, every 10 years in the United States and every 7 years in Canada). Prices for each item in this basket are calculated on a monthly basis and the CPI is computed from these prices. Here is how it works. We start by setting a period of time as the base. For example, suppose the base month is March 2012 and the basket of goods and services costs \$1000 during this period. Thus, the base is \$1000, and the CPI is set at 100. Suppose that in the next month (April 2012) the price increases to \$1010. The CPI for April 2012 is calculated in the following way:

$$\text{CPI(April 2012)} = (1010/1000) \times 100 = 101$$

If the price increases to \$1050 in the next month, the CPI is

$$\text{CPI(May 2012)} = (1050/1000) \times 100 = 105$$

The CPI, although never really being intended to serve as the official measure of inflation, has come to be interpreted in this way by the general public. Australian seniors pensions, unemployment benefits, child support payments, and other Centrelink support payments are automatically linked to the CPI and automatically indexed to the level of inflation. Despite its flaws, the CPI is used in numerous applications. One application involves adjusting prices by removing the effect of inflation, making it possible to track the 'real' changes in a time series of prices.

In Example 4.4, the figures shown are the actual prices measured in what are called current dollars. To remove the effect of inflation, we divide the annual prices by the CPI for that year and multiply by 100. These prices are then measured in constant 2012 dollars. This makes it easier to see what has happened to the prices of the goods and services of interest.

The Consumer Price Index will be discussed in more detail in Chapter 18.

IN SUMMARY

Factors that identify when to use a line chart

- 1 *Objective:* to describe a set of data
- 2 *Data type:* numerical time-series

EXERCISES

Learning the techniques

4.29 XR04-29 Over the past few years, the Australian Government has been quite determined to increase exports and reduce imports. To develop an understanding of the problem, a statistics practitioner determined the annual Australian exports and imports of merchandise trade. These data are listed in the following table.

Year	Exports (\$m)	Imports (\$m)	Year	Exports (\$m)	Imports (\$m)	Year	Exports (\$m)	Imports (\$m)
1989	44006.8	47039.5	1999	85996.6	97611.2	2009	230829.0	219482.0
1990	49078.4	51333.4	2000	97286.1	110077.6	2010	200720.1	203977.1
1991	52398.9	48912.2	2001	119539.1	118317.1	2011	245723.9	213804.1
1992	55026.9	50984.0	2002	121108.4	119648.7	2012	264016.9	239729.3
1993	60702.3	59575.4	2003	115479.4	133129.5	2013	246977.6	236499.9
1994	64548.4	64469.7	2004	109049.4	130996.7	2014	272920.6	252333.1
1995	67048.4	74619.0	2005	126823.1	149426.1	2015	254550.7	256968.3
1996	76004.6	77791.7	2006	152492.3	167364.0	2016	243422.7	263263.5
1997	78931.9	78998.1	2007	168098.5	180732.5	2017	290879.7	264008.1
1998	87773.8	90684.3	2008	180856.7	202289.5	2018	314478.5	301159.2

Source: Australian Bureau of Statistics, *Australian Economic Indicators 2019*, cat. no. 1350.0, ABS, Canberra.

Calculate the annual trade deficits (trade deficit = imports – exports) and draw line charts of the annual exports, imports and trade deficit. Briefly describe what the chart tells you.

Applying the techniques

4.30 XR04-30 Self-correcting exercise. The Aboriginal and Torres Strait Islander population is less than 3% of the total Australian population, but Aboriginal and Torres Strait Islander peoples represent a quarter of the prison population. Moreover, the relatively higher rate of Aboriginal and Torres Strait Islander deaths in custody has been a serious issue for successive Australian governments. The number of Aboriginal and Torres Strait Islander deaths and total deaths (in police and prison custody) in Australia during the period 2001–17 are shown in the following table.

Year	Deaths in custody		Australian population	
	Indigenous	Non-Indigenous	Indigenous	Non-Indigenous
2001	14	44	534718	18851743
2002	6	46	547940	19057501
2003	12	32	560973	19266182
2004	6	33	573991	19472012
2005	8	31	587486	19724057
2006	3	25	601450	20026097

2007	8	32	615303	20400818
2008	6	40	629167	20846458
2009	7	36	643049	21222574
2010	14	44	656735	21515734
2011	12	46	669881	21850417
2012	6	36	684017	22236781
2013	9	44	698583	22596776
2014	10	44	713600	22926731
2015	15	46	681386	23303196
2016	19	64	649171	23740513
2017	16	58	761300	24014151

Source: Australian Bureau of Statistics, Dec 2019 *Australian Demographic Statistics*, cat. no. 3101.0, ABS, Canberra.

- a Draw a line chart depicting the deaths in custody data.
- b Express each year's Aboriginal and Torres Strait Islander deaths as a proportion of the total deaths. Draw a line chart for these proportions and the proportion of Aboriginal and Torres Strait Islander peoples in the total Australian population.
- c Comment on the graph.

Computer applications

4.31 XR04-31 Average petrol prices in Australia and for the six states and the Northern Territory from 2002 to 2019 were recorded. Draw appropriate graphical representations of the data and describe your observations about petrol prices in Australia and by State/Territory.

4.32 XR04-32 The daily US dollar to Japanese Yen exchange rates from 1 Jan 2018 to 31 May 2019 are stored in the file.

- a Draw a line chart using these data.
- b Briefly describe what the chart tells you about the exchange rates.

Source: <https://au.investing.com/currencies/usd-jpy-historical-data>

4.33 XR04-339 The three key daily interest rates for 2-year, 3-year and 5-year bonds in Australia from 21 May 2013 to 21 June 2019 are recorded. Provide suitable graphs of the three interest rates and compare them.

Source: Various Reserve Bank of Australia Bulletins,
<https://www.rba.gov.au/statistics/tables/?v=2019-06-25-01-17-11#interest-rates>

4.34 XR04-34 The gross domestic product (GDP) is the total economic output of a country. It is an important measure of the wealth of the country. The quarterly data for the Australian GDP for March 1960 to March 2019 are stored in the file. Create a suitable graph to depict the data.

Source: Various Reserve Bank of Australia Bulletins

4.35 XR04-35 The daily new cases of coronavirus in Australia and New Zealand are recorded from January–June 2020.

- a Select an appropriate graphical method to present the data.
- b Comment on the findings.

4.36 XR04-36 The monthly sales of homes and apartments in New Zealand for the years 2013–20 are recorded. Select an appropriate graphical method to present both series on the same plot. Discuss the trends and similarities in the two series.

4.3 Describing the relationship between two or more numerical variables

In Sections 4.1 and 4.2, we presented *univariate* graphical techniques that are used to summarise single sets of numerical data. There are many situations where we wish to depict the relationship between two variables; in such cases we use *bivariate* methods. When there are more than two variables, we use *multivariate* methods.

4.3a Graphing the relationship between two numerical variables

Statistics practitioners frequently need to know how two numerical (quantitative) variables are related. For example, financial analysts need to understand how the returns of individual shares and the returns of the entire market are related. Marketing managers need to understand the relationship between sales and advertising. Economists develop statistical techniques to describe the relationship between variables such as unemployment rate and the rate of inflation. The graphical technique used to describe the relationship between two numerical (quantitative) variables is the **scatter diagram**.

To draw a scatter diagram we need data for two variables. In applications where one variable depends to some degree on the other variable, we label the dependent variable Y and the other, called the independent variable, X . For example, an individual's income depends somewhat on the number of years of education. Accordingly, we identify income as the dependent variable and label it Y , and we identify years of education as the independent variable and label it X . Here is another example, in which we consider the relationship between household income and expenditure on food. In this example, as expenditure on food is expected to depend on household income, expenditure on food is the dependent variable (Y) and household income

scatter diagram

A plot of points of one variable against another which illustrates the relationship between them.

linear relationship

One in which two variables move proportionately.

positive relationship

A relationship in which the variables move in the same direction.

negative relationship

A relationship in which the variables move in opposite directions to each other.

is the independent variable (X). In cases where there is no dependency evident, we label the variables arbitrarily.

Each pair of values of X and Y constitutes a point on the graph. A **linear relationship** is one that can be graphed with a straight line. If the two variables generally move in unison – that is, their values tend to increase together or decrease together (for example, price and quantity supplied of a consumer good) – we say that there is a **positive relationship**. If the two variables generally move in opposite directions (for example, price and quantity demanded of a consumer good) – that is, when the value of one variable increases the value of the other variable decreases – we say that there is a **negative relationship**.

To illustrate, consider the following example.

EXAMPLE 4.5

LO3

House size versus construction price

XM04-05 One of the factors that determines the building cost of a house is its size. A real estate agent wanted to know to what extent the construction price of a house is related to the size (number of squares) of the house. He took a sample of 15 houses that had been recently built in a suburb and recorded the construction price and the size of each. These data are listed in Table 4.9. Draw a scatter diagram for these data, and describe the relationship between house size and construction price.

TABLE 4.9 House size (squares) and construction price (\$'0000)

House size (squares)	Construction price (\$'0000)	House size (squares)	Construction price (\$'0000)
20	22	29	35
21	26	30	34
31	40	26	24
32	38	33	38
24	30	27	33
25	31	34	40
22	27	28	35
23	27		

Solution

In this example, as we expect the construction price of a house to be determined by the size of the house, construction price is the dependent variable, labelled Y , and house size is the independent variable, labelled X .

The fifteen pairs of values for construction price (Y) and house size (X) are plotted manually in **Figure 4.12**. The pattern of the resulting scatter diagram provides us with two pieces of information about the relationship between these two variables.

We first observe that, generally speaking, construction price (Y) tends to increase as house size (X) increases. Therefore, there is a positive relationship between the two variables. The second observation is that the relationship between construction price and house size appears to be linear. Although not all 15 points lie on a straight line, we can imagine drawing a straight line through the scatter diagram that approximates the positive linear relationship between the two variables. Finding the straight line that ‘best fits’ the scatter diagram will be addressed in Chapters 5 and 15.



**FIGURE 4.12** Scatter diagram for Example 4.5

Using the computer

Excel output for Example 4.5



COMMANDS

- 1 Type or import the data into two adjacent columns. The variable to appear on the vertical axis must be in the second column. (**XMO4-05**)
- 2 Highlight cells B1 to C16.
- 3 Click **INSERT**. In the **Charts** submenu, select the **Scatter chart** icon. Then select the first scatter chart.
- 4 Click inside the scatter diagram. Click on **Chart title** and type the appropriate chart title. On the menu bar select the **DESIGN** tab under **CHART TOOLS**, and then select **Add Chart Element** in the **Chart Layout** submenu. Add the axis titles by selecting **Axis titles** and clicking **Primary horizontal** and then **Primary vertical**. (Chart title: **Scatter Plot**; Value (X) axis: **House size (squares)**; Value (Y) axis: **Construction price (\$'000)**)
- 5 Go to the **Gridlines** menu and remove any check marks to ensure no horizontal or vertical lines will appear in the final chart. If you wish to change the scale on the axes of the chart, proceed as follows:
- 6 Right click on the Y-axis. Select **Format Axis...** and make any necessary changes under **Axis Options** in the menu that appears on the right-hand side of the screen. Click the cross at the top right corner of the menu to close it.
- 7 Repeat the same for the X-axis (if necessary).

To fit a linear trend line of the scatter plot, right click on any data point and select **Add Trendline...** Then select **Linear** under **TRENDLINE OPTIONS** in the menu that appears on the right-hand side of the screen (and if required click **Display Equation on chart**). Click the cross to **close the menu**.

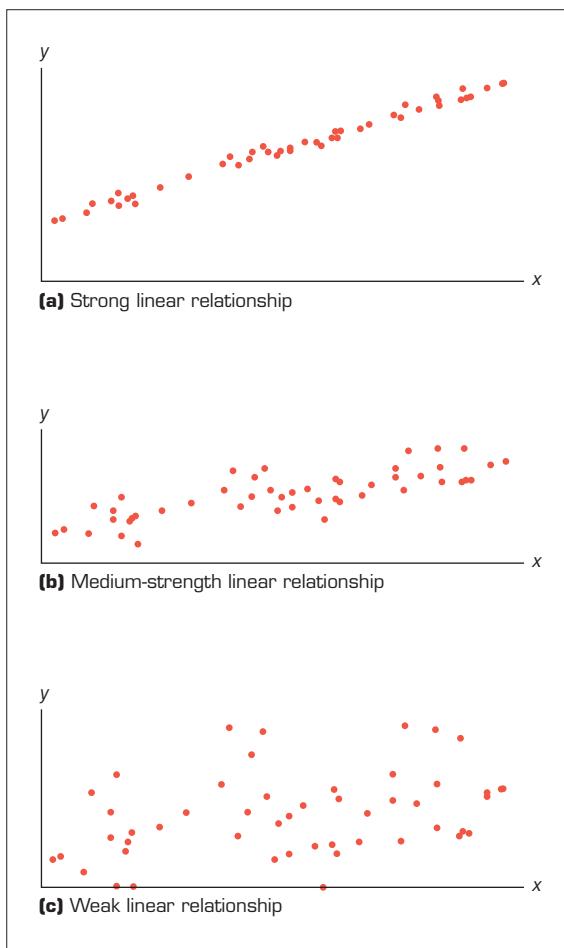
4.3b Patterns of scatter diagrams

As was the case with histograms, we frequently need to describe how two variables are related. The two most important characteristics are the strength and direction of the linear relationship.

Linearity

To determine the strength of the linear relationship, we draw a straight line through the points in such a way that the line represents the relationship. If most of the points fall close to the line, we say that there is a *linear relationship*. If most of the points appear to be scattered randomly away from a straight line, there is no or, at best, a weak linear relationship. **Figure 4.13** depicts several scatter diagrams that exhibit various levels of linearity.

FIGURE 4.13 Scatter diagrams depicting linearity



In drawing the line by hand, we would attempt to draw it so that it passes through the middle of the data. Unfortunately, different people drawing a straight line through the same set of data will produce somewhat different lines. Fortunately, statisticians have produced an objective way to draw the straight line. The method is called the *least squares method*, and it will be presented in Chapter 5 and used in Chapters 15–16.

Note that there may well be some other type of relationship, such as a quadratic or exponential one.

Direction

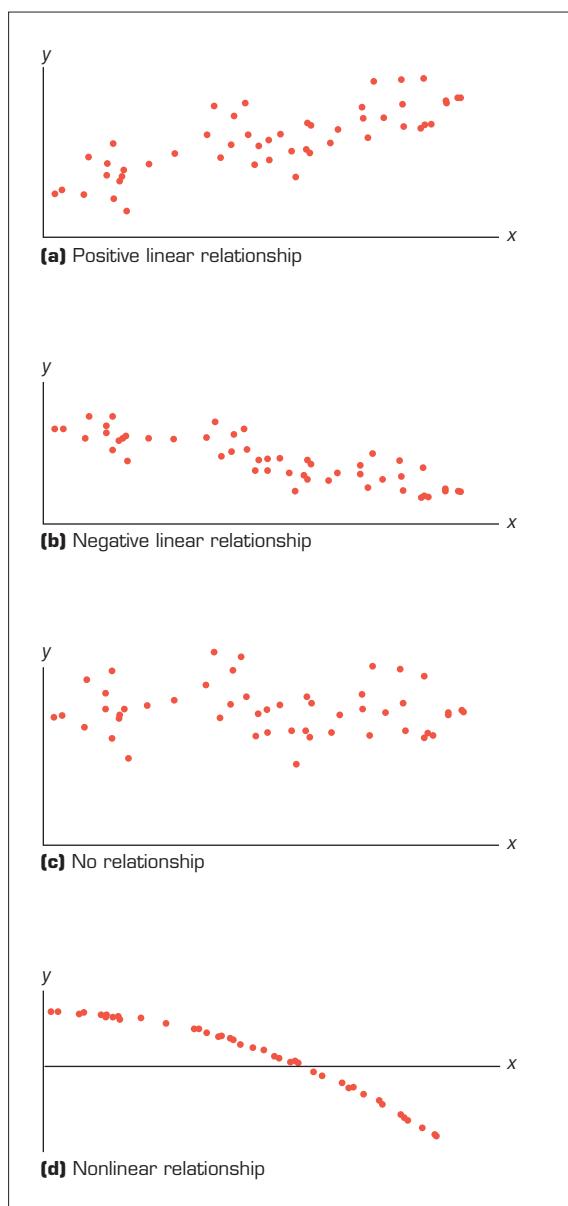
In general, if one variable increases when the other does, we say that there is a *positive linear relationship*. When the two variables tend to move in opposite directions, we describe the nature of their association as a *negative linear relationship*. (The terms positive and negative will be explained in Chapter 5.) See **Figure 4.14** for examples of scatter diagrams depicting a positive linear relationship, a negative linear relationship, no relationship, and a non-linear relationship.

4.3c Interpreting a strong linear relationship

In interpreting the results of a scatter diagram it is important to understand that if two variables are linearly related it does not mean that one is causing the other. In fact, we can never conclude that one variable causes another variable. We can express this more eloquently as ‘correlation is not causation’.

We are now in a position to address the question posed in the introduction to this chapter.

FIGURE 4.14 Scatter diagrams describing direction



SPOTLIGHT ON STATISTICS

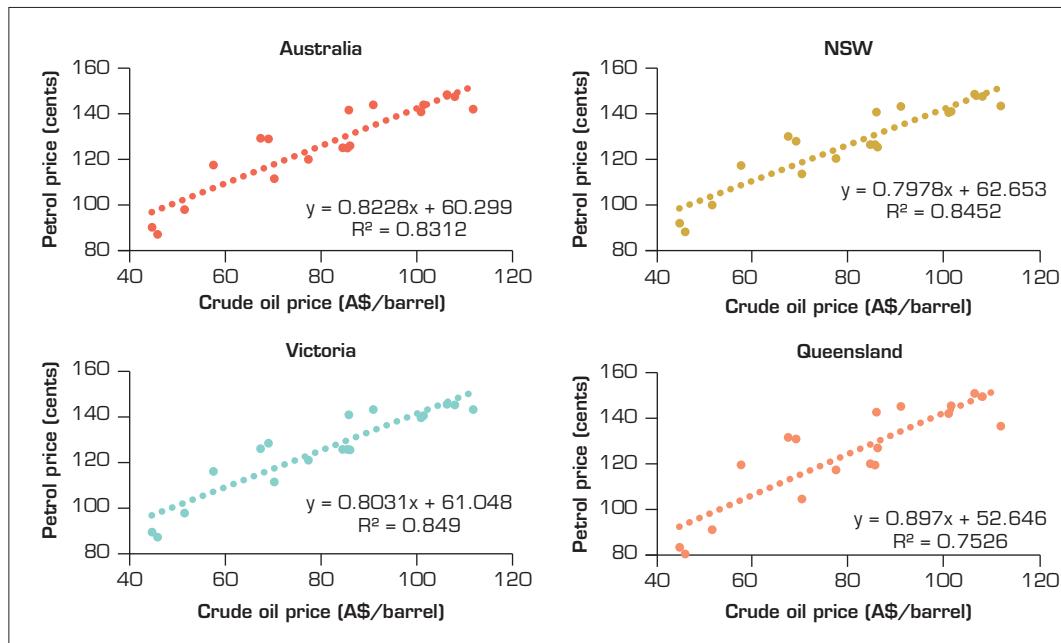
Were oil companies gouging Australian customers? Solution

To determine whether drivers' perceptions that oil companies were gouging consumers, we need to determine whether and to what extent the two variables are related. We consider the crude oil price and average petrol prices in Australia as well as three major states, New South Wales, Victoria and Queensland, individually. Part of the data are listed below.

Year	World crude oil price (A\$/barrel)	Average petrol pump price (in Australian cents/litre)			
		Australia	NSW	Victoria	Queensland
2002	45.77	87.30	88.40	87.70	80.90
2003	44.62	90.40	92.20	89.90	83.80
2004	51.41	98.20	100.20	98.20	91.50
2005	70.14	111.80	113.90	111.80	104.90
.
.
2014	106.20	148.80	149.00	145.70	151.00
2015	67.28	129.60	130.40	126.30	131.80
2016	57.44	117.80	117.60	116.40	119.80
2017	68.92	129.30	128.30	128.70	131.10
2018	90.80	144.30	143.60	143.40	145.30
2019	85.70	142.00	141.10	141.10	142.80

The appropriate statistical technique is the scatter diagram. We label the price of petrol Y and the price of crude oil X. **Figure 4.15** displays the scatter diagrams for Australia and the three states, NSW, Victoria and Queensland.

FIGURE 4.15 Scatter diagrams: Average petrol prices vs world crude oil prices, Australia and three states, 2002–19



Source: iStock.com/Dicraftsman

▶ Interpreting the results

The scatter diagrams reveal that, in Australia nationally and in NSW, Victoria and Queensland, crude oil price and petrol price are strongly, positively and linearly related. As the crude oil price increases, the petrol price also increases. However, it also means that as the crude oil price decreases, petrol price also decreases. Therefore, based on the visual display, overall there is no evidence of oil companies gouging Australians.

4.3d Graphing the relationship between three numerical variables

A *bubble chart* is an elegant two-dimensional alternative to a three-dimensional graph to depict relationships between three numerical variables. It is an extension of a scatter diagram in which the points are replaced by bubbles (discs). The first two variables are represented by the x - and y -axis variables, and the third by the bubbles. The sizes (areas) of the bubbles are proportional to the values of the third variable. For example, suppose we wish to represent the average height, weight and head circumference of newborn babies from countries in Australasia. Information about all three variables can be shown in one chart, a bubble chart.

If the third variable takes some negative values, then positive and negative values are sometimes handled by the use of full circles for positive, and empty circles for negative values.

To illustrate, consider the following example.

EXAMPLE 4.6

LO3

Australia's merchandise trade with other regions

XM04-06 The following table shows the value of Australian merchandise trade with four selected regions of the world in 2018. The imports, exports and trade deficits with the four regions, namely Africa, the Americas, Europe and Oceania, for the year 2018 are recorded in Table 4.10. Depict the information in an appropriate chart.

TABLE 4.10 Australian merchandise trade (in \$m) with four regions, 2018

Region	Exports (\$m)	Imports (\$m)	Trade deficit (\$m)
Africa	4954	5734	780
Americas	30276	60592	30316
Europe	34982	83663	48681
Oceania	18799	19622	823

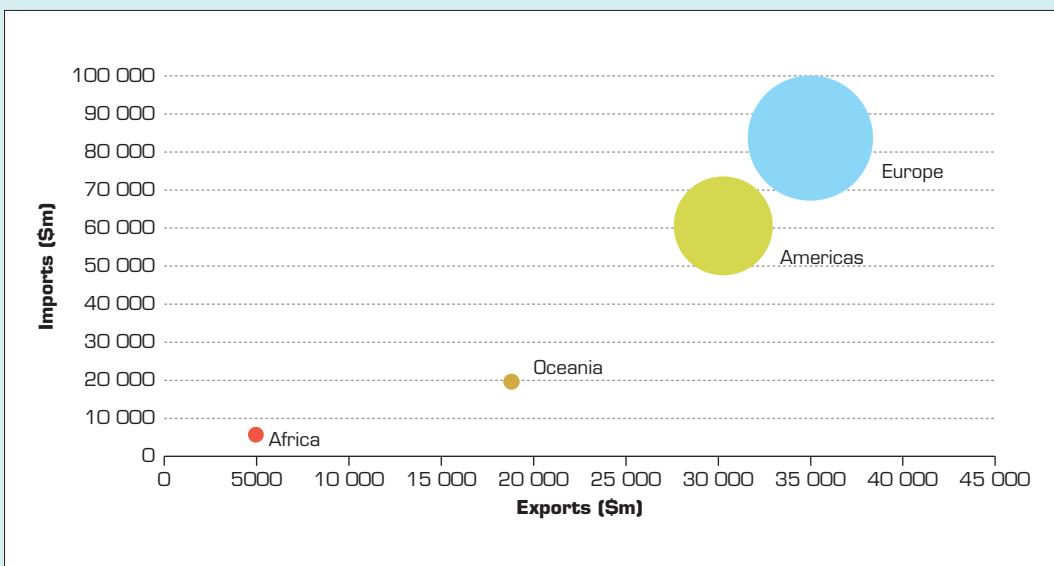
Solution

Table 4.10 provides data for Australia's imports, exports and trade deficits with four regions, Africa, Americas, Europe and Oceania, for the year 2018. We have three numerical values for each region to be displayed. This information can be easily displayed in a bubble chart, as shown in Figure 4.16, which is constructed using Excel. Imports and exports are depicted on the vertical and horizontal axes respectively, while trade deficits are expressed as bubbles of different magnitudes.

Using the computer

Excel output for Example 4.6

FIGURE 4.16 Australian merchandise trade (in \$m), Four regions, 2018



COMMANDS

- 1 Type the data in four adjacent columns or open the data file (**XM04-06**). The variable to appear on the horizontal axis must be in the second column.
 - 2 Highlight cells B3 to D6.
 - 3 Click **INSERT**. In the **Charts** submenu, select the **Scatter or Bubble chart** icon. Then select the first bubble chart.
 - 4 Click inside the diagram. Click on **Chart title** and type the appropriate chart title. On the menu bar select the **DESIGN** tab under **CHART TOOLS**, and then select **Add Chart Element** in the **Chart Layout** submenu. Add the axis titles by selecting **Axis titles** and clicking **Primary horizontal** and then **Primary vertical** (Chart title: **Bubble chart**; Value (X) axis: **Exports (\$m)**; Value (Y) axis: **Imports (\$m)**).
 - 5 Right click on the horizontal gridlines and press delete. Repeat this for the vertical gridlines.
 - 6 If there are negative values in the data for the third variable (represented by the bubbles), right click on any of the bubbles and select **Format Data Series**. Under **Series Options**, tick **Show negative bubbles**. Positive values will have fully shaded bubbles and negative values will have empty bubbles, both with sizes proportional to their magnitudes.
- To add labels for the bubbles, proceed as follows:
- 7 Right click on any of the bubbles and select **Add Data Labels**.
 - 8 Double click on any of the data labels. From the **Format Data Labels** menu appearing on the far right of the plot, click on **Label Option**.
 - 9 Select **Label Options** and tick the box for **Value From Cells**. In the **Data Label Range** box, type the cell range with country names (**A2:A5**) and click **OK**. De-select the **Y-Value** box.

This will give you a bubble chart like that in Figure 4.16.

4.3e Another visual representation of numerical data

A *heat map* is a graphical technique or data visualisation that shows the magnitude of the data as colour in a two-dimensional plot. The colour variation shows the intensity or ascending nature of the data. For example, suppose we want to compare the spending behaviour of consumers in three groups of countries (developed, developing and underdeveloped countries) by comparing the average budget shares of groups of commodities in these countries. This can be depicted graphically using a heat map.

To illustrate, consider the following example.

EXAMPLE 4.7

LO3

Consumer Price Index in 10 OECD countries

XMO4-07 Since 2000, the increase in consumer prices in Australia has been more than the increase in consumer prices in a number of other OECD (Organisation for Economic Co-operation and Development) countries. The following table shows the consumer price index (CPI) of all goods and services in Australia and a number of similar OECD countries for the years 2015–18, with base year 2015 = 100. Visually display the data showing the comparative CPI values in the 10 countries using an appropriate graphical technique.

TABLE 4.11 Consumer Price Index in 10 OECD countries (base year 2015 = 100)

Country	2015	2016	2017	2018
Australia	100.0	101.3	103.3	105.2
Canada	100.0	101.4	103.0	105.4
France	100.0	100.2	101.2	103.1
Germany	100.0	100.5	102.0	103.8
Italy	100.0	99.9	101.1	102.3
Japan	100.0	99.9	100.4	101.3
New Zealand	100.0	100.6	102.5	104.1
Sweden	100.0	101.0	102.8	104.8
United Kingdom	100.0	101.0	103.6	106.0
United States	100.0	101.3	103.4	105.9

Solution

Table 4.11 provides CPI data for 10 OECD countries. Our aim is to compare the price levels across the 10 countries with base year 2015 = 100. A heat map would provide a visual comparison of the CPI across the 10 countries with the magnitude of the values represented by different shades of colour. Here higher values are represented by darker shades. The last column provides a general trend of the data.

Using the computer

Excel output for Example 4.7

FIGURE 4.17 CPI (2015 = 100), 10 OECD countries, 2015–18



COMMANDS

- 1 Type the data in five adjacent columns or open the data file (**XM04-07**).
- 2 Highlight cells B3 to E12.
- 3 On the menu bar select **Data** and then choose **Conditional Formatting** from the **Styles** group. Select any **colour scale**; for example, **Red-White Colour Scale**.
- 4 If you want to show the movement of colour change in a line graph, in cell F2, type Trend. Click on cell F3, click on **Insert** on the menu bar and then select **Line** in the **Sparklines** group.
- 5 In the **Sparklines create** dialog box, type B2:E2 in the **Data Range** and F3 for the **Location range**. Click **OK**.
- 6 Copy the formula in cell F3 to cells F4 to F12 (by dragging the mouse down from the right-hand corner of cell F3 which appears as **+**).

This will result in the heat map shown in Figure 4.17.

We close this section by reviewing the factors that identify the use of the scatter, bubble and heat diagrams.

IN SUMMARY

Factors that identify when to use a scatter diagram

- 1 *Objective*: to describe the relationship between two variables
- 2 *Data type*: numerical (quantitative)

Factors that identify when to use a bubble chart

- 1 *Objective*: to describe the relationship between three variables
- 2 *Data type*: numerical (quantitative)

Factors that identify when to use a heat map

- 1 *Objective*: to describe the intensity or magnitude of a variable across a number of categories or periods
- 2 *Data type*: numerical (quantitative)

EXERCISES

Learning the techniques

4.37 XR04-37 For each of the following datasets, plot the points on a graph and determine whether a linear model is reasonable.

a	x	2	3	5	7	9
	y	6	9	4	7	8

b	x	1	3	5	4	7
	y	5	7	10	9	16

c	x	7	9	2	3	6
	y	4	1	6	10	5

4.38 XR04-38 The annual bonuses (\$'000) of six randomly selected employees and their years of service were recorded and are listed below.

Years (x)	1	2	3	4	5	6
Bonus (y)	6	1	9	5	17	12

- a Draw the scatter diagram for these data.
- b Draw a line through the data that approximate the relationship between years of service and annual bonus.

- 4.39 XR04-39** The observations of two variables were recorded as shown below.

x	1	2	3	4	5	6	7	8	9
y	5	28	17	14	27	33	39	26	30

- a Draw the scatter diagram for these data.
- b Draw a straight line that approximates reasonably well the relationship between the two variables.

Applying the techniques

- 4.40 XR04-40 Self-correcting exercise.** The owner of a small business in Tasmania has experienced fairly uniform monthly sales levels in previous years. This year, he decided to vary his advertising expenditures from month to month to see if that would have a significant impact on the sales level. To help assess the effect of advertising on sales level, he collected the data shown in the following table. Draw a scatter diagram for these data, and describe the relationship between advertising expenditure and sales level.

Month	1	2	3	4	5	6	7	8
Advertising expenditure, X (\$'000)	1	3	5	4	2	5	3	2
Sales level, Y (\$'000)	30	40	40	50	35	50	35	25

- 4.41 XR04-41** Because inflation reduces the purchasing power of the dollar, investors seek investments that will provide protection against inflation; that is, investments that will provide higher returns when inflation is higher. It is frequently stated that ordinary shares provide just such a hedge against inflation. The annual Australian inflation rate (as measured by percentage changes in the consumer price index) and the annual All Ordinaries Index from 1995 to 2018 are shown in the table below.

Year	Rate of inflation (%)	All Ordinaries Index	Year	Rate of inflation (%)	All Ordinaries Index
1995	4.6	2000.8	2004	2.3	3499.8
1996	2.6	2231.7	2005	2.7	4197.5
1997	0.3	2662.7	2006	3.6	4933.5
1998	0.9	2608.2	2007	2.3	6337.6
1999	1.5	2963.0	2008	4.4	5513.5
2000	4.5	3115.9	2009	1.8	4127.6
2001	4.4	3352.4	2010	2.9	4632.8
2002	3.0	3241.5	2011	3.3	4553.9
2003	2.7	3032.0	2012	1.8	4385.2

2013	2.5	5110.5	2016	1.3	5379.4
2014	2.5	5423.9	2017	1.9	5859.1
2015	1.5	5510.0	2018	1.9	6092.4

Source: <https://au.finance.yahoo.com/q/hp?s=%5EAORD>, and <http://www.rateinflation.com/inflation-rate/australia-historical-inflation-rate>

- a Draw a scatter diagram for these data with the All Ordinaries Index on the vertical axis.
- b Describe the relationship between the All Ordinaries Index and the rate of inflation over the period from 1995 to 2018.
- c Draw a straight line that approximates reasonably well the relationship between the All Ordinaries Index and the rate of inflation.

- 4.42 XR04-42** A firm's operating costs can be classified as fixed, variable or mixed. Costs for items such as telephone, electricity and maintenance are often mixed costs, meaning they have both a fixed and variable cost component. A manufacturing firm has recorded its electricity costs and the total number of hours of machine time for each of 12 months, in the table below.

- a Draw a scatter diagram for these data, with the cost of electricity on the vertical axis.
- b Describe the relationship between electricity costs and hours of machine usage.
- c Draw a straight line that approximates reasonably well the relationship between electricity cost and machine usage.
- d Use the line drawn in part (c) to estimate the monthly fixed cost of electricity and the variable cost of electricity per thousand hours of machine time.

Machine time (in '000 hours)	Cost of electricity (\$)
6	760
9	1000
8	890
7	880
10	1070
10	1030
5	700
7	910
6	745
9	950
8	870
11	1040

- 4.43 XR04-43** At a university where calculus is a prerequisite for the statistics course, the marks for calculus and statistics were recorded for a sample of 15 students. The data are as follows:

Calculus	65	58	93	68	74	81	58	85
Statistics	74	72	84	71	68	85	63	73
Calculus	88	75	63	79	80	54	72	
Statistics	79	65	62	71	74	68	73	

- 4.44 XR04-44** The growing interest in and use of the internet has forced many companies to consider ways of selling their products online. Therefore, it is of interest to these companies to determine who is using the internet. A statistics practitioner undertook a study to determine how education and internet use are connected. She took a random sample of 15 adults (20 years of age and older) and asked each to report the number of years of education they had completed and the number of hours of internet use in the previous week. These data follow.

Education (years)	11	11	8	13	17	11	11	11
Internet use (hours)	10	5	0	14	24	0	15	12
Education (years)	19	13	15	9	15	15	11	
Internet use (hours)	20	10	5	8	12	15	0	

- 4.45 XR04-45** A statistics professor formed the opinion that students who handed in quizzes and exams early outperformed students who handed in their papers later. To develop data to decide whether her opinion is valid, she recorded the amount of time (in minutes) taken by students to submit their mid-semester tests (time limit 90 minutes) and the subsequent mark for a sample of 12 students.

Time (min)	90	73	86	85	80	87
Mark	68	65	58	94	76	91
Time (min)	90	78	84	71	72	88
Mark	62	81	75	83	85	74

- a** Draw a scatter diagram of the data.
b What does the graph tell you about the relationship between the time taken to hand in the test and the subsequent mark?

Computer applications

- 4.46 XR04-46** The average annual (end of financial year) Australian gold and silver Perth Mint spot prices over a 19-year period (2000–18) were recorded. Column 1 stores the year, column 2 stores the gold price and column 3 stores the silver price.

- a** Draw a scatter diagram for these data with the gold price on the horizontal axis and the silver price on the vertical axis.
b Describe the relationship between the gold price and silver price.

- 4.47 XR04-47** An increasing number of high school students have part-time jobs, but how does this affect their performance at school? To help answer this question, a sample was taken of 300 high school students who have part-time jobs. Each student reported his or her most recent average mark and the number of hours per week at his or her job.

- a** Draw a scatter diagram of the data, including a straight line.
b Does it appear that there is a linear relationship between the two variables?

- 4.48 XR04-48** Who uses the internet? To help answer this question, a random sample of 300 adults was asked to report their age and the number of hours of internet use weekly.
- a** Draw a suitable graph to depict the data.
b Does it appear that there is a linear relationship between the two variables? If so, describe it.

Learning the techniques

- 4.49 XR04-49** Imports, Exports and Trade Surplus as a percentage of GDP for six OECD countries for 2016 is recorded. Depict the data in an appropriate chart and describe the chart.

- 4.50 XR04-50** Refer to Exercise 3.12, in which Australian exports and imports with 10 major trading partners (in A\$ millions) for 2018 are shown.
- a** Calculate the trade deficit (= Imports – Exports) for each trading partner.
b Draw a suitable graph to depict the data for the three variables, Imports, Exports and Trade deficit. Comment on the chart.

- 4.51 XR04-51** Oil reserves, oil production and oil consumption for 15 oil producing countries in 2018 are recorded. Draw a suitable graph to depict the data for the three variables. Comment on the chart.

4.52 XR04-52 Data for median (established) house prices in the capital cities of the six states and two territories in Australia for the March quarter of 2002–2019 were collected from the Australian Bureau of Statistics website (cat. no 6416.0 – Residential Property Price Indexes: Eight Capital Cities, Dec 2019, Tables 4 and 5, March Quarter 2019). Present a visual representation of the data for the capital cities and territories of Australia during the period using a heat map.

4.53 XR04-53 Apparent per capita consumption and average prices of four meat types in Sydney are recorded for the years 2001, 2005, 2015 and 2020. The four meat types are beef (rump steak), lamb (leg), pork (leg) and chicken (frozen).

- a Use a visual depiction of the trend in the average prices of the four types of meat over the time period.
- a Use a visual depiction of the movement in the per capita consumption of the four types of meat over the time period.

4.4 Graphical excellence and deception

In this chapter and in Chapter 3, we introduced a number of graphical techniques. The emphasis was on how to draw each one manually and how to command the computer to draw them. In this section, we discuss how to use graphical techniques effectively. We introduce the concept of **graphical excellence**, which is a term we apply to techniques that are informative and concise and that impart information clearly to their viewers. Additionally, we discuss an equally important concept: graphical integrity and its enemy, graphical deception.

4.4a Graphical excellence

Graphical excellence is achieved when the following characteristics apply.

- 1 **The graph presents large data sets concisely and coherently.** Graphical techniques were created to summarise and describe large data sets. Small data sets are easily summarised with a table. One or two numbers can best be presented in a sentence.
- 2 **The ideas and concepts the statistician wants to present are clearly understood by the viewer.** The chart is designed to describe what would otherwise be described in words. An excellent chart is one that can replace a thousand words and still be clearly comprehended by its readers.
- 3 **The graph encourages the viewer to compare two or more variables.** Graphs displaying only one variable provide very little information. Graphs are often best used to depict relationships between two or more variables or to explain how and why the observed results occurred.
- 4 **The display induces the viewer to address the substance of the data and not the form of the graph.** The form of the graph is supposed to help present the substance. If the form replaces the substance, the chart is not performing its function.
- 5 **There is no distortion of what the data reveal.** You cannot make statistical techniques say whatever you like. A knowledgeable reader will easily see through distortions and deception. We describe graphical deception later in this section.

Edward Tufte, professor in statistics at Yale University in the US, summarised graphical excellence this way:

- Graphical excellence is the well-designed presentation of interesting data – a matter of substance, of statistics, and of design.
- Graphical excellence is that which gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.
- Graphical excellence is nearly always multivariate.
- Graphical excellence requires telling the truth about the data.

Now let's examine the chart that has been acclaimed the best chart ever drawn.

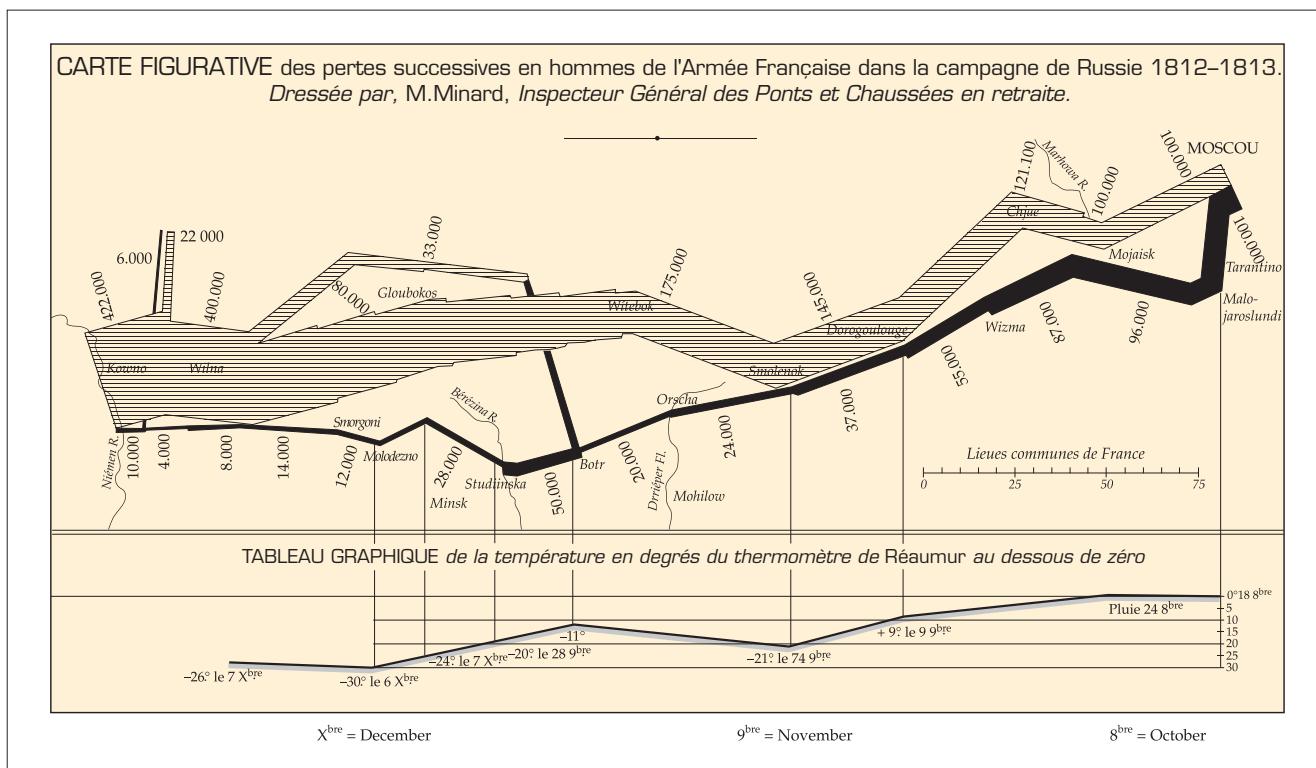
Figure 4.18 depicts Minard's graph. The striped band is a time series depicting the size of the army at various places on the map, which is also part of the chart. When Napoleon invaded Russia by crossing the Niemen River on 21 June 1812, there were 422000 soldiers.

graphical excellence

Application of graphical techniques that are informative and concise and that impart information clearly.

By the time the army reached Moscow, the number had dwindled to 100 000. At that point, the army started its retreat. The black band represents the army in retreat. At the bottom of the chart, we see the dates starting with October 1812. Just above the dates, Minard drew another time series, this one showing the temperature. It was bitterly cold and many soldiers died of exposure. As you can see, the temperature dipped to $-30^{\circ}\text{Réaumur}$ (-37.5°C) on 6 December. The chart is effective because it depicts five variables clearly and succinctly.

FIGURE 4.18 Chart depicting Napoleon's invasion and retreat from Russia in 1812



Source: Reprinted by permission, Edward R. Tufte, *The Visual Display of Quantitative Information* (Cheshire, Connecticut: Graphics Press LLC, 1983, 2001), p. 41.

4.4b Graphical deception

The use of graphs and charts is pervasive in newspapers, magazines, business and economic reports, and seminars. In large part, this is due to the increasing availability of computers and software that allow the storage, retrieval, manipulation and summary of large numbers of raw data. It is, therefore, more important than ever to be able to evaluate critically the information presented by means of graphical techniques. In the final analysis, graphical techniques merely create a visual impression, which is easy to distort. In fact, distortion is easy and commonplace in various government and financial reports. Although the heading for this section mentions deception, it is quite possible for an inexperienced person inadvertently to create distorted impressions with graphs. In any event, you should be aware of possible methods of **graphical deception**. This section illustrates a few of them.

graphical deception
Presentation of a
distorted impression
using graphs.

The first thing to watch for is a graph without a scale or with incorrect labelling on one axis. The time-series graph of a firm's sales in **Figure 4.19** might represent a growth rate of 100% or 1% over the five years depicted, depending on the vertical scale. **Figure 4.20** shows the movement in the Australian dollar. Obviously, the labelling of the y -axis has gone terribly wrong. It is best simply to ignore such graphs.

FIGURE 4.19 Graph without a vertical scale

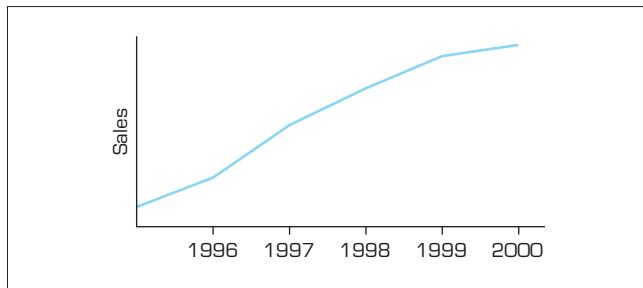
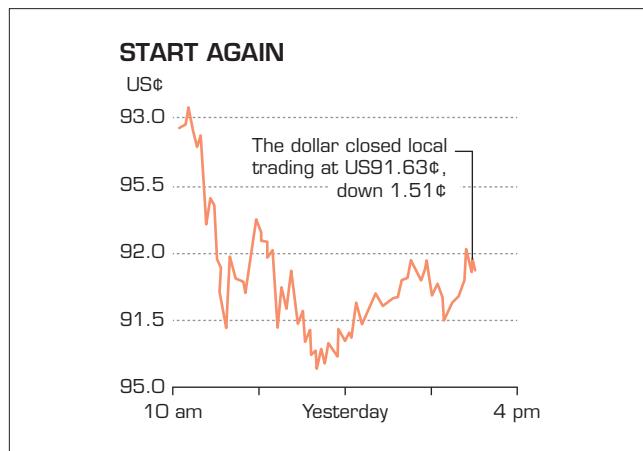


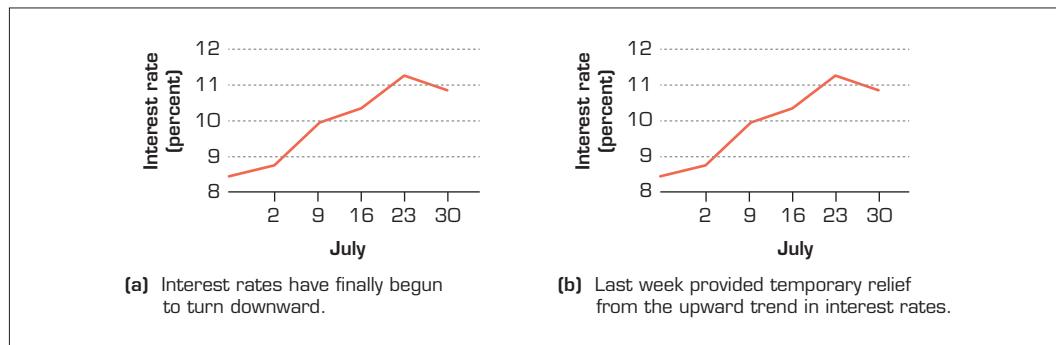
FIGURE 4.20 Value of the Australian dollar



Source: *The Weekend Australian*, 3–4 November 2007.

A second trap to avoid is being influenced by a graph's caption. Your impression of the trend in interest rates might be different depending on whether you read a newspaper carrying caption (a) or caption (b) in **Figure 4.21**.

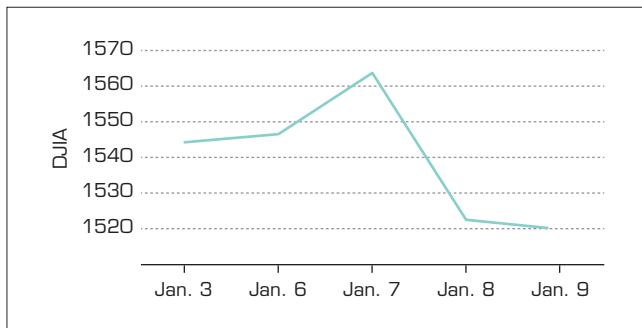
FIGURE 4.21 Different captions for the same graph



Perspective is often distorted if only absolute changes in value, rather than percentage changes, are reported. A \$1 drop in the price of your \$2 stock is relatively more distressing than a \$1 drop in the price of your \$100 stock. On 9 January 1986, many newspapers displayed graphs similar to the one shown in [Figure 4.22](#) and reported that the stock market, as measured by the Dow Jones Industrial Average (DJIA), had suffered its worst one-day loss ever on the previous day. The loss was 39 points, exceeding even the loss of Black Tuesday – 28 October 1929. Although the loss was indeed a large one, many news reports failed to mention that the 1986 level of the DJIA was much higher than the 1929 level. A better perspective on the situation could be gained by noticing that the loss on 8 January 1986 represented a 2.5% decline, whereas the decline in 1929 was 12.8%. As a point of interest, we note that the stock

market was 12% higher within two months of this historic drop and 40% higher one year later. The worst one-day loss ever, 22%, occurred on 19 October 1987.

FIGURE 4.22 Historic drop in DJIA, 1986



We now turn to some rather subtle methods of creating distorted impressions with graphs. Consider the graph in **Figure 4.23**, which depicts the growth in a firm's quarterly sales during the past year, from \$100 million to \$110 million. This 10% growth in quarterly sales can be made to appear more dramatic by stretching the vertical axis – a technique that involves changing the scale on the vertical axis so that a given dollar amount is represented by a greater height than before. As a result, the rise in sales appears to be greater, because the slope of the graph is visually (but not numerically) steeper. The expanded scale is usually accommodated by employing a break in the vertical axis, as in **Figure 4.24(a)**, or by truncating the vertical axis, as in **Figure 4.24(b)**, so that the vertical scale begins at a point greater than zero. The effect of making slopes appear steeper can also be created by shrinking the horizontal axis, so that points on the horizontal axis are moved closer together.

FIGURE 4.23 Quarterly sales for the past year – version 1

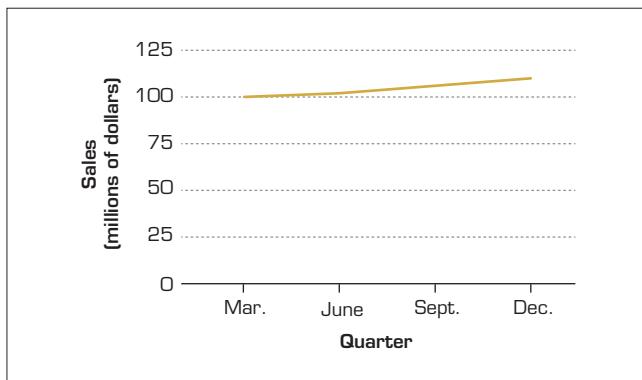
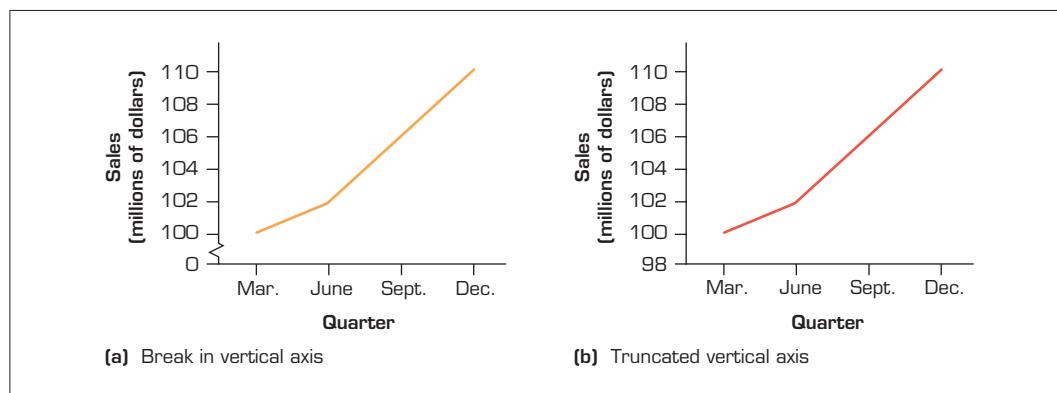
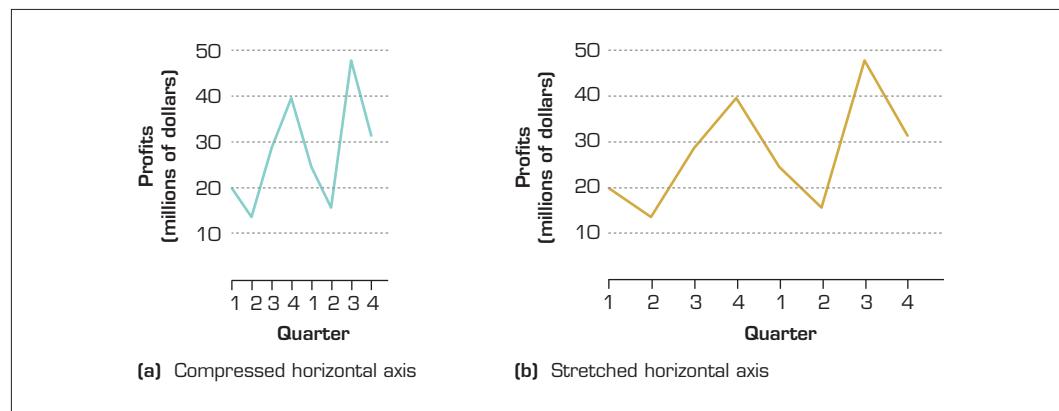


FIGURE 4.24 Quarterly sales for the past year – version 2



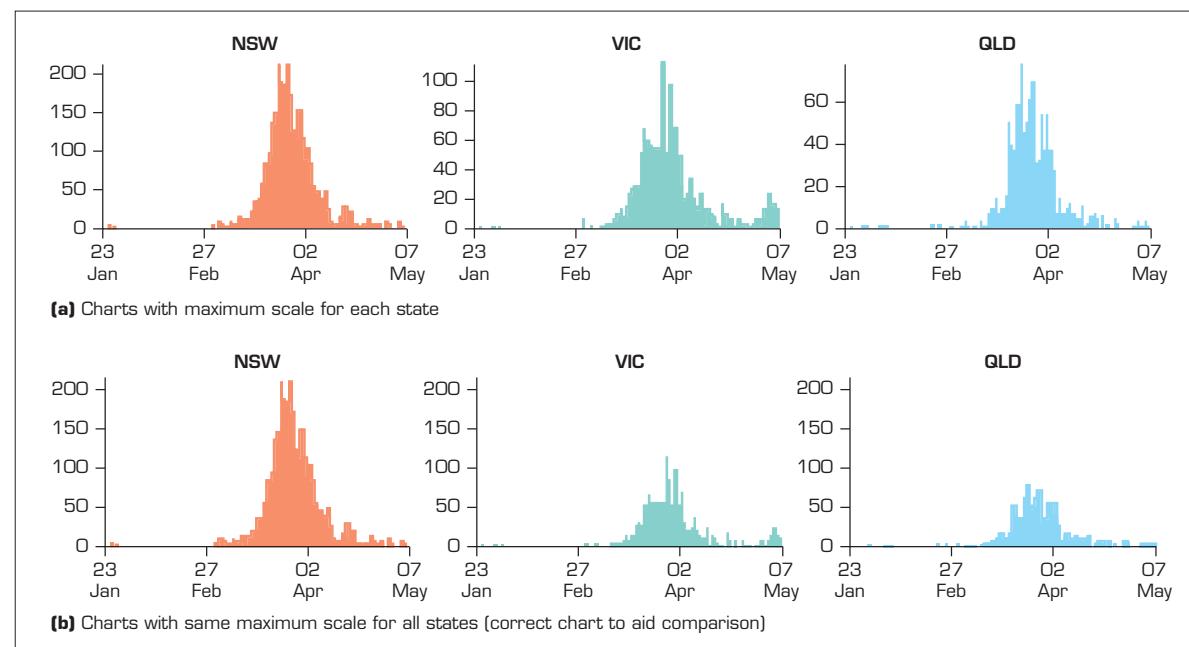
Just the opposite effect is obtained by stretching the horizontal axis – that is, spreading out the points on the horizontal axis to increase the distance between them so that slopes and trends will appear to be less steep. The graph of a firm's profits presented in [Figure 4.25\(a\)](#) shows considerable swings, both upwards and downwards in the profits from one quarter to the next. However, the firm could convey the impression of reasonable stability in profits from quarter to quarter by stretching the horizontal axis, as shown in [Figure 4.25\(b\)](#).

FIGURE 4.25 Quarterly profits over two years



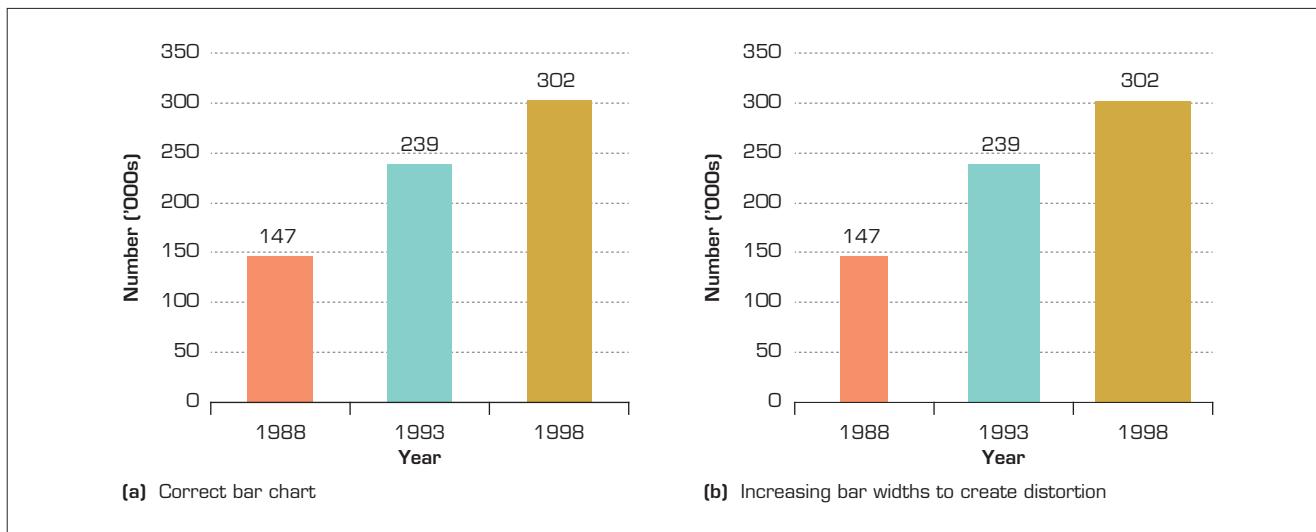
Similar illusions can be created with bar charts by stretching or shrinking the vertical or horizontal axis. One way of distorting data visually is using different scales on the vertical axis when comparing the same variable. [Figure 4.26\(a\)](#) presents the daily count of confirmed Covid-19 cases for three Australian states with maximum scale for each chart (200 for NSW, 100 for VIC and 80 for QLD). [Figure 4.26\(b\)](#) presents the same information with maximum vertical axis scale, the same 200 for all three states. A quick glance of [Figure 4.26\(a\)](#) gives the impression that the numbers of coronavirus cases in the three Australian states have a similar distribution and same magnitude, which is illusionary. However, [Figure 4.26\(b\)](#) uses the same vertical axis scale for all three charts and shows that even though the distributions look similar, the magnitude is very different in the three states.

FIGURE 4.26 Number of coronavirus cases in NSW, Victoria and Queensland, 9 May 2020



Another popular method of creating distorted impressions with bar charts is to draw the bars so that their widths are proportional to their heights. The bar chart in [Figure 4.27\(a\)](#) clearly depicts the estimated number of daily marijuana users aged 14 and over in Australia during three specific years. This chart correctly uses bars of equal width, so that both the height and the area of each bar are proportional to the number of users they represent. The growth in the number of marijuana users is exaggerated in [Figure 4.27\(b\)](#), in which the widths of the bars increase with their heights. A quick glance at this bar chart might leave the viewer with the mistaken impression that the number of marijuana users has increased four-fold over the decade, since the 1998 bar is two times the size of the 1988 bar.

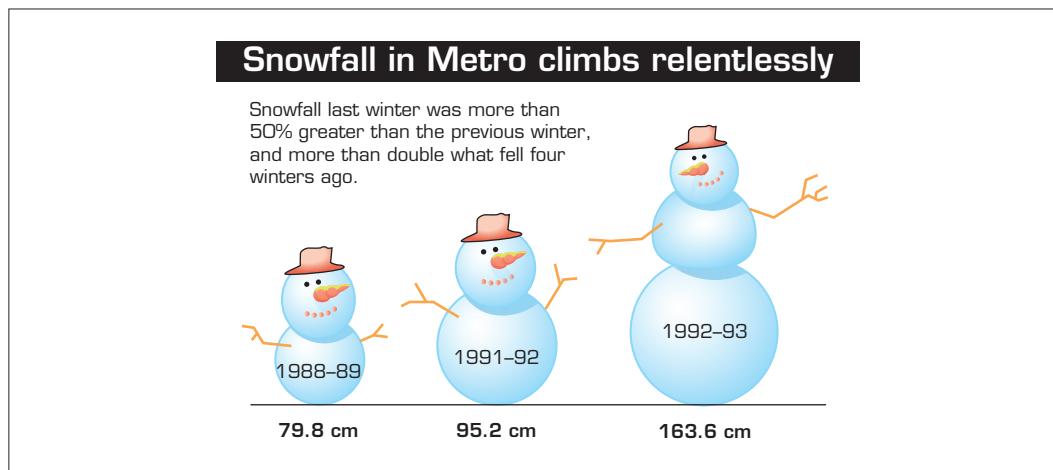
FIGURE 4.27 Estimated number of daily marijuana users aged 14 and over in Australia (in '000) for 1988, 1993, 1998



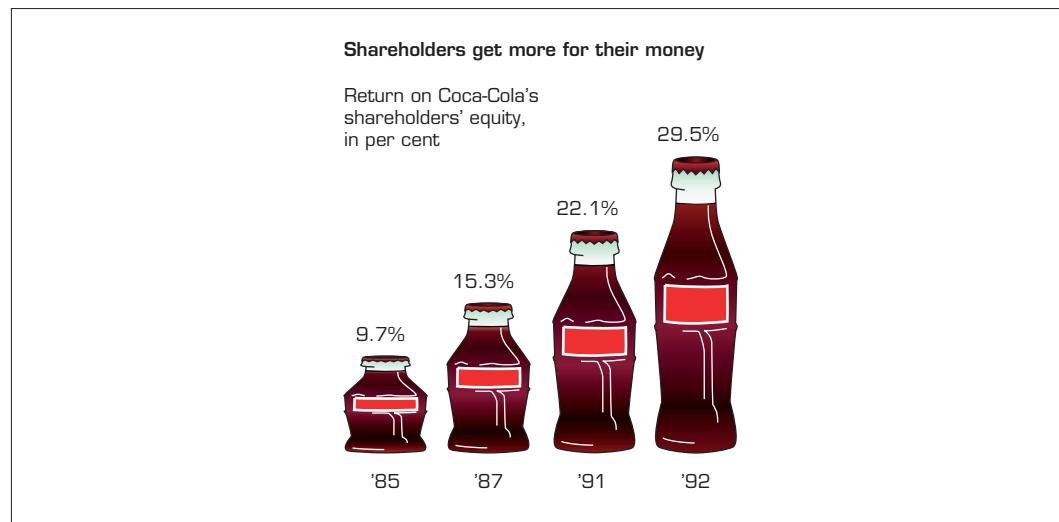
Source: Australian Institute of Health and Welfare 2002, *Australian National Drug Strategy Household Survey*, AIHW, Canberra

You should be on the lookout for size distortions, particularly in pictograms, which replace the bars with pictures of objects (such as bags of money, people or animals) to enhance the visual appeal. [Figure 4.28](#) displays the misuse of a pictogram – the snowman grows in width as well as height. The proper use of a pictogram is shown in [Figure 4.29](#), which effectively uses pictures of Coca-Cola bottles.

FIGURE 4.28 Incorrect pictogram



Source: Environment Canada, Metro Toronto Branch

FIGURE 4.29 Correct pictogram

Source: Value Line Investment Survey, 21 May 1993

The preceding examples of creating a distorted impression using graphs are not exhaustive, but they include some of the more popular methods. They should also serve to make the point that graphical techniques are used to create a visual impression, and the impression you obtain may be a distorted one unless you examine the graph with care. You are less likely to be misled if you focus your attention on the numerical values that the graph represents. Begin by carefully noting the scales on both axes; graphs with unmarked axes should be ignored completely.

4.4c Presenting statistics – Written reports and oral presentations

Throughout this book we will be presenting a variety of statistical techniques. The emphasis will be on applying the correct statistical technique and the proper interpretation of the resulting statistics. However, the ability to communicate your findings, both orally and in writing, is a critical skill. In this section we will provide general guidelines for both forms of communication. Some of the exercises and cases that appear later in this book will ask you not only to employ a statistical technique, but also to prepare a written report or design an oral presentation.

Writing a report

Just as there are many ways to write a statistics textbook, there are also many different ways to write a report. Here is our suggested method. Reports should contain the following steps.

- 1 State your objective.** In Chapter 2, we introduced statistical techniques by describing the type of information needed and the type of data produced by the experiment. For example, there are many studies to determine whether one product or service is better than one or more other similar products or services. You should simply state the purpose of the statistical analysis and include the possible results of the experiment and the decisions that may follow.
- 2 Describe the experiment.** It is important to know your readers. If there are individuals who have little knowledge of statistics, you must keep the report very simple. However, it is likely that some recipients of your report will have some knowledge of the subject and

thus will want to know how the experiment was done. In particular, they will want you to assure them that the experiment was properly conducted.

- 3 Describe your results.** If possible, include charts. In addition to replacing hundreds of words, a well-designed chart exhibits clarity and brevity. Of course, do not include graphs and charts that can more easily be replaced by a table or a sentence. Moreover, place graphs in the body of the report and not in the appendices. Be precise. This requires that you fully understand the statistical technique and the proper way to interpret the statistics.

Be honest. Some statisticians describe this component as an ethical issue; that is, it is a violation of ethical standards to purposely mislead readers of the report. In fact, it is an issue of practicality. For example, what is the purpose of a statistics practitioner lying or exaggerating the effects of a new drug? If, as a result of the statistical report, the company proceeds with the new drug and it turns out to be ineffective – or worse, dangerous – what are the consequences for the statistics practitioner and the company?

- 4 Discuss limitations of the statistical techniques.** A statistical analysis will rarely be definitive. Your report should discuss possible problems with the analysis, including violations of the required conditions, which we will describe throughout this book. It is not necessary to include computer printouts. If you do, be selective. Include only those that are critical to your report.

Making an oral presentation

The general guidelines for making presentations are similar to those for writing reports.

- 1 Know your audience.** Take the time to determine who is in your audience and what kind of information they will be expecting from you. Additionally, determine their level of statistical knowledge. If it is low, do not use statistical terms without defining them.
- 2 Restrict your main points to the objectives, conclusions and recommendations of the study.** Few listeners will be interested in the details of your statistical analysis. Your audience will be most interested in your conclusions and recommendations.
- 3 Stay within your time limit.** If your presentation is expected to exceed your time limit, stick to the main points. Hand out additional written information if necessary.
- 4 Use graphs.** It is often much easier to explain even relatively complex ideas if you provide excellent graphs.
- 5 Provide handouts.** Handouts with copies of your graphs make it easier for your audience to follow your explanations.

Study Tools

CHAPTER SUMMARY

A collection of *numerical (quantitative) data*, that are cross-sectional, can be usefully summarised by grouping observations to form a *frequency distribution*. Drawing a *stem-and-leaf display* is often helpful during preliminary analysis of the data. Either a *histogram* or a *frequency polygon* can then be used to convey the shape of the distribution. Statistics practitioners examine several aspects of the shapes of histograms. These are *symmetry*, the number of *modes*, and its resemblance to a *bell shape*.

We described the difference between *time-series data* and *cross-sectional data*. Time-series data are graphed by either *line charts* or *bar charts*. To analyse the relationship between two numerical (quantitative) variables, we draw a *scatter diagram*. We look for the direction and strength of the *linear relationship*.

The type of chart to use in a particular situation depends on the particular information the user wants to emphasise.

The data files for Examples, Exercises and Cases are provided on the companion website (accessible through <https://login.cengagebrain.com/>).

SUPPLEMENTARY EXERCISES

- 4.54 XR04-54** The number of items rejected daily by a manufacturer because of defects was recorded for the past 30 days. The results are as follows.

4	9	13	7	5	8	12	15	5	7
3	8	15	17	19	6	4	10	8	22
16	9	5	3	9	19	14	13	18	7

- a Draw a histogram.
- b Draw an ogive.
- c Describe the shape of the histogram.

- 4.55 XR04-55** The number of items returned to a leading Brisbane retailer by its customers was recorded for the last 25 days.

21	8	17	22	19
18	19	14	17	11
6	21	25	19	9
12	16	16	10	29
24	6	21	20	25

- a Draw a frequency distribution for these data. Use five class intervals, with the lower boundary of the first class being five days.
- b Draw a relative frequency histogram for these data.
- c What is the relationship between the areas under the histogram you have drawn and the relative frequencies of the observations?

- 4.56 XR04-56** The following table presents the annual average rate of unemployment among male and female Australians from 2000 to 2019.

- a Use a graphical technique to depict the information in the table.
- b Explain why you selected the particular technique you used in part (a).

Rate of unemployment (looked for full-time work), males and females, Australia, 2000–2019

Year	Males	Females	Year	Males	Females
2000	6.3	6.9	2010	4.8	6.1
2001	6.8	7.3	2011	4.7	6.0
2002	6.3	7.2	2012	4.9	6.1
2003	5.7	7.1	2013	5.6	6.6
2004	5.0	6.2	2014	5.8	7.1
2005	4.6	5.8	2015	5.8	6.9
2006	4.3	5.5	2016	5.4	6.6
2007	3.7	5.1	2017	5.2	6.4
2008	3.6	4.8	2018	5.0	6.1
2009	5.5	6.2	2019	5.0	5.9

Source: Australian Bureau of Statistics, May 2019,
Labour Force Australia, cat. no. 6202.0, ABS, Canberra.

- 4.57 XR04-57** A growing concern at universities and TAFE colleges across Australia is the number of academics who will retire in the next 5, 10 and 15 years. To examine this problem, a statistics practitioner took a random sample of 1000 academics and recorded their ages. Assuming that the academics will retire at age 65, use graphical techniques to discuss the retirement problem facing universities over the next 15 years.

4.58 XR04-58 In an effort to track the increasing prices of housing units in a large city, a statistics practitioner took a random sample of units sold this year and another sample of units sold five years ago. The data are stored in columns 1 (sales prices five years ago) and 2 (prices this year).

- Use graphical techniques to describe the two sets of data.
- Discuss similarities and differences between the two sets of data.

4.59 XR04-59 Exercise 4.41 addressed the issue of whether ordinary shares are a good hedge against inflation. In order to investigate the same issue for long-term bonds, the annual inflation rates and the annual percentage rates of return on bonds from 1995 to 2018 are shown in the table below.

Year	Inflation rate (%)	Bond return (%)	Year	Inflation rate (%)	Bond return (%)
1995	4.6	9.2	2007	2.3	6.0
1996	2.6	8.2	2008	4.4	5.8
1997	0.3	7.0	2009	1.8	5.0
1998	0.9	5.5	2010	2.9	5.4
1999	1.5	6.0	2011	3.3	4.9
2000	4.5	6.3	2012	1.8	3.4
2001	4.4	5.6	2013	2.5	3.7
2002	3.0	5.8	2014	2.5	3.7
2003	2.7	5.4	2015	1.5	2.7
2004	2.3	5.6	2016	1.3	2.3
2005	2.7	5.3	2017	1.9	2.6
2006	3.6	5.6	2018	1.9	2.7

Source: Based on Australian Bureau of Statistics data and Federal Reserve Economic Data.

- Draw a scatter diagram for these data, with the bond returns on the vertical axis.
- Describe the relationship between bond returns and inflation rates from 1995 to 2018.
- Does it appear that bonds provide a good hedge against inflation?
- Draw a straight line that approximates reasonably well the relationship between bond return and inflation rate.

4.60 XR04-60 The IQs of children born prematurely were measured when the children were five years old. Use whichever graphical technique you deem appropriate and describe what the graph tells you.

4.61 XR04-61 The average monthly value of one Australian dollar measured in American dollars for the period January 2010 to May 2019 was recorded. Produce a graph that shows how the exchange rate has varied over this period, and describe what the graph tells you.

Source: FRED® and the FRED® logo are registered trademarks of the Federal Reserve Bank of St. Louis. Used with permission. FRED® chart provided courtesy of the Federal Reserve Bank of St. Louis. © 2020 Federal Reserve Bank of St. Louis. All rights reserved.

4.62 XR04-62 One hundred and twenty-five university students were asked how many books they borrowed from the library over the previous 12 months. Their replies are recorded. Use a suitable graphical method to depict the data. What does the graph tell you?

4.63 XR04-63 The rates of unemployment in New South Wales from 1997 to 2020 are listed in the following table.

Unemployment rate, NSW, 1997–2020

Year	Rate (%)	Year	Rate (%)
1997	7.8	2009	6.5
1998	7.2	2010	5.3
1999	6.1	2011	4.9
2000	5.4	2012	5.0
2001	5.7	2013	5.6
2002	6.0	2014	5.7
2003	5.9	2015	5.6
2004	5.4	2016	5.0
2005	5.1	2017	4.7
2006	5.1	2018	4.8
2007	4.8	2019	4.5
2008	4.8	2020	6.4

Source: Australian Bureau of Statistics, May 2020, *Labour Force Australia*, cat. no. 6202.0, ABS, Canberra.

- Draw a bar chart of these data with 4.0% as the lowest point on the vertical axis.
- Draw a bar chart of these data with 0% as the lowest point on the vertical axis.
- Discuss the impressions given by the two charts.
- Which chart would you use? Explain.

4.64 XR04-64 Monthly tourist arrivals to New Zealand from February 2010 to February 2020 were recorded. Draw an appropriate chart to show how tourist arrivals to New Zealand has changed during the 10 year period. Comment on the seasonality of the data.

Case Studies

CASE 4.1 The question of global warming

C04-01a, C04-01b In the last part of the twentieth century, scientists developed the theory that the planet was warming and that the primary cause was the increasing amount of atmospheric carbon dioxide (CO_2), which is the product of burning oil, natural gas and coal (fossil fuels). Although many climatologists believe in what is commonly referred to as the greenhouse effect, there are others who do not subscribe to this theory. There are three critical questions that need to be answered in order to resolve the issue.

- 1 Is Earth actually warming?** To answer this question, we need accurate temperature measurements over a large number of years. But how do we measure the temperature before the invention of accurate thermometers? Moreover, how do we go about measuring Earth's temperature even with accurate thermometers?
- 2 If the planet is warming, is there a human cause or is it natural fluctuation?** Earth's temperature has increased and decreased many times in its long history. We've had higher temperatures and we've had lower temperatures, including various ice ages. In fact, a period called the 'little ice age' ended around the middle to the end of the nineteenth century. Then the temperature rose until about 1940, at which point it decreased until 1975. In fact, a *Newsweek* article published 28 April 1975, discussed the possibility of global cooling, which seemed to be the consensus among scientists at the time.
- 3 If the planet is warming, is CO_2 the cause?** There are greenhouse gases in the atmosphere, without which Earth would be considerably colder. These gases include methane, water vapour and carbon dioxide. All occur naturally in nature. Carbon dioxide is vital to our life on Earth because it is necessary for growing plants. The amount of CO_2 produced by fossil fuels is a relatively small proportion of all the CO_2 in the atmosphere. The generally accepted procedure is to record monthly temperature anomalies. To do so, we calculate the average for each month over many years. We then calculate any deviations between the latest month's temperature reading and its average. A positive anomaly would represent a month's temperature that is above the average. A negative anomaly indicates a month where the temperature is less than the average. One key question is how we measure the temperature.

Although there are many different sources of data, we have chosen to provide you with one, the National Climatic Data Center (NCDC), which is affiliated with the National Oceanic and Atmospheric Administration (NOAA). (Other sources tend to agree with the NCDC's data.)

C04-01a stores the monthly temperature anomalies ($^{\circ}\text{C}$) from 1880 to 2016. The best measures of CO_2 levels (ppm) in the atmosphere come from the Mauna Loa Observatory in Hawaii, which has measured this variable since March 1958. These data together with the temperature anomalies for March 1958 to May 2016 are stored in file **C04-01b**. (Note that some of the original data are missing and have been replaced by interpolated values.)

- 1 Use whichever techniques you wish to determine whether there is such a thing as global warming.
- 2 Use a graphical technique to determine whether there is a relationship between temperature anomalies and CO_2 levels.

CASE 4.2 Analysing the spread of the global coronavirus pandemic

C04-02 The coronavirus (COVID-19) pandemic outbreak was first identified in Wuhan, China, in December 2019. By June 2020 it had spread to more than 200 countries, infected more than 10 million people and resulted in the death of more than 500 000 people globally. Daily data for the number of confirmed cases and deaths are recorded. Depict graphically the number of confirmed cases and number of deaths during the six-month period for the top 10 affected countries and globally.

Source: <https://ourworldindata.org/coronavirus-source-data>

CASE 4.3 An analysis of telephone bills

C04-03 A telephone company in Melbourne wanted to gather information concerning the monthly bills of new subscribers. A survey of 200 new Melbourne residential subscribers was undertaken, and their bills for the first month were recorded. What information can you extract from the data?

CASE 4.4 An analysis of monthly retail turnover in Australia

C04-04 Analyse the monthly turnover of Australian retail turnover by state using appropriate graphical descriptive methods. The data provide seasonally adjusted as well as unadjusted series for the period April 1988 to May 2019. Compare the two series for Australia as a whole and for the 8 states and territories.

Source: *Retail Turnover by State/Territory*, ABS cat. no. 8501.0, Retail Trade, Australian Bureau of Statistics, Australia, May 2019.

CASE 4.5 Economic freedom and prosperity

C04-05 Adam Smith published *The Wealth of Nations* in 1776. In that book he argued that when institutions protect the liberty of individuals, greater prosperity results for all. Since 1995, the *Wall Street Journal* and the Heritage Foundation, a think tank in Washington, DC, have produced the *Index of Economic Freedom* for all countries in the world. The index is based on a subjective score for 10 freedoms and country attributes: business freedom, trade freedom, government spending, monetary freedom, investment freedom, financial freedom, property rights, judicial effectiveness, government integrity and labour freedom.

The scores for 2020 for 182 countries are recorded. The gross domestic product (GDP) for 2020, measured in terms of purchasing power parity (PPP), which makes it possible to compare the GDP for all countries is also recorded. Use the 2020 freedom index scores, the GDP PPP figures and a graphical technique to see how freedom and prosperity are related. Also present a graphical analysis of the relationship between the individual freedom scores and the GDP PPP.

Numerical descriptive measures

Learning objectives

This chapter presents a number of numerical descriptive measures used to summarise and describe sets of data.

At the completion of this chapter, you should be able to:

- L01** calculate mean, median and mode, and explain the relationships between them
- L02** calculate the weighted average and geometric average
- L03** calculate range, variance, standard deviation and coefficient of variation
- L04** interpret the use of standard deviation through the empirical rule and Chebyshev's theorem
- L05** explain the concepts of percentiles, deciles, quartiles and interquartile range, and show their usefulness through the application of a box plot
- L06** calculate the mean and variance when the data are already in grouped form
- L07** obtain numerical measures to calculate the direction and strength of the linear relationship between two variables
- L08** understand the use of graphical methods and numerical measures to present summary information about a data set.

CHAPTER OUTLINE

Introduction

- 5.1** Measures of central location
- 5.2** Measures of variability
- 5.3** Measures of relative standing and box plots
- 5.4** Measures of association
- 5.5** General guidelines on the exploration of data

SPOTLIGHT ON STATISTICS

Income and its allocation to food

One of the most important empirical regularities in consumption economics is Engel's law. The law states that the proportion of income allocated to food (w_F) falls with increasing income. This means that a rich consumer would spend a lesser proportion of his/her income on food than a poor consumer, who would allocate a greater proportion of his/her income to food. For example, according to the World Bank, in 2017 the real per capita GDP, in 2010 constant US\$, of India was US\$1987 and that of Australia was US\$55919. It is interesting to note that Indians, on average, allocate about one-third (33%) of their income to food expenses, while Australians allocate only one-tenth (10%) of their income to food.



Source: iStock.com/thorbjorn66

Two statisticians, Working (in 1943) and Leser (in 1963), modelled Engel's law into a linear regression framework. In order to investigate this law, we need to develop the relationship between income and the proportion of income allocated to food. We have collected data for the year 2018 from 46 countries, which are stored in file **CH05\XM05-00**. After we have presented the statistical technique, we will return to this problem and solve it (see page 186).

Introduction

In Chapter 4, we presented several graphical techniques, such as frequency distributions and histograms, to summarise numerical data into a more manageable form. These techniques revealed the approximate shape of the distribution and indicated *where* the observations were concentrated. In this chapter we continue to examine how to summarise more precisely a large set of raw data using numerical descriptive techniques so that the meaningful essentials can be extracted from it. These techniques are also critical to the development of statistical inference.

As we pointed out in Chapter 2, arithmetic calculations can be applied to numerical (quantitative) data only. Consequently, most of the techniques introduced in this chapter may only be used to describe numerical data. However, some of the techniques can be used for ordinal (ranked) data, and one of the techniques can be used for nominal (categorical) data as well.

When we introduced the histogram, we noted that there are several pieces of information that we look for. The first is the location of the centre of the data. In Section 5.1 we present measures of central location. Another important characteristic that we seek from a histogram is the spread of the data. The spread will be measured more precisely by measures of variability, which we present in Section 5.2. Section 5.3 introduces measures of relative standing and another graphical technique, the box plot. [In the Appendix A5.2, we present descriptive measures to calculate the mean and variance when the data are grouped in a frequency table.]

In Section 4.3 we introduced the scatter diagram, a graphical method used to analyse the relationship between two numerical variables. The numerical counterparts to the scatter diagram are called *measures of association*, and they are presented in Section 5.4. Finally, in Section 5.5, we complete this chapter by providing guidelines on how to explore data and retrieve information.

Recall the following terms, introduced in Chapter 1: *population*, *sample*, *parameter* and *statistic*. A parameter is a descriptive measurement about a population, and a statistic is a descriptive measurement about a sample. In this chapter we introduce several descriptive measurements. For each we will describe how to calculate both the population parameter and the sample statistic. However, in most realistic applications, populations are very large, in fact, virtually infinite. The formulas describing the calculation of parameters are not practical and are seldom used. They are provided here primarily to teach the concept and the notation. In Chapter 7 we will introduce probability distributions, which describe populations. At that time we will show how parameters are calculated from probability distributions. In general, small data sets of the type we feature in this book are samples.

5.1 Measures of central location

There are three different measures of central location that we use to describe the centre of a set of data. They are the *arithmetic mean*, the *median* and the *mode*.

5.1a Arithmetic mean

By far the most popular and useful measure of central location is the **arithmetic mean**, which we will simply refer to as the **mean**. Widely known in everyday usage as the *average*, the mean of a set of observations is defined as:

$$\text{Mean} = \frac{\text{Sum of observations}}{\text{Number of observations}}$$

If we are dealing with a population of observations, the total number of observations is denoted by N and the mean is represented by μ (the lowercase Greek letter *mu*). If the set of observations is a sample, the total number of observations is denoted by n , and the **sample mean** is represented by \bar{x} (referred to as *x-bar*).

We label the observations in a population, x_1, x_2, \dots, x_N , where x_1 is the first observation, x_2 is the second observation and so on, until x_N , with N being the population size. Similarly, for a sample, we label the observations as x_1, x_2, \dots, x_n , where n is the sample size.

mean (arithmetic mean)

The sum of a set of observations divided by the number of observations.

sample mean

The arithmetic mean of sample data.

Population mean

The mean of a population of N observations x_1, x_2, \dots, x_N is defined as:

$$\mu = \frac{\text{Sum of observations in the population}}{\text{Number of observations in the population}} = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{\sum_{i=1}^N x_i}{N}$$

Sample mean

The mean of a sample of n observations x_1, x_2, \dots, x_n is defined as:

$$\bar{x} = \frac{\text{Sum of observations in the sample}}{\text{Number of observations in the sample}} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

EXAMPLE 5.1

LO1

Average weekly study time

XM05-01 A random sample of 10 first year students at a university in Brisbane were asked how many hours they spent studying for the business data analysis course during a particular week of the survey. The results (in hours) are as follows:

15 9 13 14 13 7 20 8 12 9

Calculate the sample mean weekly study time for the course by the 10 students.

Solution

Calculating manually

Using our notation, we have $x_1 = 15$, $x_2 = 9$, ..., $x_9 = 12$, $x_{10} = 9$, and $n = 10$.

To calculate the mean, we add all the observations and divide the sum by the size of the sample. Thus, the sample mean is

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{15 + 9 + 13 + 14 + 13 + 7 + 20 + 8 + 12 + 9}{10} = \frac{120}{10} = 12 \text{ hours}$$

Interpreting the results

On average, the first-year students at the university spent 12 hours studying for the business data analysis course during the week of the survey.



Using the computer

There are several ways to calculate the mean using Excel. If we simply want to calculate the mean and no other statistics, we can proceed as follows.

COMMANDS

- 1 Type in the data (Type **Study time (hours)** in cell **A1**. Then type **15, 9, 13, 14, 13, 7, 20, 8, 12, 9** in a column (**A2:A11**) or open the data file (**XM05-01**)).
- 2 Click on any empty cell (**A12**).
- 3 Click **fx**, select **Statistical** from the **Category** drop-down menu, select the function **AVERAGE** and click **OK**.
- 4 In the **Number1** box, type the input range of the data (**A2:A11**) and click **OK**. The mean will appear in the active cell.

Alternatively, type the following command in any empty active cell, **=AVERAGE([Input range])**. For example, **=AVERAGE(A2:A11)**. The active cell will show the mean as 12.

The mean is a popular measure because it is simple to calculate and interpret, and it lends itself to mathematical manipulation. More importantly for decision-makers, it is generally the best measure of central location for purposes of statistical inference. One serious drawback is that it is unduly influenced by extreme observations, also known as outliers. For example, if the sample of 10 observations in Example 5.1 is enlarged to include an eleventh observation that has a value of 89 hours, the mean of the resulting sample of eleven observations is $209/11 = 19$ hours. Adding a single relatively large value to the original sample of observations substantially increases the value of the mean from 12 to 19 hours, making this statistic a poor measure of central location of the given data set. This is one reason why we sometimes resort to another measure of central location, the *median*.

5.1b Median

median

The middle value of a set of observations when they are arranged in order of magnitude.

The second most popular measure of central location is the **median**. The median of a set of data is the middle value of the arranged values (in ascending or descending order) of that data set. The population median and the sample median are both computed in the same way.

When there is an even number of observations, there will be two middle values. In this case, the median is determined by averaging the two middle values of the data arranged in ascending or descending order.

The median has intuitive appeal as a measure of central location: at most, half the observations fall below the median, and at most, half fall above. Because of the distorting effect of extreme observations on the mean, the median is often the preferred measure in such situations as salary negotiations.

EXAMPLE 5.2

L01

Median salary of workers in a travel agency

XM05-02 The annual salaries (in \$'000) of the seven junior office workers in a travel agency are as follows:

64 67 62 68 66 62 65

Find the median salary of the employees.





Solution

Calculating manually

When the seven salaries are ordered in ascending order, the data appear as follows:

62	62	64	65	66	67	68
----	----	----	----	----	----	----

The median salary is, therefore, the middle (fourth) observation, \$65 000.

The mean value of the observations is \$64 857. In this example, the mean and the median salaries are similar and both represent the centre of the data.

Interpreting the results

Half of the salaries are below \$65 000 and half of the salaries are above \$65 000.

Using the computer

COMMANDS

To calculate the median, we follow the commands used in Example 5.1 but substitute **MEDIAN** in place of **AVERAGE**.
The active cell should show the median as 65.

EXAMPLE 5.3

L01

Median salary of workers in a travel agency – continued

XM05-03 Consider Example 5.2. Suppose now that the data include the director's salary of \$145 000. That is, the annual salaries (in \$'000) of all employees in the travel agency, including the director, are as follows:

64	67	62	68	66	62	65	145
----	----	----	----	----	----	----	-----

Find the median salary of all employees.

Solution

Calculating manually

When the eight salaries are ordered in ascending order, the data appear as follows:

62	62	64	65	66	67	68	145
----	----	----	----	----	----	----	-----

As there is an even number of observations, the median is the average of the two middle observations, 65 and 66. That is, the median is:

$$\text{Median} = \frac{65 + 66}{2} = 65.5$$

The median salary of the employees is \$65 500.

This median salary of \$65 500 is similar to the median salary of \$65 000 in Example 5.2, even though we now have the outlier salary of the director included. That is, adding an extreme value (of \$145 000) did not have much effect on the median. However, the mean salary of the employees is now \$74 875. This value is much higher than the mean salary (\$64 857) in Example 5.2. Therefore, when there is an extreme observation in the data, the median value is the preferred central measure of a typical salary, rather than the mean value.

The median is the most appropriate measure of central location to use when the data under consideration are ordinal (ranked), rather than numerical. Such a situation arises whenever items are simply ranked, such as according to preference, degree of ability or degree of difficulty. For example, if 11 statistical problems are ranked from 1 to 11 according to their degree of difficulty, statistical problem 6 is the problem of median difficulty.

5.1c Mode

mode

The most frequently occurring value in a set of data.

A third measure of central location is the **mode**. The population mode and the sample mode are both computed in the same way.

When data have been organised into a histogram, we are often interested in knowing which class has the largest number of observations. We refer to that class as the *modal class*. In Chapter 4, we discussed modal classes in the context of describing the shape of a histogram.

The mode is useful when the level of demand for an item is of interest – perhaps for the purpose of making purchasing or production decisions – and the item is produced in various standard sizes. But when the number of possible data values is quite large, the mode ceases to be useful as a measure of central location. Sometimes no single value occurs more than once, making all the observations modes and providing no useful information. This, in fact, is the situation with the electricity bills data in Table 4.1 (page 87). In such a case, it is more useful to group the data into classes and refer to the class with the largest frequency as the *modal class*. A distribution is then said to be *unimodal* if there is only one such class, and *bimodal* if there are two such classes. Although the *midpoint of the modal class* is sometimes referred to as the mode, it does not identify the observation that occurs most frequently (as the true mode does) but the observation about which there is the greatest clustering of values. Thus, it corresponds graphically to the highest point on the frequency polygon.

EXAMPLE 5.4

LO1

Mode weight of men at a gym

XM05-04 The manager of a gym in Melbourne observed the weight (in kilograms) of a sample of 14 male adult patrons:

85 80 77 96 85 68 77 85 94 85 82 76 80 76

Find the mode weight of the male adult patrons.

Solution

Calculating manually

The frequency table for the above data is shown.

x_i	68	76	77	80	82	85	94	96
f_i	1	2	2	2	1	4	1	1

The mode of these weights is 85 kg (which occurs with the highest frequency of four). This fact is of interest to the manager, together with the facts that the mean weight is 81.9 kg and the median weight is 81 kg.

Interpreting the results

The mode weight of male adult gym patrons is 85 kg. That is, the most frequent weight of male adult patrons of the gym is 85 kg. This mode weight is slightly greater than the mean or median weights.

Using the computer

COMMANDS

To calculate the mode, we follow the commands used in Example 5.1 but substitute **MODE** in place of **AVERAGE**. The active cell should show the mode as 85.

Note that if there is more than one mode, Excel shows only one of them, without indicating that there are other modes. If there are no values occurring more than once, then Excel will print a #N/A.

We now look at another example for which we will calculate all three measures of central location – first manually and then using the computer.

EXAMPLE 5.5

L01

Weight of children in their first year of school

XM05-05 The weights (in whole kilogram) of a sample of 10 children in a Year 1 class at an Adelaide school were recorded as follows:

19 14 25 16 20 15 18 25 16 22

Calculate the three measures of central location.

Solution**Calculating manually**

The mean is calculated as:

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{n} = \frac{19 + 14 + 25 + 16 + 20 + 15 + 18 + 25 + 16 + 22}{10} = \frac{190}{10} = 19 \text{ kg}$$

The median is determined by placing the observations in ascending order as follows:

14 15 16 16 18 19 20 22 25 25

Because there is an even number of observations, the median is the average of the two middle numbers, which are 18 and 19. Thus, the median is 18.5 kg.

The mode is the observation that occurs most frequently. The observations 16 and 25 both occur twice. All other observations occur only once. Consequently, there are two modes: 16 kg and 25 kg.

Interpreting the results

The mean and median weight of a sample of 10 Year 1 students are 19 kg and 18.5 kg, respectively. The mode weights are 16 kg and 25 kg.

Using the computer

Excel can produce the measures of central location and a number of other summary statistics (which we will introduce in later sections) using the **Data Analysis** tool. The Excel commands to do this are given below.

Excel output for Example 5.5

	C	D
1	<i>Weight (kg)</i>	
2		
3	Mean	19
4	Standard Error	1.256
5	Median	18.5
6	Mode	25
7	Standard Deviation	3.972
8	Sample Variance	15.778
9	Kurtosis	-1.080
10	Skewness	0.479
11	Range	11
12	Minimum	14
13	Maximum	25
14	Sum	190
15	Count	10

Note that Excel recognises only one mode: 25. Obviously, this is not correct.



COMMANDS

- 1 Type the data in a column or open the data file (**XM05-05**). Type **Weight (kg)** in cell A1. Then type **19, 14, 25, 16, 20, 15, 18, 25, 16, 22** in a column (**A2:A11**).
- 2 Click **DATA** and then under the **Analysis** submenu, click **Data Analysis** and in the drop-down menu select **Descriptive Statistics**. Click **OK**.
- 3 Type in the **input range** (include the cell containing the variable name.) Tick **Labels in First Row**. (**A1:A11**)
- 4 To store the output on the same sheet, under **Output options** click **Output Range:** and type a starting cell reference for the output (**C1**). Tick the box for **Summary Statistics**. Click **OK**.

EXAMPLE 5.6

L01

Summary central measures of students' exam marks

XM05-06 A statistics lecturer wants to report the results of a mid-semester exam taken by his class of 100 students. The marks are shown in **Table 5.1**. Find the mean, median and mode of these data and describe what information they provide.

TABLE 5.1 Mid-semester exam marks for class of 100 students

94	88	65	73	84	83	95	38	97	72	83	87	94	93	73	78	86	81	92	30	79	90	69	96	39
84	90	8	59	74	94	90	95	70	81	91	75	82	83	65	34	89	57	98	93	83	99	42	99	51
95	90	84	48	81	96	91	96	83	41	100	25	48	71	89	61	77	43	85	75	64	64	93	86	84
94	80	55	84	66	34	98	72	11	38	85	77	96	50	71	37	16	76	18	73	99	85	53	69	66

Solution

Using the computer

We will use Excel to calculate the three measures of central location.

Excel output for Example 5.6

	A	B
15	Marks	
16		
17	Mean	73.18
18	Standard Error	2.244
19	Median	81
20	Mode	84
21	Standard Deviation	22.443
22	Sample Variance	503.684
23	Kurtosis	0.474
24	Skewness	-1.099
25	Range	92
26	Minimum	8
27	Maximum	100
28	Sum	7318
29	Count	100





COMMANDS

The Excel commands are the same as those for Example 5.5, with data in file **XM05-06**, input data range **A1:A101** and output starting cell reference **C1**. The mean and median are 73.18 and 81.0, respectively. As was the case in Example 5.5, Excel identifies only one mode, 84. In fact there are two, 83 and 84, each of which occurs five times. The rest of the output will be discussed later.

Interpreting the results

Most students want to know the mean, which is generally interpreted as measuring the ‘average’ student’s performance. The median tells us that half the class received a mark greater than 81% and the other half had marks below 81%. Which is the better measure? The answer depends on what we want to measure.

If we want an overall measure of how well the class performed, the mean should be used. Because the mean is calculated by adding all the marks and dividing by the number of students, the mean provides a number based on the total marks of the class and thus provides a measure of the class performance. However, if we want to know the standing of a student’s mark relative to the middle mark, the median gives us a mark that truly represents the centre of the data. The marks of half the class were above the median and those of half the class were below it.

If the marks are classified by letter grade, where A = 80–100, B = 70–79, C = 60–69, D = 50–59, and F = 0–49, we can count the frequency of each grade. Because we are now interested in the number of students receiving each type of grade, the mode becomes a logical measure to calculate. As we pointed out in Chapter 4, we designate the category or class with the largest number of observations as the modal class. As you can see from **Table 5.2**, the modal class is A.

TABLE 5.2 Marks converted to letter grades

Grade	Frequency
A	53
B	16
C	9
D	6
F	16

5.1d Mean, median, mode: Which is best?

With three measures to choose from, which one should we use? There are several factors to consider when making our choice of measure of central location. The mean is generally our first choice. However, there are several circumstances under which the median is more useful. The mode is seldom the best measure of central location. One advantage of the median is that it is not as sensitive as the mean to extreme values. (See discussions below Example 5.1.)

Consider the data in Example 5.2. The mean and median salaries are \$64 857 and \$65 000, respectively. When an eighth (extreme) observation of \$145 000 is included in Example 5.3, the mean has increased from \$64 857 to \$74 875, but the median has stayed about the same. When there is a relatively small number of extreme observations (either very small or very large, but not both), the median usually produces a better measure of the centre of the data.

To illustrate another advantage of the median over the mean, suppose you and your classmates have taken a statistics test and the lecturer is returning the graded tests. What piece of information is most important to you? The answer, of course, is *your* mark. What is the next most important piece of information? The answer is how well you performed relative to the class. Most students ask their instructor for the class mean. This is the wrong statistic to request. It is the *median* you should ask for, because it divides the class into halves.

This information allows you to identify into which half your mark falls. The median provides this information; the mean does not. Nevertheless, the mean can also be useful in this scenario. If there are several classes taking the course, the class means can be compared to determine which class performed best (or worst).

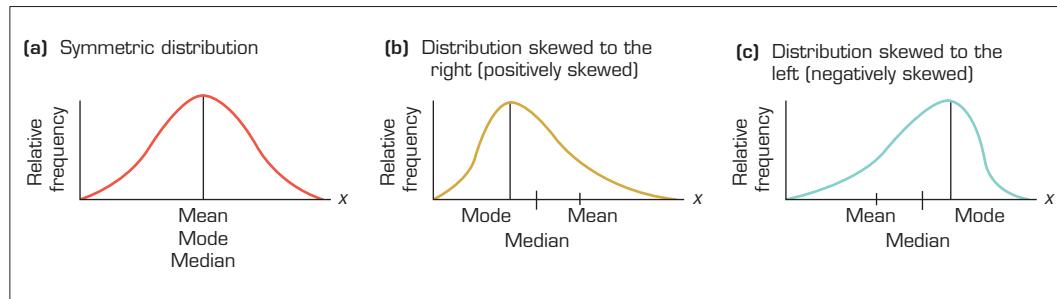
Because the mean is the best measure of central location for the purpose of statistical inference, it will be used extensively from Chapter 10 onwards. But for descriptive purposes, it is usually best to report the values of all three measures, since each conveys somewhat different information. Moreover, the relative positions of the mean and the median provide some information about the shape of the distribution of the observations.

5.1e Relationship between mean, median and mode

The relationship between the three measures of central location can be observed from the smoothed relative frequency polygons in **Figure 5.1**. If the distribution is symmetrical and unimodal, the three measures coincide, as shown in **Figure 5.1(a)**. If a distribution is not symmetric, it is called a *skewed distribution*. The distribution in **Figure 5.1(b)** is **skewed to the right**, or positively skewed, since it has a long tail extending off to the right (indicating the presence of a small proportion of relatively large, extreme values) but only a short tail extending to the left. Distributions of incomes commonly exhibit such positive **skewness**. As mentioned earlier, these extreme values pull the mean to the right more than they pull the median. A mean value greater than the median therefore provides some evidence of positive skewness.

The distribution in **Figure 5.1(c)** is **skewed to the left**, or negatively skewed, since it has a long tail to the left but a short tail to the right. The distribution of age at death exhibits such negative skewness. Once again, the extreme values affect the mean more than they do the median, so the mean value is pulled more noticeably in the direction of the skew. A mean value less than the median is an indication of negative skew.

FIGURE 5.1 Relationships between mean, median and mode



5.1f Measures of central location for ordinal and nominal data

When the data are numerical, we can use any of the three measures of central location. However, for ordinal and nominal data the calculation of the mean is not valid. Because the calculation of the median begins with placing the data in order, this statistic is appropriate for ordinal data. The mode, which is determined by counting the frequency of each observation, is appropriate for nominal data. However, nominal data do not have a 'centre', so we cannot interpret the mode of nominal data in that way.

5.1g Other types of measures of central location (Optional)

Apart from these three popular measures of central location, there are other measures that are used for specific purposes. In this section we consider two of these measures: weighted mean and geometric mean.

Weighted mean

When we calculate an arithmetic mean, we give equal weight to each data point. The **weighted mean** or weighted arithmetic mean is similar to an arithmetic mean, except that some data points contribute more ‘weight’ than others. A weighted average assigns weights that determine the relative importance of each data point. If all the weights are equal, then the weighted mean equals the arithmetic mean.

weighted mean

The sum of a set of observations multiplied by their corresponding weights.

The weighted mean of a set of observations is defined as:

$$\text{Weighted mean} = \text{Sum of (weight * corresponding observation)}$$

We label the observations in a sample, x_1, x_2, \dots, x_n , with the corresponding weights, w_1, w_2, \dots, w_n , where w_i is the weight for observation x_i . The sum of all the weights should equal 1, with n being the sample size.

Weighted mean

The weighted mean of a population of N observations x_1, x_2, \dots, x_N , with corresponding weights w_1, w_2, \dots, w_N , is defined as:

$$\mu_w = w_1x_1 + w_2x_2 + \dots + w_Nx_N = \sum_{i=1}^N w_i x_i \text{ where } \sum_{i=1}^N w_i = 1$$

The weighted mean of a sample of n observations x_1, x_2, \dots, x_n , with corresponding weights w_1, w_2, \dots, w_n , is defined as:

$$\bar{x}_w = w_1x_1 + w_2x_2 + \dots + w_nx_n = \sum_{i=1}^n w_i x_i \text{ where } \sum_{i=1}^n w_i = 1$$

EXAMPLE 5.7

LO2

Returns from a portfolio with unequally weighted shares

XMO5-07 Consider 4 annual returns from a share portfolio, 14, 8, 11 and 4 per cent, with weights 0.2, 0.3, 0.4 and 0.1, respectively. Calculate the weighted mean of the returns.

Solution

Calculating manually

Using our notation, we have $x_1 = 14$, $x_2 = 8$, $x_3 = 11$ and $x_4 = 4$; $w_1 = 0.2$, $w_2 = 0.3$, $w_3 = 0.4$ and $w_4 = 0.1$; $n = 4$.

The sample weighted mean of the returns is

$$\begin{aligned}\bar{x}_w &= \sum_{i=1}^4 w_i x_i = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 \\ &= (0.2)(14) + (0.3)(8) + (0.4)(11) + (0.1)(4) \\ &= 2.8 + 2.4 + 4.4 + 0.4 = 10.0\end{aligned}$$

The arithmetic mean of the four returns is $(14 + 8 + 11 + 4)/4 = 9.25\%$, and if we give different levels of importance to the observations the weighted mean is 10.0%.





Interpreting the results

When the four returns contribute unequally in the portfolio, then the centre of the four returns is 10%. However, if the four returns contribute equally in the portfolio, then the centre of the returns is 9.25%.

Using the computer

To calculate the weighted mean using Excel, we can proceed as follows.

COMMANDS

- 1 Type the data into one or more columns of a new Excel spreadsheet or open the data file. Type the titles **Return** and **Weight** in cells **A1** and **B1**. Type the values and the corresponding weights in two adjacent columns (**A2:A5 and B2:B5**) or open Excel file (**XM05-07**).
- 2 Click on any empty cell (e.g. **A6**).
- 3 Click inset function **fx**, select **Math & Trig** from the category drop-down menu, then select the function **SUMPRODUCT** and click **OK**.
- 4 In the **Array1** box, type the input range of the returns (**A2:A5**) and, in **Array2** box type the range of the weights (**B2:B5**) and click **OK**. The weighted mean will appear in the active cell.

Alternatively, type the following command in any empty active cell, **=SUMPRODUCT([Array1, Array2])**; for example, **=SUMPRODUCT(A2:A5,B2:B5)**. The active cell will show the weighted mean as 10.

Geometric mean

The arithmetic mean is the single most popular and useful measure of central location. We noted certain situations for which the median is a better measure of central location. However, there is another circumstance where neither the mean nor the median is the best measure. When the variable is a growth rate or rate of change, such as the value of an investment over periods of time, we need another measure.

The geometric mean is another measure of central location. It is calculated as the n th root of the product of a given number of observations. The population geometric mean is denoted by μ_g and the sample geometric mean by \bar{x}_g .

Geometric mean

The geometric mean of a population of N observations x_1, x_2, \dots, x_N is defined as:

$$\text{Population geometric mean: } \mu_g = \sqrt[N]{x_1 x_2 \dots x_N} = (x_1 x_2 \dots x_N)^{1/N}$$

The geometric mean of a sample of n observations x_1, x_2, \dots, x_n is defined as:

$$\text{Sample geometric mean: } \bar{x}_g = \sqrt[n]{x_1 x_2 \dots x_n} = (x_1 x_2 \dots x_n)^{1/n}$$

geometric mean

The n th root of the product of all observations.

Geometric mean rate of return

One of the major applications in business of geometric mean is in rate of return of investment. Using the geometric mean, we can calculate the average compounding rate of return on investment over time.

For example, consider an investment of \$100 at a rate of return r_1 in the first year, rate of return r_2 in the second year, ..., and at a rate of return r_n in the n th year.

- 1 At the end of year 1, the value of investment $A_1 = 100(1 + r_1)$
- 2 At the end of year 2, the value of investment $A_2 = A_1(1 + r_2) = 100(1 + r_1)(1 + r_2)$
- 3 At the end of year 3, the value of investment $A_3 = A_2(1 + r_3) = 100(1 + r_1)(1 + r_2)(1 + r_3)$
...
- 4 At the end of the n th year, the value of investment, $A_n = 100(1 + r_1)(1 + r_2)(1 + r_3)\dots(1 + r_n)$

If the *average compounding rate of return* over the n years is r_g , then A_n can also be calculated as

$$A_n = 100(1 + r_g)^n$$

Therefore:

$$100(1 + r_g)^n = 100(1 + r_1)(1 + r_2)(1 + r_3)\dots(1 + r_n)$$

Simplifying the equation we get:

$$(1 + r_g)^n = (1 + r_1)(1 + r_2)(1 + r_3)\dots(1 + r_n)$$

Therefore, the average compounding rate of return is

$$r_g = [(1 + r_1)(1 + r_2)(1 + r_3)\dots(1 + r_n)]^{(1/n)} - 1$$

Average compounding rate of return

The average compounding rate of return over n years for an investment with rates of return for the 1st, 2nd ..., n th year, r_1, r_2, \dots, r_n is defined as:

$$r_g = [(1 + r_1)(1 + r_2)\dots(1 + r_n)]^{(1/n)} - 1$$

EXAMPLE 5.8

LO2

Rate of growth of a 2-year investment

XM05-08 Suppose you make a 2-year investment of \$1000, and it grows to \$2000 during the first year. During the second year, however, the investment suffers a loss and becomes \$1000. Calculate the average growth rate of return using (a) arithmetic mean and (b) geometric mean. Comment on both means calculated above.

Solution

Calculating manually

The investment of \$1000 is for 2 years. Let r_1 and r_2 be the rates of return in year 1 and year 2, respectively.

At the end of year 1, the value of investment $A_1 = 1000(1 + r_1) = 2000$.

Therefore, we can find r_1 as

$$r_1 = \frac{2000}{1000} - 1 = 1 \text{ or } 100\%$$

At the end of year 2 the value of investment

$$A_2 = A_1(1 + r_2) = 2000(1 + r_2) = 1000$$

Therefore, we can find r_2 as

$$r_2 = \frac{1000}{2000} - 1 = -0.5 \text{ or } -50\%$$





The rates of return for years 1 and 2 are $r_1 = 100\%$ and $r_2 = -50\%$, respectively.

It is worth noting that since the original investment of \$1000 became \$1000 again at the end of year 2, the average rate of return should be 0%.

(a) Using the arithmetic mean:

$$\text{Average rate of return} = (1/2)[1 + (-0.5)] = 0.25 \text{ or } 25\%.$$

(b) Using the geometric mean:

Average compounding rate of return

$$r_g = [(1 + r_1)(1 + r_2)]^{(1/2)} - 1 = [(1 + 1)(1 + (-0.5))]^{(1/2)} - 1 = 0 \text{ or } 0\%$$

Interpreting the results

The average rate of return based on the arithmetic mean is 25% and that based on the geometric mean is 0%. As can be seen, the rate of return based on the arithmetic mean calculation is obviously misleading. Because there was no change in the value of the investment from the beginning to the end of the 2-year period, the 'average' compounding rate of return is 0%. As you can see, this is the value of the geometric mean, which correctly reflects the reality.

Using the computer

To calculate the geometric mean using Excel, we follow the commands used in Example 5.1 but substitute GEOMEAN in place of AVERAGE. We can proceed as follows.

COMMANDS

- 1 Type the values of $1 + R_i$ in a column of a new Excel spreadsheet or open the data file. (Type **1+Rate** as column title in cell **A1**. Type **2.0, 0.5** in a column (**A2:A3**) or open file (**XM05-08**)).
- 2 Click on any empty cell (**C1**).
- 3 Click insert function **fx**, select **Statistical** from the **Category** drop-down menu, select the function **GEOMEAN** and click **OK**.
- 4 In the **Number1** box, type the input range of the data (**A2:A3**) and click **OK**, then in the same formula bar at the end of the formula, type **-1** and click **OK**. The geometric mean will appear in the active cell.

Alternatively, type the following command in any empty active cell, **=GEOMEAN([Input range])-1**. For example, **=GEOMEAN(A2:A3)-1**. The active cell will show the mean as 0.

EXAMPLE 5.9

L01

Rate of growth of an investment

XM05-09 The annual rate of return from an investment over the last 5 years is 10%, 20%, -20%, 20% and 5%. Find the average compounding annual rate of return.

Solution

Calculating manually

The investment is for 5 years. Let r_1, r_2, r_3, r_4 and r_5 be the rates of return in years 1–5. Therefore, $r_1 = 0.1$, $r_2 = 0.2$, $r_3 = -0.2$, $r_4 = 0.2$ and $r_5 = 0.05$. The appropriate central measure for the average annual compounding rate of return in this situation is calculated using the geometric mean rate of return.



Average compounding annual rate of return

$$r_g = [(1 + r_1)(1 + r_2)(1 + r_3)(1 + r_4)(1 + r_5)]^{(1/5)} - 1$$

$$= [(1 + 0.1)(1 + 0.2)(1 + (-0.2))(1 + 0.2)(1 + 0.05)]^{(1/5)} - 1 = 0.0588 \text{ or } 5.88\%$$

Note that we can also calculate the value of the investment after 5 years. For example, if the initial investment was for \$1000, the value of the investment after 5 years, A_5 , can be calculated as

$$A_5 = 1000(1 + r_g)^5 = 1000(1 + 0.0588)^5 = \$1330.56$$

Using the computer

The Excel commands are the same as those for Example 5.8, with data in file **XM05-09**, input data range **A2:A6** and output starting cell reference **C1**. The active cell will show the average compounding growth rate of the investment over the 5 years as 0.0588.

Here is a summary of the numerical techniques introduced in this section and when to use them.

IN SUMMARY

Factors that identify when to calculate the mean

- 1** *Objective:* to describe a single set of data
- 2** *Type of data:* numerical (quantitative)
- 3** *Descriptive measurement:* central location

Factors that identify when to calculate the median

- 1** *Objective:* to describe a single set of data
- 2** *Type of data:* ordinal or numerical (with extreme observations)
- 3** *Descriptive measurement:* central location

Factors that identify when to calculate the mode

- 1** *Objective:* to describe a single set of data
- 2** *Type of data:* numerical, ordinal or nominal
- 3** *Descriptive measurement:* central location

Factors that identify when to calculate the weighted mean

- 1** *Objective:* to describe a single set of data (with weights)
- 2** *Type of data:* numerical (with weights sum to 1)
- 3** *Descriptive measurement:* central location

Factors that identify when to calculate the geometric mean

- 1** *Objective:* to describe a single set of data
- 2** *Type of data:* numerical, growth rates
- 3** *Descriptive measurement:* central location

EXERCISES

Learning the techniques

- 5.1 XR05-01** The following data represent the number of defective products found by a quality control inspector at a roof-tile manufacturing company, during each eight-hour shift of each day for a week:

	Mon	Tues	Wed	Thurs	Fri
Shift 1	24	17	35	15	19
Shift 2	21	13	15	20	18

Find the mean number of defective products found per shift.

- 5.2 XR05-02** A sample of 12 students was asked how much cash they had in their wallets. Their responses (in \$A) are listed in the table.

60	30	20	5	110	50
65	35	40	85	45	10

Determine the mean, the median and the mode for these data.

- 5.3 XR05-03** A random sample of 12 joggers was asked to report the number of kilometres they ran last week. Their responses are listed in the table.

5.5	7.2	1.6	22.0	8.7	2.8
5.3	3.4	12.5	18.6	8.3	6.6

- a Calculate the three statistics that measure the central location.
- b Briefly describe what each statistic tells you.

- 5.4 XR05-04** The prices (in \$'000) of 10 randomly selected houses listed for sale in Robertson, Sunnybank and Sunnybank Hills areas in Queensland during March 2019 are listed below:

97	95	91	270	95	116	97	108	99	97
----	----	----	-----	----	-----	----	-----	----	----

- a Calculate the three statistics that measure the central location.
- b Briefly describe what each statistic tells you.

- 5.5 XR05-05** The lecturers at a university are required to submit their final examination paper to the exams and timetabling office 10 days before the end of teaching for that semester. The exam coordinator sampled 20 lecturers and recorded the number of days before the final exam when each submitted his or her exam paper. The results are given in the table.

2	14	4	13	6	12	11	6	7	7
8	10	12	12	6	13	6	14	5	2

- a Calculate the mean, the median and the mode.
- b Briefly describe what each statistic in part (a) tells you.

- 5.6 XR05-06** Consider a sample of five observations 4, 8, 9, 11 and 12 with weights 0.1, 0.2, 0.3, 0.3 and 0.1 respectively. Calculate the mean and the weighted mean of the data.

- 5.7 XR05-07** Consider a sample of eight observations 44, 58, 49, 55, 50, 40, 45 and 60 with weights 0.1, 0.15, 0.1, 0.1, 0.25, 0.15, 0.05 and 0.1 respectively. Calculate the mean and the weighted mean of the data.

- 5.8 XR05-08** Compute the average compounding rate of return for the following rates of return.

0.25	-0.10	0.50
------	-------	------

- 5.9 XR05-09** The following returns were realised on an investment over a 5-year period.

Year	1	2	3	4	5
Rate of return	0.10	0.22	0.06	-0.05	0.20

- a Calculate the mean and median of the returns.
- b Calculate the average compounding rate of return.
- c Which one of the three statistics computed in parts (a) and (b) best describes the return over the 5-year period? Explain.

- 5.10 XR05-10** An investment you made 5 years ago has realised the following rates of return.

Year	1	2	3	4	5
Rate of return	-0.15	-0.20	0.15	-0.08	0.50

- a Calculate the mean and median of the rates of return.
- b Calculate the average compounding rate of return.
- c Which one of the three statistics computed in parts (a) and (b) best describes the return over the 5-year period? Explain.

Applying the techniques

- 5.11 XR05-11 Self-correcting exercise.** The ages of the seven employees of a suburban pizza shop are as follows:

22	21	67	22	23	20	21
----	----	----	----	----	----	----

- a Calculate the mean, the median and the mode of the employees' ages.
- b How would these three measures of central location be affected if the oldest employee retired?

- 5.12 XR05-12** The weights of five athletes waiting for a flight at Auckland airport to travel to a world sporting event are given.

70	74	72	71	73
----	----	----	----	----

- a Calculate the mean and the median weight of the five athletes.
- b Later two athletes from the weightlifting team, weighing 98kg and 116kg, joined the group of five athletes. Calculate the mean and the median weight of the seven athletes.
- c Compare the means and medians obtained in parts (a) and (b).

- 5.13 XR05-13** Twenty families in a New South Wales country town were asked how many cars they owned. Their responses are summarised in the following table.

Number of cars	0	1	2	3	4
Number of families	3	10	4	2	1

Determine the mean, the median and the mode of the number of cars owned per family.

- 5.14 XR05-14** The compensation received by the highest-paid executives of 12 international companies (in '\$000) were as follows:

820	2215	803	850	856	801
899	846	902	856	911	947

- a Find the mean and the median compensation of the executives.
- b Comment on the skewness of the distribution of the executives' compensations.
- c Repeat part (a), ignoring the highest compensation. Is the mean or the median more affected by dropping the highest compensation?

- 5.15 XR05-15** Data for the share price of nine investments in Australia on 1 April 2019 are listed below.

Company	Price (\$)
Caltex Australia	26.21
ANZ Bank	26.03
Westpac	25.92
Ansell Ltd	25.42
NAB	25.27
JB Hi-Fi Ltd	24.95
AGL Energy Ltd	21.77
Telstra Corp	3.32
AMP Ltd	2.10

- a Calculate the mean and median share price of the nine companies and compare the values.
- b Repeat part (a) after eliminating AMP Ltd and Telstra Corp from the data. How are the mean and the median affected by this change?

- 5.16 XR05-16** The following table shows the value of market cap (\$billion) of 10 companies listed on the Australian Securities Exchange (as at closing of trade, 31 March 2019).

- a Calculate the mean market cap value of the top 10 companies.
- b Calculate the median market cap value.
- c Repeat parts (a) and (b) after eliminating the market cap for AGL Ltd. Compare the results.

Company	Market cap (\$billion)
Commonwealth Bank	125.1
BHP Billiton Ltd	113.4
Westpac Bank	89.4
CSL Ltd	88.3
ANZ Bank	73.7
National Australia Bank	71.0
Woolworths Ltd	40.1
Telstra Corp	39.5
Wesfarmers	39.3
AGL Ltd	14.3

- 5.17 XR05-17** Sporting competitions that use judges' scores to determine a competitor's performance often drop the lowest and the highest scores before calculating the mean score, in order to diminish the effect of extreme values on the mean. Competitors A and B receive the following scores.

A	6.0	7.0	7.25	7.25	7.5	7.5	7.5
B	7.0	7.0	7.0	7.25	7.5	7.5	8.5

- a Compare the performances of competitors A and B based on the mean of their scores, both before and after dropping the extreme scores for each competitor.
- b Repeat part (a), calculating the median instead of the mean.

- 5.18 XR05-18** Suppose the commodities consumed by Australian households can be grouped into 10 commodity groups: food, alcoholic beverages, clothing, housing, durables, health care, transport and communication, recreation, education and all others. The following table presents the rate of growth in prices of the 10 commodities and the share of total income Australian households allocate to each of the commodities. Calculate the overall growth in consumer prices in Australia.

Commodity	Rate of price increase (%)	Expenditure share
Food	3.25	0.105
Alcoholic beverages	6.03	0.038
Clothing	1.23	0.043
Housing	4.21	0.203
Durables	1.42	0.056
Health care	3.72	0.050
Transport and communication	2.10	0.140
Recreation	2.58	0.176
Education	5.42	0.046
All others	3.16	0.143

5.19 XR05-19 An investment of \$1000 you made 4 years ago was worth \$1200 after the first year, \$1200 after the second year, \$1500 after the third year, and \$2000 today.

- a Calculate the annual rates of return.
- b Calculate the mean and median of the rates of return.
- c Calculate the geometric mean.
- d Discuss whether the mean, median, or geometric mean is the best measure of the performance of the investment.

5.20 XR05-20 Suppose that you bought a stock 6 years ago at \$12. The price of the stock at the end of each year is shown in the table.

Year	1	2	3	4	5	6
Stock price (\$)	10	14	15	22	30	25

- a Calculate the rate of return for each year.
- b Calculate the mean and median of the rates of return.
- c Calculate the geometric mean of the rates of return.
- d Explain why the geometric mean is the best statistic to use to describe what happened to the price of the stock over the 6-year period.

Computer applications

5.21 XR05-21 A company found that the average number of days until the bills were paid by its customers was 15 days. To determine whether changing the colour of its invoices would improve the speed of payment, a company selected 200 customers at random and sent their invoices on blue paper. The number of days until the bills were paid was recorded. Calculate the mean and the median of these data. Report what you have discovered.

5.22 XR05-22 The amount of time taken by 100 respondents to complete a telephone survey is recorded. (Times are rounded to the nearest whole minute.)

- a Use a software package to calculate the mean, the median and the mode.
- b Describe briefly what each measure tells you about the data.

5.23 Example 4.1 in Chapter 4 deals with the problem of graphically summarising 200 electricity bills. Recall that the data are stored in file **XM04-01**. Use your software to calculate the mean, the median and the mode of these data and interpret their values.

5.24 XR05-24 The summer incomes of a sample of 125 business students are recorded.

- a Calculate the mean and the median of these data.
- b What do the two measures of central location tell you about the students' summer incomes?
- c Which measure would you use to summarise the data? Explain.

5.25 Refer to Exercise 4.16, in which the annual incomes of 200 junior trainees in a fast-food chain were stored in file **XR04-16**.

- a Determine the mean and the median of the sample.
- b Briefly describe what each statistic tells you.

5.26 Refer to Exercise 4.18, in which the prices (in thousands of dollars) of homes in a suburb in Adelaide sold last year were stored in file **XR04-18**.

- a Determine the mean and the median of this sample.
- b What information about the prices have you learnt from the statistics in part (a)?

5.27 XR05-27 The owner of a hardware store that sells electrical wire by the metre is considering selling the wire in pre-cut lengths to save on labour costs. A sample of wire lengths sold over the course of one week is recorded.

- a Calculate the mean, the median and the mode for these data.
- b For each measure of central location calculated in part (a), discuss the weaknesses in providing useful information to the store owner.
- c How might the store owner decide upon the lengths to pre-cut?

5.28 XR05-28 The amount of time (in seconds) taken to perform a spot weld on a car under production was recorded for 50 workers.

- a Calculate the mean, the median and the mode of these data.
- b Discuss what information you have discovered from the statistics calculated in part (a).

5.29 XR05-29 Employee training and education have become important factors in the success of many firms. A survey undertaken by a statistical agency attempts to measure the amount of training undertaken by employees of a sample of 144 firms with more than 500 employees. The firms were asked to report the number of hours of employee training for the six-month period from May to October.

- a Calculate the mean and the median.
- b Interpret the statistics you calculated.

5.30 XR05-30 The incomes of a sample of 125 part-time MBA students are recorded.

- a Determine the mean and the median of these data.
- b What do these two statistics tell you about the incomes of part-time MBA students?

5.31 XR05-31 In an effort to slow down traffic in a 70 km/h maximum speed zone, engineers painted a solid line one metre from the curb along the entire length of a road and filled the space with diagonal lines. The lines made the road look narrower. A sample of 120 car speeds was taken after the lines were drawn and are recorded.

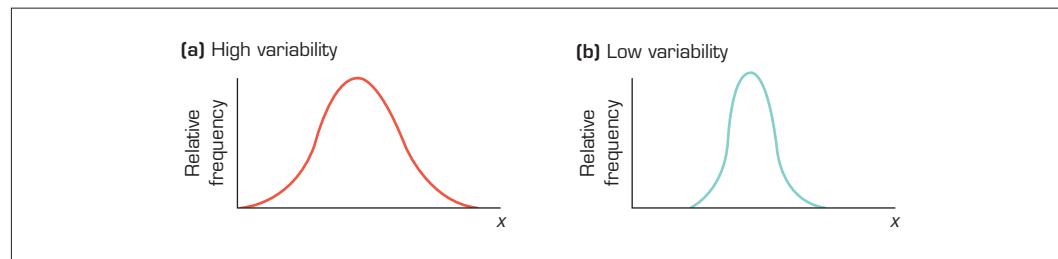
- a Compute the mean, the median and the mode of these data.
- b Briefly describe the information you acquired from each statistic calculated in part (a).

5.2 Measures of variability

We are now able to calculate three measures of central location. But these measures fail to tell the whole story about a distribution of observations.

Once we know the average value of a set of observations, our next question should be ‘How typical is the average value of all observations in the data?'; in other words: ‘How spread out are the observations around their average value?'. Are the observations highly variable and widely dispersed around the average value, as depicted by the smoothed relative frequency polygon in **Figure 5.2(a)**, or do they exhibit low variability and cluster about the average value, as in **Figure 5.2(b)**?

FIGURE 5.2 Smoothed relative frequency polygons



The concept of variability is of fundamental importance in statistical inference. It is therefore important that we be able to measure the degree of variability in a set of observations. In this section, we introduce four measures of variability. Let us start with the simplest measure of variability, the **range**.

range

The difference between largest and smallest observations.

5.2a Range

The advantage of the range is its simplicity. The disadvantage is also its simplicity. Because the range is calculated from only two observations, it tells us nothing about the other observations. Consider the following two sets of data:

Set 1	4	4	4	4	4	50
Set 2	4	8	15	24	39	50

The range of both sets is 46. The two sets of data are completely different and yet their ranges are the same. To measure variability, we need other statistics that incorporate all the data and not just two observations, the smallest and the largest.

5.2b Variance

variance

A measure of variability of a numerical data set.

Variance and its related measure, the *standard deviation*, are arguably the most important statistics (other than the mean). They are used to measure the variability in a set of numerical data. The variance and standard deviation take into account all the data and, as you will discover, play a vital role in almost all statistical inference procedures.

Consider two very small populations, each consisting of five observations:

Population A	8	9	10	11	12
Population B	4	7	10	13	16

The mean of both populations is 10, as you can easily verify. The population values are plotted along the horizontal x -axis in **Figure 5.3**. Visual inspection of these graphs indicates that the observations for population B are more widely dispersed than those for population A. We are searching for a measure of dispersion that confirms this notion and takes into account each observation in the population.

TABLE 5.3A Sum of mean deviations of observations from the mean and mean absolute deviations

Population A			Population B		
x_i	Deviation $(x_i - \mu_A)$	Absolute deviation $ x_i - \mu_A $	x_i	Deviation $x_i - \mu_B$	Absolute deviation $ x_i - \mu_B $
8	$(8 - 10) = -2$	2	4	$(4 - 10) = -6$	6
9	$(9 - 10) = -1$	1	7	$(7 - 10) = -3$	3
10	$(10 - 10) = 0$	0	10	$(10 - 10) = 0$	0
11	$(11 - 10) = 1$	1	13	$(13 - 10) = 3$	3
12	$(12 - 10) = 2$	2	16	$(16 - 10) = 6$	6
Sum	0	6	Sum	0	18

deviation

Difference between an observation and the mean of the set of data it belongs to.

Consider the five observations in populations A and B. To obtain a measure of their dispersion, we might begin by calculating the **deviation** of each value from the mean.

That is:

$$\text{Population A: Sum of absolute deviation} = \sum_{i=1}^5 |x_i - \mu_A| = 6$$

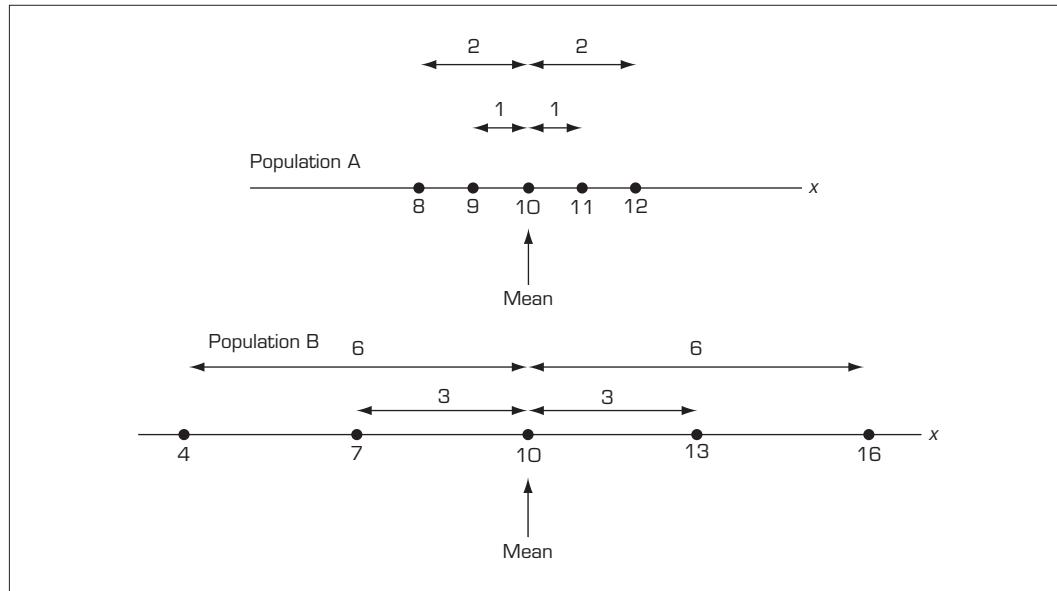
$$\text{Population B: Sum of absolute deviation} = \sum_{i=1}^5 |x_i - \mu_B| = 18$$

The deviations are represented by the double-pointed arrows above the x -axis in **Figure 5.3**. It might at first seem reasonable to take the average of these deviations as a measure of dispersion, but the average of deviation from the mean is always zero because the sum of deviations from the mean is always zero. This difficulty can be overcome, however, by taking the average of the absolute deviations or the squared deviations as the required measure of dispersion.

If we use the absolute deviations,¹ the measure of dispersion, calculated as the sum of the absolute deviations from the mean and divided by the number of observations, is called the *mean absolute deviation* (MAD):

$$\text{MAD} = \frac{1}{n} \sum_{i=1}^n |x_i - \mu|$$

FIGURE 5.3 Deviations of observations from the mean



For population A:

$$\text{MAD}_A = \frac{1}{5} [| -2 | + | -1 | + | 0 | + | 1 | + | 2 |] = \frac{1}{5} [2 + 1 + 0 + 1 + 2] = \frac{6}{5} = 1.2$$

Similarly, for population B:

$$\text{MAD}_B = \frac{1}{5} [| -6 | + | -3 | + | 0 | + | 3 | + | 6 |] = \frac{1}{5} [6 + 3 + 0 + 3 + 6] = \frac{18}{5} = 3.6$$

The MAD of population B therefore exceeds the MAD of population A, confirming our initial visual impression that the values in population B are more dispersed than those in population A.

Mean absolute deviation (MAD)

The mean absolute deviation of a population of N observations x_1, x_2, \dots, x_N with a mean of μ is defined as:

$$\text{Population MAD} = \frac{\sum_{i=1}^N |x_i - \mu|}{N}$$

The mean absolute deviation of a sample of n observations x_1, x_2, \dots, x_n with a mean of \bar{x} is defined as:

$$\text{Sample MAD} = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

In theory, MAD can be used as a measure of dispersion, but due to its limited utility in statistical inference it is seldom calculated.

¹ The notation used for absolute value of a real number x is $|x|$. For example, the absolute value of -2 is written as $|-2|$ and its resulting value is 2 .

The other and the most popular measure of dispersion, or variability, of a population of observations is called the variance; it is denoted by σ^2 (where σ is the lowercase Greek letter *sigma*).

Consider again the five observations in populations A and B. Instead of using the absolute deviations, we could calculate the squared value of the deviation of each observation from the mean, to obtain a measure of their dispersion:

TABLE 5.3B Sum of squared mean deviations of observations from the mean

Population A			Population B		
x_i	$x_i - \mu_A$	$(x_i - \mu_A)^2$	x_i	$x_i - \mu_B$	$(x_i - \mu_B)^2$
8	$(8 - 10) = -2$	4	4	$(4 - 10) = -6$	36
9	$(9 - 10) = -1$	1	7	$(7 - 10) = -3$	9
10	$(10 - 10) = 0$	0	10	$(10 - 10) = 0$	0
11	$(11 - 10) = 1$	1	13	$(13 - 10) = 3$	9
12	$(12 - 10) = 2$	4	16	$(16 - 10) = 6$	36
Sum	0	10	Sum	0	90

That is:

$$\text{Population A: Sum of squared mean deviations} = \sum_{i=1}^5 (x_i - \mu_A)^2 = 10$$

$$\text{Population B: Sum of squared mean deviations} = \sum_{i=1}^5 (x_i - \mu_B)^2 = 90$$

Variance of a population

The **variance of a population** of N observations x_1, x_2, \dots, x_N with a mean of μ is defined as

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_N - \mu)^2}{N} = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

Alternatively, for ease of manual calculations, the population variance can be written in the following short-cut forms:

$$\sigma^2 = \frac{1}{N} \left[\sum_{i=1}^N x_i^2 - \frac{\left(\sum_{i=1}^N x_i \right)^2}{N} \right] \text{ or } \sigma^2 = \frac{1}{N} \left[\sum_{i=1}^N x_i^2 - N\mu^2 \right]$$

Variance of a population

Sum of the squared deviations (from the population mean) of all of the items of data in the population divided by the number of items.

We suggest that, rather than blindly memorising this formula, you think of the variance as being the *mean squared deviation*; that is, the mean of the squared deviations of the observations from their mean μ . This should help you both to remember and to interpret the formula for the variance.

Consider again the five observations in populations A and B. Letting σ_A^2 denote the variance of population A, we obtain:

$$\begin{aligned} \sigma_A^2 &= \frac{(8-10)^2 + (9-10)^2 + (10-10)^2 + (11-10)^2 + (12-10)^2}{5} \\ &= \frac{10}{5} \\ &= 2 \end{aligned}$$

Proceeding in an analogous manner for population B, we obtain:

$$\begin{aligned}\sigma_B^2 &= \frac{(4-10)^2 + (7-10)^2 + (10-10)^2 + (13-10)^2 + (16-10)^2}{5} \\ &= \frac{90}{5} \\ &= 18\end{aligned}$$

The variance of population B therefore exceeds the variance of population A, consistent with our initial visual impression that the values in population B are more dispersed than those in population A.

Now suppose that you are working with a sample, rather than with a population. If you are given a sample of n observations, your interest in calculating the variance of the sample (denoted by s^2) lies in obtaining a good estimate of the population variance (σ^2). Although it would seem reasonable to define the sample variance s^2 as the average of the squared deviations of the sample observations from their mean \bar{x} , doing so tends to underestimate the population variance σ^2 . This problem can be rectified, however, by defining the sample variance s^2 as the sum of the squared deviations divided by $n - 1$, rather than by n . Further discussion of this point is provided in Chapter 10.

Variance of a sample

The **variance of a sample** of n observations x_1, x_2, \dots, x_n having mean \bar{x} is defined as

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

Alternatively, for ease of manual calculations, the sample variance can be written in the following short-cut forms:

$$s^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \right] \text{ or } s^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right]$$

variance of a sample

Sum of the squared deviations (from the sample mean) of all the items of data in a sample, divided by one less than the number of items.

EXAMPLE 5.10

L01

Variation in ANZ credit card annual account fees

XMO5-10 The following are the credit card annual account fees (in A\$) charged by the ANZ bank for a sample of 8 ANZ credit cards.

Card type	Annual account fees
ANZ Low Rate	58
ANZ Low Rate Platinum	99
ANZ First Free Days	30
ANZ First Low Interest	26
ANZ Free Days MasterCard	26
ANZ Low Interest MasterCard	26
ANZ Platinum	87
ANZ Visa PAYCARD	24



- Find the mean and the variance of the annual account fees of the various ANZ credit cards.

Solution

Calculating manually

Card type	x_i	x_i^2
ANZ Low Rate	58	3364
ANZ Low Rate Platinum	99	9801
ANZ First Free Days	30	900
ANZ First Low Interest	26	676
ANZ Free Days MasterCard	26	676
ANZ Low Interest MasterCard	26	676
ANZ Platinum	87	7569
ANZ Visa PAYCARD	24	576
Sum	376	24238

That is,

$$\sum_{i=1}^8 x_i^2 = 24238 \quad \sum_{i=1}^8 x_i = 376$$

The mean of this sample of 8 observations is

$$\begin{aligned}\bar{x} &= \frac{1}{8} \sum_{i=1}^8 x_i \\ &= \frac{376}{8} \\ &= \$47\end{aligned}$$

Therefore, the average annual account fee is \$47.

To find the sample variance, first use the alternative formulas given above. For this we calculate:

$$\begin{aligned}s^2 &= \frac{1}{(8-1)} \left[\sum_{i=1}^8 x_i^2 - \frac{\left(\sum_{i=1}^8 x_i \right)^2}{8} \right] \\ &= \frac{1}{7} \left[24238 - \frac{(376)^2}{8} \right] \\ &= 938 (\$)^2\end{aligned}$$

The variance is 938 (\$)².

Using the computer

COMMANDS

Follow the commands used to calculate the mean in Example 5.1, except type **VAR** instead of **AVERAGE**.

For Example 5.10, type =**VAR(B2:B9)** into any empty cell. The active cell would show the variance as 938.

5.2c Standard deviation

Because calculating variance involves squaring the original observations, the unit attached to a variance is the square of the unit attached to the original observations. In Example 5.10, as our original observations are expressed in dollars, the variance is expressed in dollars-squared.

Although variance is a useful measure of relative variability of two sets of observations, statisticians often want a measure of variability that is expressed in the same units as the original observations, just as the mean is. Such a measure, known as the **standard deviation**, can be obtained simply by taking the square root of the variance.

standard deviation
Square root of the variance.

Standard deviation

The *standard deviation* of a set of observations is the positive square root of the variance of the observations.

$$\text{Population standard deviation: } \sigma = \sqrt{\sigma^2}$$

$$\text{Sample standard deviation: } s = \sqrt{s^2}$$

For example, the standard deviation of the sample of annual account fees in Example 5.10 is

$$s = \sqrt{s^2}$$

$$= \sqrt{938(\$)^2} = \$30.63$$

Note that the unit associated with the standard deviation (\$) is the same unit as the original observations of the data (\$) as well as the mean (\$).

One important application of variance (or, alternatively, of standard deviation) arises in finance, where variance is the most popular numerical measure of risk. For example, we might be concerned with the variance of a firm's sales, profits or return on investment. In all cases, the underlying assumption is that a large variance corresponds to a higher level of risk. The next example illustrates this important application of variance.

Using the computer

COMMANDS

Follow the instructions in Example 5.1 to calculate the mean, except type **STDEV** instead of **AVERAGE**.

For Example 5.10, we type the following into any empty cell: **=STDEV(B2:B9)**. The active cell would store the standard deviation as 30.627.

EXAMPLE 5.11

LO3

Comparing the risk of investments

XMO5-11 Unit trusts are becoming an increasingly popular investment alternative among small investors. To help investors decide which trust to invest in, financial magazines regularly report the average annual rate of return achieved by each of a number of unit trusts over the past 10 years. They also indicate each trust's level of risk, by classifying the historical variability of each trust's rate of return as high, intermediate or low.

If the annual (percentage) rates of return over the past 10 years for two unit trusts are as follows, which trust would you classify as having the higher level of risk?

Trust A	12.3	-2.2	24.9	1.3	37.6	46.9	28.4	9.2	7.1	34.5
Trust B	15.1	0.2	9.4	15.2	30.8	28.3	21.2	13.7	1.7	14.4



Solution

Calculating manually

For each trust, we must find the variance of the rates of return of the sample.

For Trust A, we have:

$$\sum_{i=1}^{10} x_i = 12.3 + -2.2 + \dots + 34.5 = 200.0$$

$$\sum_{i=1}^{10} x_i^2 = (12.3)^2 + (-2.2)^2 + \dots + (34.5)^2 = 6523.06$$

The variance for Trust A is therefore:

$$\begin{aligned}s_A^2 &= \frac{1}{9} \left[\sum_{i=1}^{10} x_i^2 - \frac{\left(\sum_{i=1}^{10} x_i \right)^2}{10} \right] \\&= \frac{1}{9} \left[6523.06 - \frac{(200)^2}{10} \right] \\&= 280.34 (\%)^2\end{aligned}$$

For Trust B we have:

$$\sum_{i=1}^{10} x_i = 15.1 + 0.2 + \dots + 14.4 = 150.0$$

$$\sum_{i=1}^{10} x_i^2 = (15.1)^2 + (0.2)^2 + \dots + (14.4)^2 = 3144.36$$

The variance for Trust B is therefore:

$$\begin{aligned}s_B^2 &= \frac{1}{9} \left[\sum_{i=1}^{10} x_i^2 - \frac{\left(\sum_{i=1}^{10} x_i \right)^2}{10} \right] \\&= \frac{1}{9} \left[3144.36 - \frac{(150.0)^2}{10} \right] \\&= 99.38 (\%)^2\end{aligned}$$

Note that, since the calculation of s^2 involves squaring the original observations, the sample variance is expressed in $(\%)^2$, which is the square of the unit (%) used to express the original observations of rate of return.





Using the computer

Excel output for Example 5.11

	C	D	E	F
1	Trust A		Trust B	
2				
3	Mean	20	Mean	15
4	Standard Error	5.29471	Standard Error	3.15235
5	Median	18.6	Median	14.75
6	Mode	#N/A	Mode	#N/A
7	Standard Deviation	16.7434	Standard Deviation	9.96862
8	Sample Variance	280.34	Sample Variance	99.3733
9	Kurtosis	-1.3419	Kurtosis	-0.46394
10	Skewness	0.21697	Skewness	0.10695
11	Range	49.1	Range	30.6
12	Minimum	-2.2	Minimum	0.2
13	Maximum	46.9	Maximum	30.8
14	Sum	200	Sum	150
15	Count	10	Count	10

COMMANDS

Use the commands described in Example 5.5 to produce descriptive statistics for the data. Excel prints the range, the sample standard deviation and the sample variance, as well as a variety of other statistics, some of which we will present in this book.

Interpreting the results

From the sample data, we conclude that Trust A has the higher level of risk as measured by variance, since the variance of its rates of return exceeds that of the rates of return Trust B.

Notice also that Trust A has produced a higher average rate of return over the past 10 years. Specifically, the mean rates of return for Trusts A and B were:

$$\bar{x}_A = \frac{200.0}{10} = 20\% \text{ and } \bar{x}_B = \frac{150.0}{10} = 15\%$$

This result is in keeping with our intuitive notion that an investment that involves a higher level of risk should produce a higher average rate of return.

Notice that, alternatively, we could have used standard deviation as our measure of variability. For instance, the standard deviations of the rates of return of the samples for Trusts A and B are:

$$s_A = \sqrt{s_A^2} = \sqrt{280.34} = 16.74\%$$

and

$$s_B = \sqrt{s_B^2} = \sqrt{99.38} = 9.97\%$$

As you can see, the observations in Trust A are more variable than those in Trust B, whether we use variance or standard deviation as our measure of variability. But standard deviation is the more useful measure of variability in situations where the measure is to be used in conjunction with the mean to make a statement about a single population, as we shall see below.

5.2d Interpreting the standard deviation

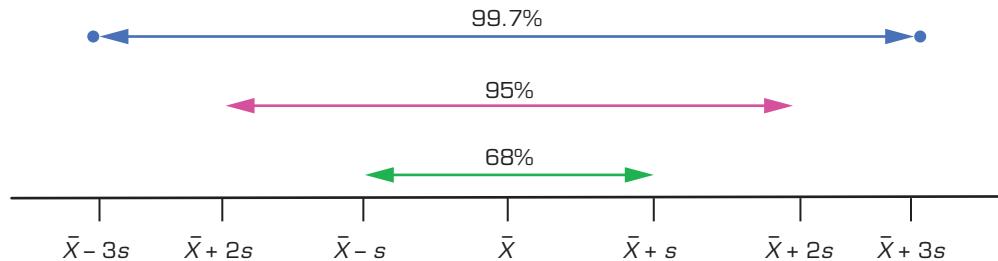
Knowing the mean and standard deviation allows the statistics practitioner to extract useful information from the data. The information depends on the shape of the histogram. If the histogram is bell shaped, we can use the **empirical rule**.

empirical rule

When the distribution is bell shaped, the percentage of observations that fall within 1, 2 and 3 standard deviations (SDs) from the mean are 68%, 95% and 99.7% respectively.

Empirical rule

- 1 Approximately 68% of all observations fall within one standard deviation of the mean.
- 2 Approximately 95% of all observations fall within two standard deviations of the mean.
- 3 Approximately 99.7% of all observations fall within three standard deviations of the mean.

**EXAMPLE 5.12**

LO4

Using the empirical rule to interpret a standard deviation

After an analysis of the returns on an investment, a statistics practitioner discovered that the histogram is bell shaped and that the mean and standard deviation are 10% and 3% respectively. What can you say about the way the returns are distributed?

Solution

Because the histogram is bell shaped, we can apply the empirical rule.

- 1 Approximately 68% of the returns lie between 7% (the mean minus one standard deviation = 10 – 3) and 13% (the mean plus one standard deviation = 10 + 3).
- 2 Approximately 95% of the returns lie between 4% (the mean minus two standard deviations = 10 – 2[3]) and 16% (the mean plus two standard deviations = 10 + 2[3]).
- 3 Approximately 99.7% of the returns lie between 1% (the mean minus three standard deviations = 10 – 3[3]) and 19% (the mean plus three standard deviations = 10 + 3[3]).

As a final point, the empirical rule forms the basis for a crude method of approximating the standard deviation of a sample of observations that has a mound-shaped distribution. Since most of the sample observations (about 95%) fall within 2 standard deviations of the mean, the range of the observations is approximately equal to $4s$. Once we have found the range of the observations, we can calculate an approximate standard deviation of the sample as

$$s = \frac{\text{Range}}{4}$$

This *range approximation of the standard deviation* is useful as a quick check to ensure that our calculated value of s is reasonable, or ‘in the ballpark’. In Example 5.10, the range of the annual credit card account fees is \$99 – \$24 = \$75, so $75/4 = \$18.75$ is an approximation of s . In this case, the range approximation \$18.75 is not very close to \$30.63, our calculated value of s . One reason for this discrepancy may be that the distribution of the data may not be mound-shaped, in which case the range approximation of the standard deviation is not valid.

A more general interpretation of the standard deviation is derived from **Chebyshev’s theorem**, which applies to all shapes of histograms.

Chebyshev’s theorem

The proportion of observations that lie within k standard deviations of the mean is at least $(1 - 1/k^2)$.

Chebyshev's theorem

The proportion of observations that lie within k standard deviations of the mean is at least

$$1 - \frac{1}{k^2} \text{ for } k > 1$$

When $k = 2$, Chebyshev's theorem states that at least three-quarters (75%) of all observations lie within two standard deviations of the mean. With $k = 3$, Chebyshev's theorem states that at least eight-ninths (88.9%) of all observations lie within three standard deviations of the mean.

Note that the empirical rule provides approximate proportions based on a symmetrical distribution, whereas Chebyshev's theorem provides lower bounds on the proportions contained in the intervals based on any distribution.

EXAMPLE 5.13

LO4

Using Chebyshev's theorem to interpret a standard deviation

XMO5-13 The durations (in minutes) of a sample of 30 telephone calls placed by a firm in Melbourne in a given week are recorded in **Table 5.4**.

TABLE 5.4 Duration of telephone calls (in minutes) of a Melbourne firm

11.8	3.6	16.6	13.5	4.8	8.3
8.9	9.1	7.7	2.3	12.1	6.1
10.2	8.0	11.4	6.8	9.6	19.5
15.3	12.3	8.5	15.9	18.7	11.7
6.2	11.2	10.4	7.2	5.5	14.5

The 30 telephone call durations have a mean of $\bar{x} = 10.26$ and a standard deviation of $s = 4.29$. Given no other information about the distribution of the durations, Chebyshev's theorem asserts that at least three-quarters or 75% of the durations lie within two standard deviations of the mean:

$$[\bar{x} - 2s, \bar{x} + 2s] = [10.26 - 2(4.29), 10.26 + 2(4.29)] = [1.68, 18.84]$$

In fact, all but the largest of the 30 durations fall in this interval; that is, the interval actually contains 96.7% (29 out of 30) of the telephone call durations – a percentage well above the lower bound asserted by Chebyshev's theorem.

5.2e Coefficient of variation

Is a standard deviation of 10 a large number indicating great variability, or is it a small number indicating little variability? The answer depends somewhat on the magnitude of the observations in the data set. If the magnitudes of the observations are in the millions, a standard deviation of 10 would probably be considered a small number. On the other hand, if the magnitudes of the observations are in the hundreds, a standard deviation of 10 would be seen as a large number. This logic lies behind yet another measure of variability, the **coefficient of variation**.

coefficient of variation

Standard deviation divided by the mean.

Coefficient of variation

The *coefficient of variation* of a set of observations is the standard deviation of the observations divided by their mean.

$$\text{Population coefficient of variation: } cv = \frac{\sigma}{\mu}$$

$$\text{Sample coefficient of variation: } cv = \frac{s}{\bar{x}}$$

The coefficient of variation is usually multiplied by 100 and reported as a percentage, which effectively expresses the standard deviation as a percentage of the mean.

In Example 5.11, the average rates of return for Trusts A and B are 20% and 15% respectively. The corresponding standard deviations are 16.74% and 9.97% respectively. If the decision is made based on better average rate of return, then Trust A is preferred to Trust B. However, if the decision is made on the basis of risk (standard deviation), then Trust B would be preferred to Trust A. If the decision is to be made based on both higher return and lower risk, then the trust to be selected will be based on the value of the coefficient of variation.

EXAMPLE 5.14

Relative variation in the rates of return

Calculate the coefficient of variation of the rates of return for Trusts A and B in Example 5.11.

Solution

The means and standard deviations of the rates of return for Trusts A and B in Example 5.11 are $\bar{x}_A = 20$, $s_A = 16.74$; $\bar{x}_B = 15$, $s_B = 9.97$. Trust A has a higher rate of return and a larger variation and so a higher risk.

The coefficients of variation of the sample rates of return for Trusts A and B in Example 5.11 are:

$$cv_A = \frac{s_A}{\bar{x}_A} = \frac{16.74}{20} = 0.837$$

$$cv_B = \frac{s_B}{\bar{x}_B} = \frac{9.97}{15} = 0.665$$

Thus, in percentages, for the Trust A and Trust B returns, the coefficients of variation are 83.7% and 66.5% respectively. In this particular case, comparing coefficients of variation and comparing standard deviations lead to the same conclusion: the observations in sample A are more variable and therefore Trust B is preferred to Trust A. But if the mean return for Trust A in Example 5.11 was, for example, 27% with the same standard deviation, $s_A = 16.74\%$, the coefficient of variation of the return for Trust A would then be:

$$cv_A = \frac{s_A}{\bar{x}_A} = \frac{16.74}{27} = 0.62$$

Therefore in that case, Trust A would be preferred to Trust B.

5.2f Measures of variability for ordinal and nominal data

The measures of variability introduced in this section can only be used for numerical data. The next section will feature a measure that can be used to describe the variability of ordinal data. There are no measures of variability for nominal data.

We complete this section by reviewing the factors that identify the use of measures of variability.

IN SUMMARY

Factors that identify when to calculate the range, variance, standard deviation and coefficient of variation

- 1 *Objective:* to describe a single set of data
- 2 *Type of data:* numerical
- 3 *Descriptive measurement:* variability

EXERCISES

Learning the techniques

- 5.32** a Is it possible for a standard deviation to be negative? Explain.
 b Is it possible for the standard deviation of a set of data to be larger than its variance? Explain.
 c Is it possible for the standard deviation of a set of data to be zero? Explain.
 d Is it possible for a coefficient of variation to be negative? Explain.
 e Is it possible for the coefficient of variation of a set of data to be zero? Explain.

- 5.33 XR05-33** Calculate the range, variance, standard deviation and coefficient of variation for the following sample of data:

5	7	12	14	15	15	17	20	21	24
---	---	----	----	----	----	----	----	----	----

Refer to the sample of data in the table and try to answer each of the following questions without performing any calculations. Then verify your answers by performing the necessary calculations.

- a If we drop the largest value from the sample, what will happen to the mean, variance, standard deviation and coefficient of variation?
- b If each value is increased by 2, what will happen to the mean, variance, standard deviation and coefficient of variation?
- c If each value is multiplied by 3, what will happen to the mean, variance, standard deviation and coefficient of variation?

- 5.34 XR05-34** Calculate the mean, variance, standard deviation and coefficient of variation for each of the following samples of data:

- a 14 7 8 11 5
- b -3 -2 -1 0 1 2 3
- c 4 4 8 8
- d 5 5 5 5

- 5.35 XR05-35** Examine the three samples of data shown below:

1	27	39	22	36	31
2	32	28	33	30	27
3	34	47	16	49	39

- a Without performing any calculations, indicate which sample has the greatest variability and which sample has the least variability. Explain why.
- b Calculate the variance, standard deviation and coefficient of variation of the three samples. Was your answer in part (a) correct?
- c Calculate $\sum(x_i - \bar{x})$ for the three samples. What can you infer about this calculation in general?

- 5.36 XR05-36** The number of hours a student spent studying for a statistics course over the past seven days was recorded as follows.

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

Calculate the range, mean, variance, standard deviation and coefficient of variation for these data. Express each answer in appropriate units.

5.37 XR05-37 The unemployment rates of the Australian states and territories during September 2018 for males and females aged 15 and over are presented below. Calculate the mean, standard deviation and coefficient of variation of the unemployment rates for males and females. Discuss the variability in unemployment between the Australian states and territories for males and females during September 2018.

State/Territory	Male	Female
New South Wales	4.3	4.5
Victoria	4.2	5.0
Queensland	6.0	5.9
South Australia	5.5	5.5
Western Australia	5.9	6.2
Tasmania	5.2	6.4
Northern Territory	4.5	3.6
Australian Capital Territory	3.3	4.0

Source: Australian Bureau of Statistics, *Labour Force Australia*, September 2018, cat. no. 6202.0, ABS, Canberra, Labour Underutilisation (aged 15 and over).

State/Territory	Number of deaths (12 months ended in March)				
	2016	2017	2018	2019	2020
New South Wales	362	367	406	361	331
Victoria	265	268	258	235	262
Queensland	252	250	256	229	219
South Australia	101	84	102	91	116
Western Australia	168	178	160	166	149
Tasmania	34	31	38	30	37
Northern Territory	50	44	38	42	35
Australian Capital Territory	15	12	5	7	5

Source: © Commonwealth of Australia CC BY 3.0 Australia <https://creativecommons.org/licenses/by/3.0/au/>

Calculate the mean and the standard deviation of road deaths in each state and territory of Australia over the years 2016–20. Comment on the average number of road deaths and its variation over the years.

5.40 XR05-40 The entry price of a sample of 15 Australian managed funds are listed in the table.

0.05	0.84	0.99	1.16	1.36	0.92	1.02	0.89
0.97	1.12	0.93	0.97	0.86	0.91	2.98	

- a Calculate the mean and standard deviation of the entry price.
- b Calculate the median entry price of the funds.
- c Repeat parts (a) and (b) after eliminating the first and last values.

5.38 In Exercise 4.8, you were asked to depict graphically the distribution of the following sample of statistics exam completion times (in minutes):

61	86	61	58	70	75	66	77	66	64
73	91	65	59	86	82	48	67	55	77
80	58	94	78	62	79	83	54	52	45

- a Find the mean and the median of this set of completion times.
- b Using the frequency distribution in Exercise 4.8, determine the mode as the midpoint of the modal class.
- c Locate the mean, the median and the mode on the relative frequency histogram constructed in Exercise 4.8.
- d Use the shortcut formula to find the variance of this sample of completion times.

5.39 XR05-39 The number of road deaths during 12 months ended in March from 2016 to 2020 in the six Australian states and two territories are listed in the table.

5.41 The mean and standard deviation of the wages of 1000 factory workers are \$25 600 and \$2200 respectively. If the wages have a mound-shaped distribution, how many workers receive wages of between:

- a \$23 400 and \$27 800?
- b \$21 200 and \$30 000?
- c \$19 000 and \$32 200?

5.42 A set of data whose histogram is bell shaped yields a mean and a standard deviation of 50 and 4 respectively. Approximately what proportion of observations are between:

- a 46 and 54?
- b 42 and 58?
- c 38 and 62?

5.43 A statistics practitioner determined that the mean and the standard deviation of a data set were 120 and 30 respectively. What can you say about the proportions of observations that lie within each of the following intervals?

- a 90 and 150
- b 60 and 180
- c 30 and 210

Applying the techniques

5.44 XR05-44 Self-correcting exercise. The 15 shares in your portfolio had the following percentage changes in value over the past year:

3	0	6	-5	-2	5	-18	20
14	18	-10	10	50	-20	14	

- a Calculate μ , σ^2 and σ for this population of data. Express each answer in appropriate units.
- b Calculate the range and the median.

5.45 XR05-45 The owner of a hardware shop that sells electrical wire by the metre is considering selling the wire in pre-cut lengths in order to reduce labour costs. A sample of lengths (in metres) of wire sold over the course of one day produced the following data:

3	7	4	2.5	3	20	5	5	15	3.5	3
---	---	---	-----	---	----	---	---	----	-----	---

- a Find the mean, the variance and the standard deviation.
- b Calculate the range and the approximate standard deviation.

5.46 XR05-46 The following 20 values represent the number of seconds required to complete one spot-weld by a sample of 20 automated welders on a company's production line:

2.1	2.7	2.6	2.8	2.3	2.5	2.6	2.4	2.6	2.7
2.4	2.6	2.8	2.5	2.6	2.4	2.9	2.4	2.7	2.3

- a Calculate the variance and the standard deviation for this sample of 20 observations.
- b Use the range approximation of the standard deviation s to check your calculations in part (a). What assumption must you make in order to use this approximation?

5.47 Last year, the rates of return on the investments in a large portfolio had an approximately mound-shaped

distribution, with a mean of 20% and a standard deviation of 10%.

- a What proportion of the investments had a return of between 10% and 30%?
- b What proportion of the investments had a return that was between 0% and 40%?
- c What proportion of the investments had a return of between -10% and 50%?
- d What proportion of the investments had a positive return?

(Hint: A mound-shaped distribution is symmetrical.)

5.48 XR04-48 Consider once again the annual percentage rates of return over the past 10 years of unit trusts A and B in Example 5.11. Suppose that 10 years ago you formed a portfolio by investing equal amounts of money in each of the two trusts. The rate of return you would have earned on the portfolio over the first year would then have been $0.5(12.3) + 0.5(15.1) = 13.7\%$.

- a Calculate the rate of return earned on the portfolio for each of the 10 years.
- b Find the mean return on the portfolio over the past 10 years.
- c Find the standard deviation of the portfolio returns over the past 10 years.
- d Find the coefficient of variation of the portfolio returns over the past 10 years.
- e Rank the three possible investments (Trust A, Trust B and the portfolio) according to their average returns, riskiness (as measured by standard deviation) and coefficient of variation over the past 10 years.

Computer applications

5.49 XR05-49 The dividend yields on shareholders' funds of 120 Australian companies are recorded in the data file. Column 1 stores the dividend yields for the first 1–40 companies ranked by market capitalisation, column 2 stores those for the middle 41–80 companies and column 3 for the bottom 81–120 companies. To understand the meaning of these yields, consider the 3.43% yield that was realised on the first company's shares. This means that \$100 invested in shares of that company at the beginning

of the year would have yielded a profit of \$3.43 over the year, leaving you with a total of \$103.43 at the year's end.

- a Find the mean, median, range, standard deviation and coefficient of variation of the yield for the first 40 Australian companies.
- b Repeat part (a) for the middle 41–80 companies.
- c Repeat part (a) for the bottom 81–120 companies.
- d Is there any major difference in the dividend yield of the three groups of companies?

- 5.50 XR05-50** The gross performance of two groups of Australian managed funds with a rating of 3 and 4 stars, over a 1-year and 3-year investment period for a sample of five trusts are shown in the following table. (Gross performance of a unit trust is measured as gross returns compounded over the given period, assuming all income and growth is reinvested. A 15% return over three years would mean an average of 15% performance for each of three years.)

Gross performance (%)

Period	Managed funds (4-star rated)				
1-year	7.31	9.82	10.15	12.42	9.65
3-year	5.82	6.42	6.17	6.59	3.87
Period	Managed funds (3-star rated)				
1-year	10.44	8.83	9.68	10.62	11.25
3-year	5.86	5.20	4.88	6.10	6.40

- a Find the mean, median, range, standard deviation and coefficient of variation of this sample of 4-star rated managed funds.
- b Repeat part (a) for the 3-star rated managed funds.
- c Which type of 1-year investment (4-star rated or 3-star rated managed funds) appears to have:
 - i the higher level of risk?
 - ii the higher average performance?
- d Compare the risk and average performance for the two periods of investment for the 4-star rated managed funds.
- e Repeat part (d) for the 3-star rated managed funds.

- 5.51** Refer to Exercise 5.22, in which the amount of time taken by 100 respondents to complete a telephone survey is stored in column 1 of file **XR05-22**.
- a Use a software package to calculate the variance and the standard deviation.
 - b Use a software package to draw the histogram.

- 5.52** Example 4.1 dealt with the problem of graphically summarising 200 electricity bills. Recall that the data were stored in file **XMO4-01**. Use your software to calculate several measures of dispersion.

- 5.53 XR05-53** A sample of 400 visitors to an exhibition was timed to determine how long each took to view the exhibit. Three samples were taken: one in the morning (134), the second in the afternoon (133) and the third in the evening (133). These data are stored in columns 1, 2 and 3 respectively of the data file.

- a Determine the mean and the median of each sample.
- b Determine the range, the variance and the standard deviation of each sample.
- c Discuss the similarities and differences among the three samples.
- d What are the implications of your findings?

- 5.54** Refer to Exercise 4.20. Recall that the number of customers entering a bank between 10 a.m. and 3 p.m. for the past 100 working days was recorded in columns 1 to 5 of file **XRO4-20**.

- a For each hour from 10 a.m. to 3 p.m., determine the mean and the standard deviation.
- b Briefly describe what the statistics calculated in part (a) tell you.

- 5.55 XR05-55** Refer to Exercise 5.49, which deals with the dividend yields on shareholders' funds of 120 Australian companies.

- a Calculate the standard deviation s of the dividend yields of the 120 companies.
- b Use the range approximation of s to check your answers to part (a).
- c Interpret the results.

- 5.56 XR05-56** Many traffic experts argue that the most important factor in traffic accidents is not the average speed of cars but the amount of variation. Suppose that the speeds of a sample of 200 cars were recorded over a stretch of highway that has seen numerous accidents. Calculate the variance and the standard deviation of the speeds, and interpret the results.

- 5.57 XR05-57** Three men were trying to make the football team. The coach had each of them kick the ball 50 times and the distances were recorded.
- a Calculate the mean, variance and standard deviation for each man.
 - b What do these statistics tell you about each man's performance?

5.58 XR05-58 Variance is often used to measure quality of products in a production line. Suppose that a sample is taken of steel rods that are supposed to be exactly 100 cm long. The length of each rod is determined, and the results are recorded. Calculate the mean, variance and standard deviation. Briefly describe what these statistics tell you.

5.59 XR05-59 To learn more about the size of withdrawals at an ATM, the bank manager took a sample of 75 withdrawals and recorded the amounts. Determine the mean and standard deviation of these data, and describe what these two statistics tell you about the withdrawal amounts.

5.60 XR05-60 Everyone is familiar with waiting in lines. For example, people wait in line at a supermarket to go through the checkout. There are two factors that determine how long a queue becomes: one is the speed of service, the other is the number of arrivals at the checkout. The mean number of arrivals is an important number, but so is the standard deviation. Suppose that a consultant for the supermarket counts the number of arrivals per hour during a sample of 150 hours.

- Calculate the standard deviation of the number of arrivals.
- Assuming that the histogram of the number of arrivals is bell shaped, interpret the standard deviation.

5.3 Measures of relative standing and box plots

The measures of central location (Section 5.1) and measures of variability (Section 5.2) provide the statistics practitioner with useful information about the location and dispersion of a set of observations. The measures in this section describe another aspect of the shape of the distribution of data, as well as providing information about the position of particular observations relative to the entire data set. We have already presented one measure of relative standing, the median, which is also a measure of central location. Recall that the median divides the data set into halves, allowing the statistics practitioner to determine in which half of the data set each observation lies. The statistics we are about to introduce will give you much more detailed information.

5.3a Percentiles

Percentile

The p th percentile of a set of observations is the value for which at most $p\%$ of the observations are less than that value and at most $(100 - p)\%$ of the observations are greater than that value, when the observations are arranged in an ascending order.

The p th **percentile** is defined in much the same manner as the median, which divides a series of observations in such a way that at most 50% of the observations are smaller than the median and at most 50% of the observations are greater. In fact, the median is simply the 50th percentile. Suppose, for example, that your statistics mid-semester examination mark is reported to be in the 60th percentile. This means that at most 60% of the student marks are below or equal to yours and at most 40% of the marks are above yours.

Just as we have a special name for the percentile that divides the ordered set of observations in half, we have special names for percentiles that divide the ordered set of observations into quarters, fifths and tenths: they are **quartiles**, quintiles and deciles respectively.

Note that quartiles are dividers – values that divide the entire range of observations into four equal quarters. In practice, however, the word *quartile* is sometimes used to refer to one of these quarters. An observation *in the first quartile* is in the bottom 25% of the observations, whereas an observation *in the upper quartile* is among the top 25%.

percentile

The p th percentile is the value for which $p\%$ of observations are less than that value and $(100 - p)\%$ are greater than that value.

quartiles

The 25th (Q_1), 50th (Q_2 , or median), and 75th (Q_3) percentiles.

The following list identifies some of the most commonly used percentiles, together with notation for the quartiles:

$O_1 =$	first (lower) decile	= 10th percentile
$O_2 =$	first (lower) quartile	= 25th percentile
$O_3 =$	second (middle) quartile	= median (50th percentile)
	third (upper) quartile	= 75th percentile
	ninth (upper) decile	= 90th percentile

5.3b Locating percentiles

The following formula allows us to *approximate* the location of any percentile.

Location of a percentile

$$L_p = (n+1) \frac{p}{100}$$

where L_p is the location of the p th percentile.

EXAMPLE 5.15

LO5

Calculating percentiles

XM05-15 Calculate the 25th, 50th and 75th percentiles (first, second and third quartiles) of the following data:

10 17 22 15 43 24 18 10 19 32

Solution

Calculating manually

Placing the 10 observations in ascending order, we get:

10 10 15 17 18 19 22 24 32 43

The location of the 25th percentile is:

$$L_{25} = (10+1) \frac{25}{100} = (11)(0.25) = 2.75$$

The 25th percentile is three-quarters of the distance between the second (which is 10) and the third (which is 15) observations. Three-quarters of the distance between 10 and 15 is:

$$(0.75)(15 - 10) = 3.75$$

Because the second observation is 10, the 25th percentile is $10 + 3.75 = 13.75$.

To locate the 50th percentile, we substitute $p = 50$ into the formula and produce

$$L_{50} = (10+1) \frac{50}{100} = (11)(0.50) = 5.5$$

This means that the 50th percentile is halfway between the fifth and sixth observations. The fifth and sixth observations are 18 and 19 respectively. The 50th percentile is $18 + 0.5(19 - 18) = 18.5$, which is also the median.

The location of the 75th percentile is:

$$L_{75} = (10+1) \frac{75}{100} = (11)(0.75) = 8.25$$



Thus, it is located one-quarter of the distance between the eighth and the ninth observations, which are 24 and 32 respectively. One-quarter of the distance is:

$$(0.25)(32 - 24) = 2$$

which means that the 75th percentile is $24 + 2 = 26$.

Using the computer

COMMANDS

- 1 Type in the data or open the data file (**XM05-15**).
- 2 Click the cell reference for the output (**C1**).
- 3 Click insert function **fx** and select **All** or **Statistical** from the **Category** drop-down menu, then select the function **PERCENTILE.EXC** and click **OK**.
- 4 Type the input range (exclude the cell containing the name of the variable) into the **Array** box (**A2:A11**).
- 5 Type the percentile as a decimal (**p/100**) into the **K** box and click **OK**. (**0.25** for 25th percentile or **0.40** for 40th percentile or **0.75** for 75th percentile)

Alternatively, the p th percentile of an array of data can be calculated by replacing steps 3–5 with typing the following command in cell C1: **=PERCENTILE.EXC(Array, p/100)**.

EXAMPLE 5.16

LO5

Calculating quartiles – Telephone bills

XM05-16 The telephone bill amounts of 200 new customers in Melbourne were recorded. Determine the quartiles.

Solution

Using the computer

COMMANDS

- 1 Type in the data or open the data file (**XM05-16**).
- 2 Click the cell reference for the output (**C1**).
- 3 Click insert function **fx** and select **All** or **Statistical** from the **Category** drop-down menu, then select the function **QUARTILE.EXC** and click **OK**.
- 4 Type the input range (exclude the cell containing the name of the variable) into the **Array** box (**A2:A201**).
- 5 Type the quartile number into the **QUART** box (**1** or **2** or **3**) and click **OK**.

Alternatively, steps 3–5 can be replaced by typing **=QUARTILE.EXC(A2:A201,1)** in the activated cell C1. Repeat the procedure for quartiles 2 and 3. Excel calculates the third and first quartiles in the following way. The 3rd quartile is 84.825, which is the number such that 150 numbers are below it and 50 numbers are above it. The 1st quartile is 9.385, which is the number such that 50 numbers are below it and 150 numbers are above it. The 2nd quartile or the median is 26.905.

We can often get an idea of the shape of the histogram from the quartiles. For example, if the first and second quartiles are closer to each other than are the second and third quartiles, the histogram is positively skewed. If the first and second quartiles are further apart than the second and third quartiles, the histogram is negatively skewed. However, if the difference between the first and second quartiles is approximately equal to the difference between the second and third quartiles, the histogram is not necessarily symmetric. We need to examine the entire distribution to draw that conclusion.

5.3c Interquartile range (IQR)

interquartile range (IQR)

The difference between the first and third quartiles.

Interquartile range

$$\text{Interquartile range} = Q_3 - Q_1$$

The interquartile range measures the spread of the middle 50% of the observations. Large values of this statistic mean that the first and third quartiles are far apart, indicating a high level of variability.

For Example 5.15, using the first and third quartiles, we find:

$$\text{IQR} = Q_3 - Q_1 = 26 - 13.75 = 12.25$$

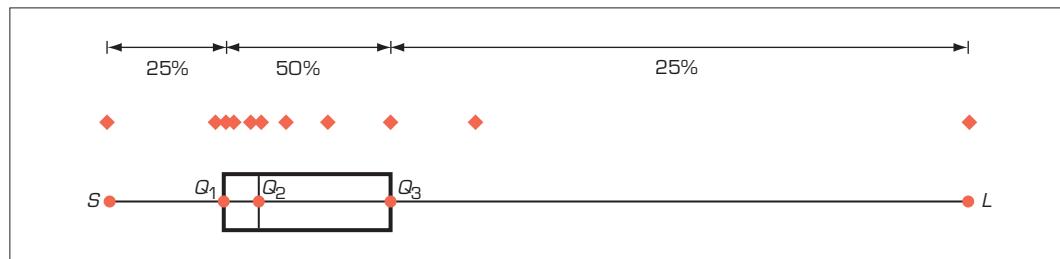
5.3d Box plots

box plot

Pictorial display showing the five summary statistics, the three quartiles and the largest and smallest values of the data.

1	Smallest	S
2	Lower quartile	Q_1
3	Median	Q_2
4	Upper quartile	Q_3
5	Largest	L

FIGURE 5.4 Display of five summary values – box plot



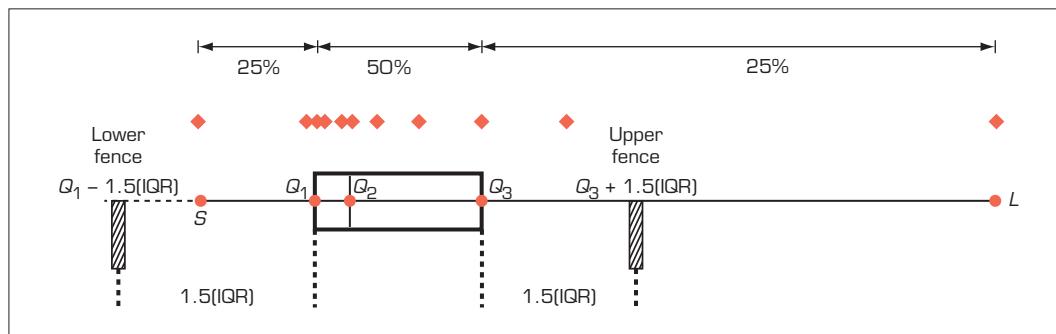
A box plot can be used to (a) identify the shape of the distribution (symmetric, positively skewed or negatively skewed) of a given data set and (b) to identify outliers in a data set.

Strictly speaking, **Figure 5.4** is *not* a box plot, although it is close to being one. One small adjustment to the whiskers is needed, which involves the concept of an *outlier*, the name given to an unusually large or small value in a data set. For our purposes, we will define an *outlier* as a value located at a distance of more than 1.5(IQR) from the box, see **Figure 5.5**.² Therefore, any value less than $Q_1 - 1.5(\text{IQR})$ or greater than $Q_3 + 1.5(\text{IQR})$ is considered an outlier. That is, values falling outside the interval $[Q_1 - 1.5(\text{IQR}), Q_3 + 1.5(\text{IQR})]$ are outliers and the endpoints of this interval are called *lower* and *upper fences*.

² A value located at a distance between 1.5(IQR) and 3(IQR) from the box is sometimes called a *moderate outlier* and a value located at a distance of more than 3(IQR) from the box is called an *extreme outlier*.

outlier

An observation more than $Q_3 + 1.5(\text{IQR})$ or less than $Q_1 - 1.5(\text{IQR})$.

FIGURE 5.5 Box plot with upper and lower fences**lower fence** $Q_1 - 1.5(\text{IQR})$ **upper fence** $Q_3 + 1.5(\text{IQR})$ **EXAMPLE 5.17**

LO5

Identifying the outliers – Share prices of 11 stocks at the ASX

XM05-17 Consider the share prices of 11 stocks at the Australian Securities Exchange (ASX) presented in Table 5.5. Display the data in the form of a box plot and identify any outliers in the data.

TABLE 5.5 Share prices of 11 stocks (in \$A), Australian Securities Exchange (31 March 2019)

Company	Price (\$)		Rank
Telstra Corporation	3.32	← Smallest (S)	1
JB Hi-Fi Ltd	24.95		2
National Australia Bank	25.27	← Q_1	3
Ansell Ltd	25.42		4
Westpac Banking Group	25.92		5
ANZ Banking Group Ltd	26.03	← Median (Q_2)	6
Caltex Australia	26.21		7
Woolworths Group Ltd	30.40		8
Wesfarmers Ltd	34.65	← Q_3	9
Flight Centre	42.05		10
Commonwealth Bank	70.64	← Largest (L)	11

Source: www.marketindex.com.au/asx100**Solution**

Consider the data in **Table 5.5**, the share price of 11 stocks at the Australian Securities Exchange (ASX). The first step in creating a box plot is to rank the data and note the smallest and largest values. (The share prices of the stocks in Table 5.5 are already ranked from smallest to largest.) The smallest value in this case is \$3.32 and the largest is \$70.64.

The next step is to identify the other three values to be displayed. The stock with the median value (\$26.03) is ANZ, because its rank (6) is the middle rank. The lower quartile (Q_1) is the value of 25.27, because at most 25% of the values are smaller and at most 75% of the values are larger. Similarly, the upper quartile (Q_3) is the value of 34.65, because at most 75% of the values are smaller and at most 25% of the values are larger.

1	Smallest (S)	3.32
2	Lower quartile (Q_1)	25.27
3	Median (Q_2)	26.03
4	Upper quartile (Q_3)	34.65
5	Largest (L)	70.64

The five values of interest are plotted in **Figure 5.6**. In this plot, a box with endpoints Q_1 and Q_3 is used to represent the middle half (middle 50%) of the data. Notice that about a quarter of the data fall along the line (or *whisker*) to the left of the box, and about a quarter of the data fall along the whisker to the right of the box. The location of the

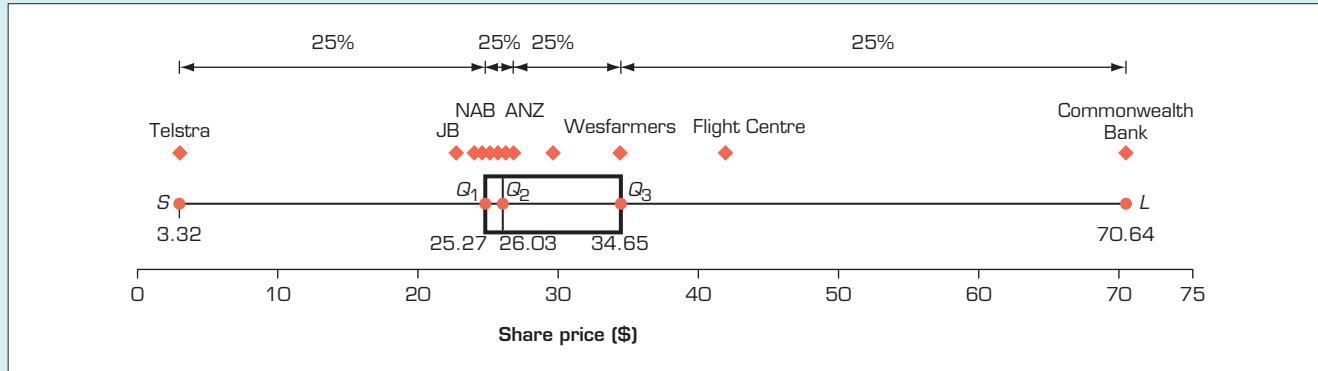




median is indicated by the vertical line inside the box. The shares corresponding to some of the plotted values are also identified. The length of the box is given by the interquartile range.

$$\text{IQR} = Q_3 - Q_1 = 34.65 - 25.27 = 9.38$$

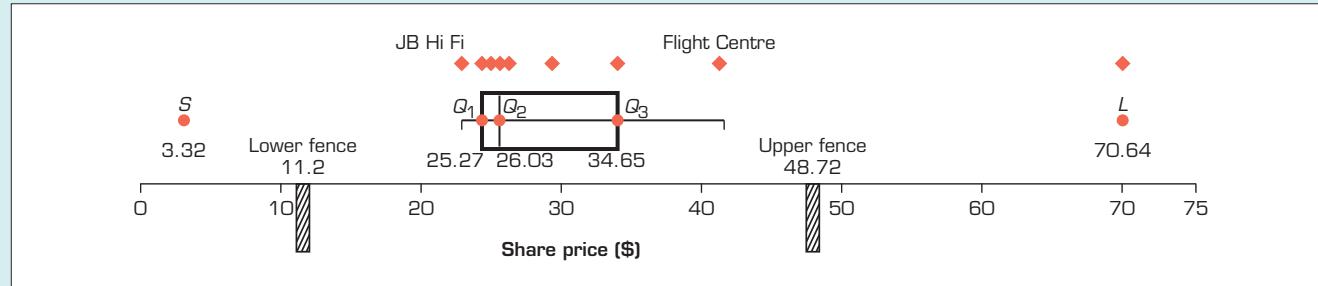
FIGURE 5.6 Display of five summary values – Box plot



In this example, $1.5(\text{IQR}) = 1.5(9.38) = 14.07$. Therefore, an outlier is any value outside the interval $[25.27 - 14.07, 34.65 + 14.07] = [11.2, 48.72]$. The endpoints of this interval are the fences, outside of which all values are outliers. That is, there are only two outliers in this example, namely, Telstra (\$3.32) and Commonwealth Bank (\$70.64).

The lines, or whiskers, emanating from each end of the box in a box plot should extend to the most extreme value that is not an outlier; in this case, to the values of JB Hi-Fi Ltd (24.95) and Flight Centre (42.05). The resulting box plot is shown in **Figure 5.7**.

FIGURE 5.7 Box plot and fences for the share prices of the 11 stocks



Using the computer

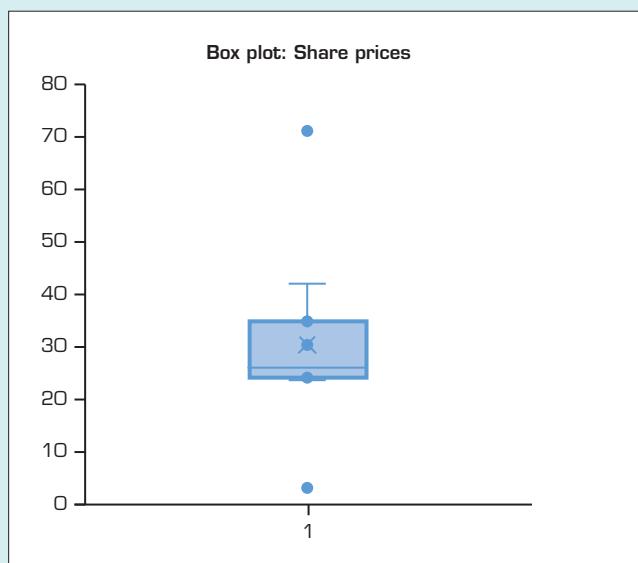
Although the statistics in a box plot can be calculated manually and the box plot can be drawn by hand, it is far easier to let the computer do the work.³ To show how the computer outputs box plots, we have used Excel to print the box plot for the data in **Table 5.5**. The Excel box plot output will show the minimum, the maximum, the three quartiles and the outliers (if ticked).

³ The Box plot option is available only in Microsoft Excel 2016 or higher. Otherwise, we recommend that the quartiles, minimum and maximum can be calculated using Excel and box plot can be hand-drawn. Interquartile range and fences can be calculated using these and outliers can be identified. Alternatively, XLSTAT can be used.





Excel box plot for Example 5.17 with outliers



COMMANDS

- 1 Type in the data in one column or open the data file (**XM05-17**).
- 2 Select/Highlight the column of data (**B2:B12**). On the top menu bar select **Insert** and then, under **Charts**, select **Insert Statistic chart**. Select the option **Box and Whisker**. The box plot for the selected data will now appear. Outliers may already be visible, If not, to identify the outliers, right click inside the box and select **Format Data Series** and then tick **Show outlier points**.

Notice that the quartiles produced in the box plot are not exactly the same as those produced by descriptive statistics. The methods of determining these statistics vary slightly.

Interpreting the results

From the box plot in **Figure 5.7**, we can quickly grasp several points concerning the distribution of the 11 share prices of Australian stocks. The values range from \$3.32 to \$70.64, with about half being smaller than \$26.03 and about half larger than \$26.03. About half the values lie between \$25.27 and \$34.65, with about a quarter below \$25.27 and a quarter above \$34.65. The distribution is somewhat skewed slightly to the right (mean = 30.4 and the median = 26.03) and there is one outlier, \$3.32, to the left and \$70.64 to the right.

Notice that the shape of a box plot is not heavily influenced by a few extreme observations, because the median and the other quartiles are not unduly influenced by extreme observations, in contrast with means and variances.

5.3e Outliers

In our discussion of the box plots, we introduced the notion of an *outlier*: an unusually large or small value in a sample. Because an outlier is considerably removed from the main body of a sample, its validity is suspect and some investigation is needed to check that it is not the result of an error in measuring, recording or transcribing the value. As well as providing a graphical summary of a data set, a box plot is useful for identifying outliers before performing further statistical analysis on a data set.

EXAMPLE 5.18

L05

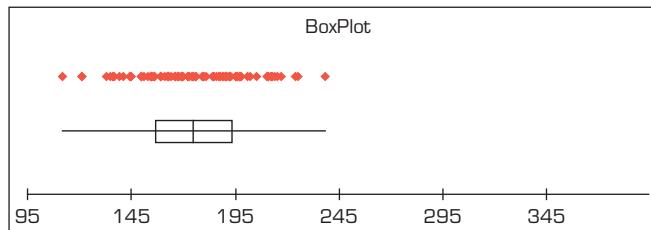
Box plot and fast-food restaurants

XM05-18 Many fast-food restaurants have drive-through windows, offering drivers the advantages of quick service. To measure how fast this service is, an organisation called QSR organised a study in which the length of time taken by a sample of drive-through customers at each of five restaurants was recorded. Compare the five sets of data using a box plot, and interpret the results.

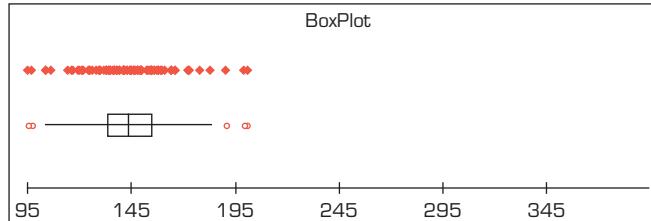
Solution**Calculating manually**

Using the hand calculation method outlined in Example 5.15, we find the following box plots for the five restaurants.

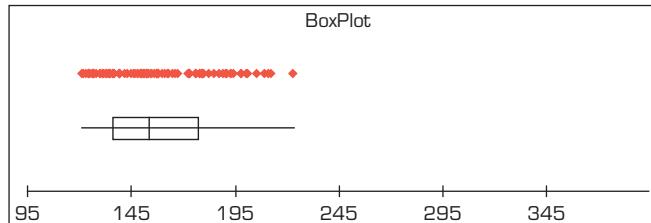
	A	B
2	Restaurant 1	
3	Smallest = 112	
4	Q1 = 156.75	
5	Median = 175	
6	Q3 = 192.75	
7	Largest = 238	
8	IQR = 36	
9	Outliers: None	



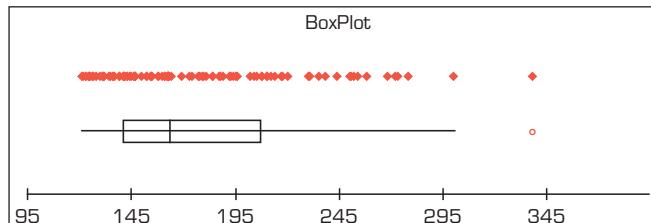
	A	B
15	Restaurant 2	
16	Smallest = 95	
17	Q1 = 133	
18	Median = 143.5	
19	Q3 = 155	
20	Largest = 201	
21	IQR = 22	
22	Outliers: 201, 199, 190, 97, 95,	



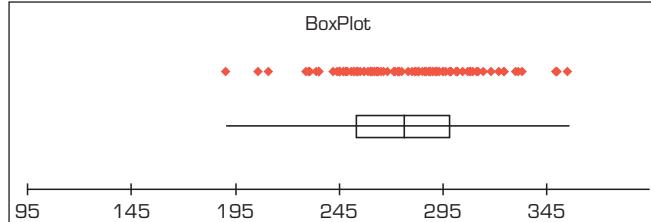
	A	B
28	Restaurant 3	
29	Smallest = 121	
30	Q1 = 136	
31	Median = 153	
32	Q3 = 177.5	
33	Largest = 223	
34	IQR = 41.5	
35	Outliers: None	



	A	B
41	Restaurant 4	
42	Smallest = 121	
43	Q1 = 141.25	
44	Median = 163	
45	Q3 = 207.25	
46	Largest = 338	
47	IQR = 66	
48	Outliers: 2338	



	A	B
54	Restaurant 5	
55	Smallest = 190	
56	Q1 = 253.25	
57	Median = 276.5	
58	Q3 = 297.5	
59	Largest = 355	
60	IQR = 44.25	
61	Outliers: None	



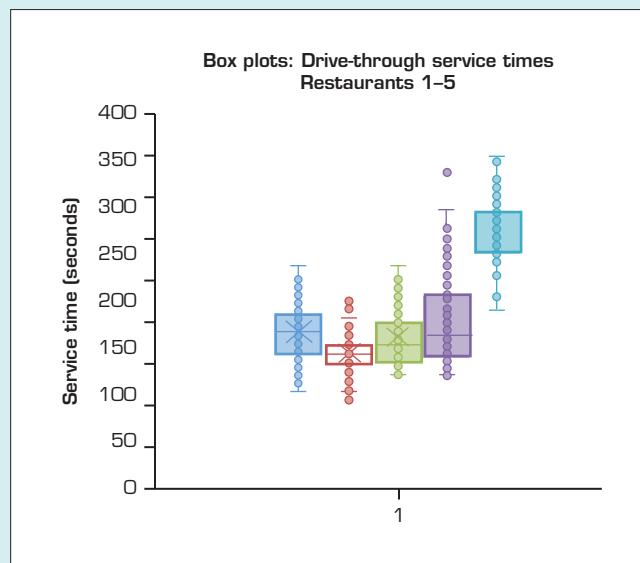
We will also use the computer to produce the box plots.





Using the computer

Excel box plots: Drive-through service times for restaurants 1–5



COMMANDS

Follow the Excel commands on page 175 with data file **XM05-18** and specify the input range as **A1:E101**.

Interpreting the results

The times for restaurant 2 appear to be the lowest and most consistent. The service times for restaurant 4 display considerably more variability. The slowest service times are for restaurant 5. The service times for restaurants 1, 2 and 5 seem to be symmetric, but the times for restaurants 3 and 4 are positively skewed. The box plots also identify some outliers in the data for restaurant 2 (small and large) and restaurant 4 (large).

5.3f Measures of relative standing and variability for ordinal data

Because the measures of relative standing are calculated by ordering the data, these statistics are appropriate for ordinal as well as for numerical data. Furthermore, because the interquartile range is calculated by taking the difference between the upper and lower quartiles, it too can be used to measure the variability of ordinal data.

Below are the factors that tell us when to use the techniques presented in this section.

IN SUMMARY

Factors that identify when to calculate percentiles and quartiles

- 1** *Objective:* to describe a set of data
- 2** *Type of data:* numerical or ordinal
- 3** *Descriptive measurement:* relative standing

IN SUMMARY

Factors that identify when to calculate the interquartile range

- 1 *Objective*: to describe a set of data
- 2 *Type of data*: numerical or ordinal
- 3 *Descriptive measurement*: variability

EXERCISES

Learning the techniques

- 5.61 XR05-61** Calculate the first, second and third quartiles of the following sample.

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

- 5.62 XR05-62** Find the third and eighth deciles (30th and 80th percentiles) of the following data set.

26	23	29	31	24	22	15	31	30	20
----	----	----	----	----	----	----	----	----	----

- 5.63 XR05-63** Find the first and second quintiles (20th and 40th percentiles) of the following data.

52	61	88	43	64	71	39	73	51	60
----	----	----	----	----	----	----	----	----	----

- 5.64 XR05-64** Determine the first, second and third quartiles of the following data.

10.5	14.7	15.3	17.7	15.9	12.2	10.0	
14.1	13.9	18.5	13.9	15.1	14.7		

- 5.65 XR05-65** Calculate the third and sixth deciles of the accompanying data.

7	18	12	17	29	18	4	27	
30	2	4	10	21	5	8		

- 5.66** Refer to Exercise 5.61. Determine the interquartile range.

- 5.67** Refer to Exercise 5.64. Determine the interquartile range.

- 5.68 XR05-68** Calculate the interquartile range for the following data:

5	8	14	6	21	11	9	10	18	2
---	---	----	---	----	----	---	----	----	---

Applying the techniques

- 5.69 Self-correcting exercises.** Refer to Exercise 5.44, which gives the percentage changes in value for

15 shares. Calculate the 20th percentile and the 60th percentile for these data.

- 5.70 XR05-70** Draw the box plot of the following set of data:

9	28	15	21	12	22	29	20
23	31	11	19	24	16	13	

- 5.71 XR05-71** Draw the box plot of the following set of data:

65	80	39	22	74	61	63	46	72	34
30	34	69	31	46	39	57	79	89	41

- 5.72** Refer to Example 5.11.

- a Draw the box plot for each sample (Trust).
- b Discuss the similarities and differences between the returns for Trust A and Trust B.

Computer applications

- 5.73 XR05-73** A sample of 100 observations were recorded.

- a Use a software package to draw the box plot.
- b What are the values of the quartiles?
- c What information can you extract from the box plot?

- 5.74 XR05-74** Refer to Example 5.6 for which the 100 test marks are recorded.

- a Draw the box plot.
- b What are the quartiles?
- c Are there any outliers?
- d What does the box plot tell you about the marks for the statistics exam?

- 5.75 XR05-75** Refer to Exercise 4.20, in which the number of customers entering a bank during each hour of operation for 100 days was recorded.

- a Draw the box plot for each hour of operation.
- b What are the quartiles, and what do they tell you about the number of customers arriving each hour?

5.76 XR05-76 The career-counselling centre at a university wanted to learn more about the starting salaries of the university's graduates. Each graduate was asked to report the highest salary offer they had received. The survey also asked each graduate to report the degree they had completed and their starting salary. Draw box plots to compare the four groups of starting salaries. Report your findings.

5.77 XR05-77 A random sample of marathon runners was drawn and their times (in minutes) to complete the race were recorded.

- a Draw the box plot.
- b What are the quartiles?
- c Identify any outliers.
- d What information does the box plot provide?

5.78 XR05-78 Do members of private golf courses play faster than players on public courses? The amount of time taken for a sample of private-course and public-course golfers was recorded.

- a Draw box plots for each sample.
- b What do the box plots tell you?
- c Are there any outliers in the sample?

5.79 XR05-79 For many restaurants, the amount of time customers linger over coffee and dessert negatively affects profits. To learn more about this variable, a sample of 200 restaurant groups was observed and the amount of time customers spent in the restaurant was recorded.

- a Calculate the quartiles of these data.
- b What do these statistics tell you about the amount of time spent in this restaurant?

5.80 XR05-80 The tourism industry sponsored a poll that asked a random sample of people how much they spent in preparation for their last holiday travel. Determine the quartiles and describe what they tell you.

5.4 Measures of association

In Chapter 2, we introduced the scatter diagram, a graphical technique that describes the relationship between two variables. In this section, we present two numerical measures of the strength and direction of the *linear* relationship depicted in a scatter diagram. The two measures of association are the *covariance* and the *coefficient of correlation*. Later in this section and again in Chapter 15, we will learn how to estimate a linear relationship between two variables, the *least squares line*, and another measure of association known as the *coefficient of determination* to assess the quality of the linear relationship between the two variables.

5.4a Covariance

As we did in Chapter 2, we label the two variables as X and Y . We denote the population means of variables X and Y as μ_x and μ_y , respectively, and N as the size of the population. If considering a sample, let \bar{X} and \bar{Y} be the sample means of variables X and Y respectively, and n be the number of pairs of observations in the sample. The *covariance* between X and Y is then defined as a measure of how two variables are linearly related.

covariance

A measure of how two variables are linearly related.

Covariance

$$\text{Population covariance} = \sigma_{xy} = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{N}$$

$$\text{Sample covariance} = s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

The denominator in the calculation of the sample covariance is $n - 1$, not the more logical n for the same reason we divide by $n - 1$ to calculate the sample variance (see page 157).

To illustrate how covariance measures association, consider the following three sets of sample data. Notice that the values of x are the same in all three sets, and that the values of y are the same but in a different order.

Set 1	x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	
	2	13	-3	-7	21	
	6	20	1	0	0	
	7	27	2	7	14	
$n = 3$	$\bar{x} = 5$	$\bar{y} = 20$			Sum = 35	$s_{xy} = 35/(3 - 1) = 17.5$

Set 2	x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	
	2	27	-3	7	-21	
	6	20	1	0	0	
	7	13	2	-7	-14	
$n = 3$	$\bar{x} = 5$	$\bar{y} = 20$			Sum = -35	$s_{xy} = -35/(3 - 1) = -17.5$

Set 3	x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	
	2	20	-3	0	0	
	6	27	1	7	7	
	7	13	2	-7	-14	
$n = 3$	$\bar{x} = 5$	$\bar{y} = 20$			Sum = -7	$s_{xy} = -7/(3 - 1) = -3.5$

In set 1, as x increases, so does y . In this case, when x is larger than its mean, y is at least as large as its mean. Thus, $(x_i - \bar{x})$ and $(y_i - \bar{y})$ have the same sign or zero, which means that the product is either positive or zero. Consequently, the covariance is a positive number. In general, if two variables move in the same direction (both increase or both decrease), the covariance will be a large positive number.

Next, consider set 2. As x increases, y decreases. Thus, the signs of $(x_i - \bar{x})$ and $(y_i - \bar{y})$ are opposite. As a result, the covariance is a negative number. If, as one variable increases, the other decreases, the covariance will be a large negative number.

Now consider set 3. As x increases, y exhibits no particular pattern. One product is positive, one is negative and the third is zero. Consequently, the covariance is a small number.

Generally speaking, if the two variables are unrelated (as one increases, the other shows no pattern), the covariance will be close to zero (either positive or negative).

If you plan to calculate the sample covariance manually, here is an alternative shortcut formula.

Shortcut formula for sample covariance

$$s_{xy} = \frac{1}{n-1} \left[\sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n} \right] \text{ or } s_{xy} = \frac{1}{n-1} \left[\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} \right]$$

As a measure of association, covariance suffers from a major drawback. It is usually difficult to judge the strength of the relationship from the covariance. For example, suppose that you have been told that the covariance of two variables is 250. What does this tell you about the relationship between the two variables? The sign, which is positive, tells you that as one increases, the other also generally increases. However, the degree to which the two variables move together is difficult to ascertain because we don't know whether 250 is a large number. To overcome this shortcoming, statisticians have produced another measure of association, which is based on the covariance. It is called the *coefficient of correlation*.

5.4b Coefficient of correlation

The **coefficient of correlation (Pearson)** is defined as the covariance divided by the standard deviations of the variables. Let X and Y be two random variables.

Population coefficient of correlation

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

where σ_{xy} is the population covariance, and σ_x and σ_y are the population standard deviations of X and Y respectively.

coefficient of correlation (Pearson)

A measurement of the strength and direction of a linear relationship between two numerical variables.

Sample coefficient of correlation

$$r = \frac{s_{xy}}{s_x s_y}$$

where s_{xy} is the sample covariance, and s_x and s_y are the sample standard deviations of X and Y respectively.

The population parameter is denoted by the Greek letter *rho*, ρ . The advantage that the coefficient of correlation has over the covariance is that the former has set lower and upper limits. The limits are -1 and $+1$ respectively.

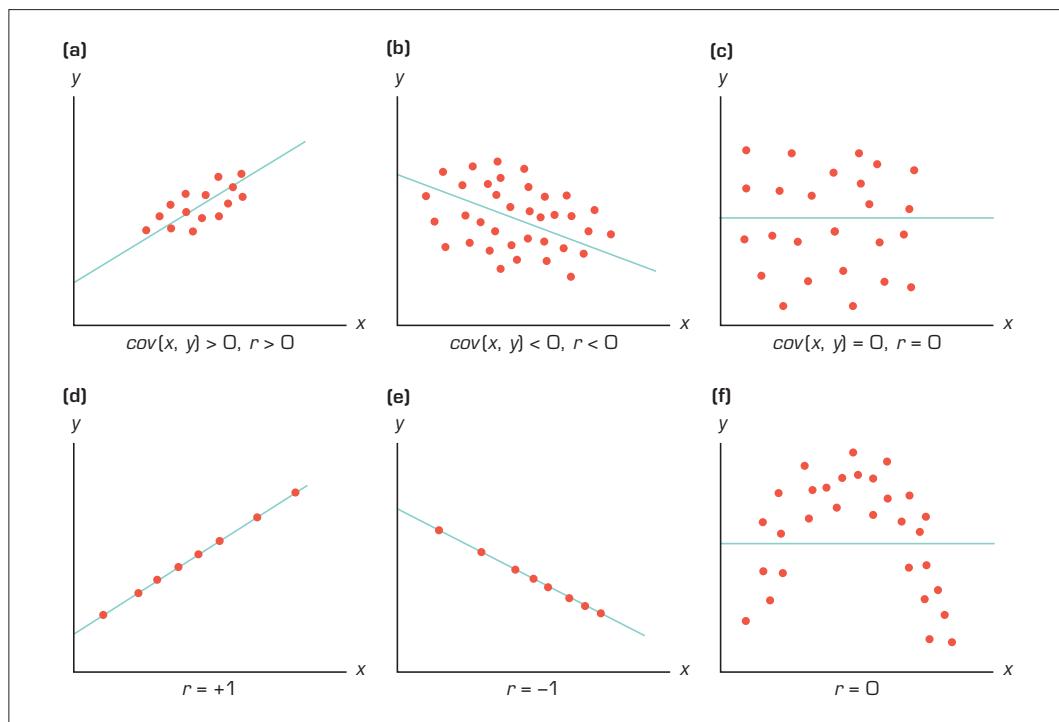
$$-1 \leq r \leq +1 \quad \text{and} \quad -1 \leq \rho \leq +1$$

The sign of the coefficient of correlation will be the same as the sign of the covariance and is interpreted in the same way. The strength of the association is gauged by the value of ρ or r . For example, a correlation close to $+1$ indicates two variables that are very strongly positively related. The closer the correlation is to 1 , the closer the relationship is to being described by a straight line sloping upwards from left to right. **Figure 5.8(a)** depicts a scatter diagram of one such case. When the coefficient of correlation equals $+1$, there is a perfect positive relationship. **Figure 5.8(d)** exhibits the behaviour of two variables that are positively perfectly correlated and have a correlation coefficient of $+1$.

A correlation close to -1 tells us that there is a strong negative relationship – as one variable increases, the other decreases. See **Figure 5.8(b)** for an illustrative scatter diagram. A perfect straight line sloping downwards would produce a correlation of -1 . **Figure 5.8(e)** depicts a scatter diagram of two perfectly negatively correlated variables.

A correlation close to zero indicates that no straight-line relationship exists. It may mean no pattern, such as the scatter diagram depicted in **Figure 5.8(c)**, or a relationship that is not a straight line as seen in **Figure 5.8(f)**. When the coefficient of correlation equals 0, there is no linear relationship. All other values of correlation are judged in relation to the three values.

FIGURE 5.8 Covariance and correlation for various scatter diagrams



The drawback to the coefficient of correlation is that except for three values -1 , 0 and $+1$, we cannot interpret the correlation. For example, suppose we calculated the coefficient of correlation to be -0.4 . What does this tell us? It tells us two things. The minus sign tells us that the relationship is negative and since 0.4 is closer to 0 than to 1 , we judge that the linear relationship is weak. In many applications we need a better interpretation than the ‘linear relationship is weak’. Fortunately, there is yet another measure of the strength of a linear relationship, which gives us more information. It is the *coefficient of determination*, which we introduce later in this section.

EXAMPLE 5.19

LO7

Calculating the coefficient of correlation

Calculate the coefficient of correlation for data sets 1–3 on page 180.

Solution

We have already calculated the means and the covariances (page 180):

$$\begin{aligned} \text{Set 1: } & \bar{x} = 5; \bar{y} = 20; \text{cov}(x,y) = s_{xy} = 17.5 \\ \text{Set 2: } & \bar{x} = 5; \bar{y} = 20; \text{cov}(x,y) = s_{xy} = -17.5 \\ \text{Set 3: } & \bar{x} = 5; \bar{y} = 20; \text{cov}(x,y) = s_{xy} = -3.5 \end{aligned}$$

Now we need to calculate only the standard deviations (s_x and s_y) of X and Y .

$$\text{Set 1: } \bar{x} = \frac{2+6+7}{3} = 5.0$$





$$\bar{y} = \frac{13+20+27}{3} = 20.0$$

$$s_x^2 = \frac{(2-5)^2 + (6-5)^2 + (7-5)^2}{3-1} = \frac{9+1+4}{2} = 7.0$$

$$s_y^2 = \frac{(13-20)^2 + (20-20)^2 + (27-20)^2}{3-1} = \frac{49+0+49}{2} = 49.0$$

The standard deviations are:

$$s_x = \sqrt{7.0} = 2.65$$

$$s_y = \sqrt{49.0} = 7.00$$

As the $(x_i - \bar{x})$ values are the same for all three data sets, their standard deviations are also the same: $s_x = 2.65$ for all three. Similarly, the standard deviations of Y are also the same: $s_y = 7.00$ for all three data sets.

The coefficients of correlation are:

$$\text{Set 1: } r = \frac{s_{xy}}{s_x s_y} = \frac{17.5}{(2.65)(7.0)} = 0.943$$

$$\text{Set 2: } r = \frac{s_{xy}}{s_x s_y} = \frac{-17.5}{(2.65)(7.0)} = -0.943$$

$$\text{Set 3: } r = \frac{s_{xy}}{s_x s_y} = \frac{-3.5}{(2.65)(7.0)} = -0.189$$

It is now easier to see the strength as well as the direction of the linear relationship between X and Y .

EXAMPLE 5.20

LO6

Construction price versus house size

XM05-20 Refer to Example 4.5. A real estate agent wanted to know to what extent the construction price of a house is related to the size (number of squares) of the house. He took a sample of 15 houses that were recently built in an outer suburb in Perth, and recorded the construction price and the size of each. These data are listed in **Table 5.6**. Calculate the covariance and the coefficient of correlation, to measure how the two variables are related.

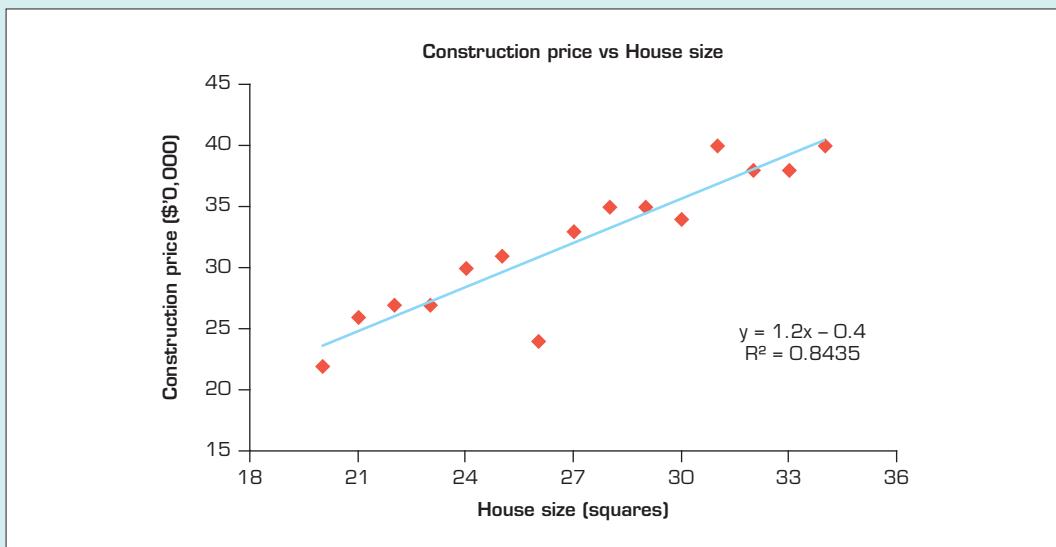
TABLE 5.6 House size and construction price

House size (squares)	Construction price (\$'0000)	House size (squares)	Construction price (\$'0000)
20	22	29	35
21	26	30	34
31	40	26	24
32	38	33	38
24	30	27	33
25	31	34	40
22	27	28	35
23	27		

Solution

Visualising graphically

We begin by constructing a scatter diagram of construction price against the house size as given below.



The scatter diagram clearly shows a strong, positive linear relationship between the two variables. In order to measure the strength of this relationship, we calculate the coefficient of correlation.

Calculating manually

TABLE 5.7 Calculations for Example 5.20

House	x	y	x ²	y ²	xy
1	20	22	400	484	440
2	21	26	441	676	546
3	31	40	961	1600	1240
4	32	38	1024	1444	1216
5	24	30	576	900	720
6	25	31	625	961	775
7	22	27	484	729	594
8	23	27	529	729	621
9	29	35	841	1225	1015
10	30	34	900	1156	1020
11	26	24	676	576	624
12	33	38	1089	1444	1254
13	27	33	729	1089	891
14	34	40	1156	1600	1360
15	28	35	784	1225	980
Total	405	480	11215	15838	13296

We then calculate the sample means and standard deviations of X and Y. They are:

$$\bar{x} = 27.0, \quad s_x = 4.47$$

$$\bar{y} = 32.0, \quad s_y = 5.84$$

We apply the shortcut formula on page 180 to calculate the covariance between X and Y and the coefficient of correlation. This requires the following sums: Σx , Σy , Σx^2 , Σy^2 and Σxy . **Table 5.7** describes these calculations.

Thus, the covariance is:

$$s_{xy} = \frac{1}{n-1} \left[\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n} \right]$$

$$= \frac{1}{14} \left[13296 - \frac{405 \times 480}{15} \right] = \frac{336}{14} = 24$$

The coefficient of correlation is:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{24}{4.47 \times 5.84} = 0.918$$

Interpreting the results

The covariance provides very little useful information other than telling us that the two variables are positively related. The coefficient of correlation informs us that there is a strong positive relationship. This information can be extremely useful to real estate agents, insurance brokers and all potential home purchasers.

Using the computer

Excel output for Example 5.20

Variance–covariance matrix (population)⁴

	A	B	C
1		House size	Price
2	House size	18.667	
3	Price	22.400	31.867

Correlation matrix

	A	B	C
1		House size	Price
2	House size	1	
3	Price	0.9184	1

COMMANDS

- Type data in two columns or open the data file (**XM05-20**).
- Click **DATA**, **Data Analysis** and **Covariance**. Click **OK**.
- Specify the coordinates of the data (**A1:B16**). Click the square for **Labels in First Row**.
- To store the output on the same Excel worksheet, click the circle for **Output Range** and type the output starting cell reference (**C1**). Click **OK**.

To calculate the coefficient of correlation, repeat the steps above, but click **Correlation** instead of **Covariance**.

This method is very useful when you have two variables and you would like to compute the coefficient of correlation and/or the covariance for each pair of variables, to produce the correlation matrix and the variance–covariance matrix.

Alternatively, to calculate the sample covariance, you can use the inbuilt statistical functions by typing the following in an empty cell:

=COVARIANCE.S([Input range of one variable];[Input range of the second variable])

Replace **COVARIANCE.S** by **CORREL** to calculate the coefficient of correlation.

For the above example, we would enter =COVARIANCE.S(A2:A16; B2:B16) and =CORREL(A2:A16; B2:B16). Sample variance can be calculated using the function **VAR.S(array)**.

For the sample covariance, we multiply the population covariance from the output by $(n/n - 1)$. That is, we observe that sample covariance $s_{xy} = 22.4 \times (15/14) = 24$. The coefficient of correlation $r = 0.9184$.

4 Unfortunately Excel calculates the population variances and covariance. If you are using sample data, you can convert these population parameters to sample statistics by multiplying the variance–covariance matrix entries by $n/(n - 1)$. However, sample covariance and variance can be calculated using the alternative method with functions COVARIANCE.S and VAR.S respectively. (Population values can be obtained using COVARIANCE.P and VAR.P.)

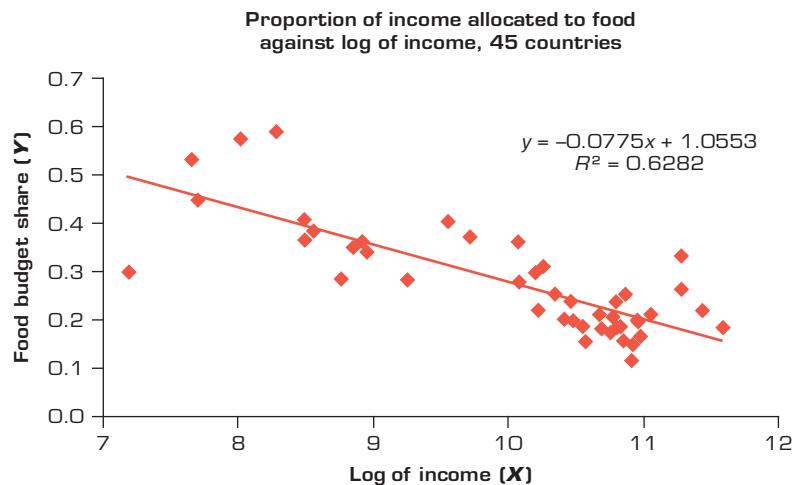
SPOTLIGHT ON STATISTICS

Income and its allocation to food: Solution

Now we can provide a solution to the opening example. Working and Leser modelled Engel's law in the form of a linear relationship between the proportion of income allocated to food and the logarithm of income. We define the logarithm of income as X and the proportion of income allocated to food (budget share) as Y . Using data for 45 countries for the two variables, first we present a scatter diagram and then calculate the coefficient of correlation. As can be seen from the scatter diagram, the points are scattered around a downwards sloping straight line, and therefore there is a negative linear relationship between Y and X . We can perform the manual calculation or use Excel to obtain the coefficient of correlation between the log of income and the proportion of income allocated to food. This correlation value is -0.793 . This means that there is a strong negative linear relationship between the log of income and the proportion of income allocated to food. That is, the food share of consumers' income falls with increasing income.



Source: iStock.com/thorbjorn66



5.4c Interpreting correlation

Because of its importance, we remind you about the correct interpretation of the analysis of the relationship between two numerical variables that we discussed in Chapter 4. In other words, if two variables are linearly related, it does not mean that X causes Y . It may mean that another variable causes both X and Y or that Y causes X . Remember: correlation is not causation.

5.4d Estimating the linear relationship

When we presented the scatter diagram in Section 4.3, we pointed out that we were interested in measuring the strength and direction of the *linear relationship*. Both can be judged more easily by drawing a straight line through the data. However, if different people draw a line through the same data set, it is likely that each person's line will differ from all the others. Moreover, we often need to know the equation of the line. Consequently, we need an objective method of producing a straight line. Such a method has been developed; it is called the **least squares method**.

least squares method

A method of deriving an estimated linear equation (straight line) which best fits the data.

The least squares method produces a straight line drawn through the points so that the sum of squared deviations between the points and the line is minimised. The line is represented by the following equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where $\hat{\beta}_0$ is the y -intercept (where the line intersects the y -axis), $\hat{\beta}_1$ is the slope (defined as rise/run), and \hat{y} is the value of y determined by the line. The coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ are derived using calculus so that we minimise the sum of squared errors (SSE):

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Estimated least squares line coefficients

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

As $\hat{\beta}_1$ is always required for the calculation of $\hat{\beta}_0$, $\hat{\beta}_1$ is always calculated first.

REAL-LIFE APPLICATIONS

Breakeven analysis

Breakeven analysis is an extremely important business tool, one that you will likely encounter repeatedly in your course of studies. It can be used to determine the volume of sales your business needs to start making a profit.

Breakeven analysis is especially useful when managers are attempting to determine the appropriate price for the company's products and services.

A company's profit can be calculated simply as

$$\text{Profit} = (\text{Price per unit} - \text{Variable cost per unit}) \times (\text{Number of units sold}) - \text{Fixed costs}$$

The breakeven point is the number of units sold such that the profit is 0. Thus, the breakeven point is calculated as:

$$\text{Number of units sold} = \frac{\text{Fixed cost}}{\text{Price} - \text{Variable cost}}$$

Managers can use the formula to help determine the price that will produce a profit. However, to do so requires knowledge of the fixed and variable costs. For example, suppose that a bakery sells only loaves of bread. The bread sells for \$3.20 a loaf, the variable

cost is \$1.20, and the fixed annual costs are \$25 000. The breakeven point is:

$$\text{Number of units sold} = \frac{25000}{3.20 - 1.20} = 12500$$

The bakery must sell more than 12 500 loaves per year to make a profit.

In the next application box we discuss fixed and variable costs.



Source: Shutterstock.com/Vladimir Gerasimov

REAL-LIFE APPLICATIONS

Fixed and variable costs

Fixed costs are costs that must be paid whether or not any units are produced. These costs are 'fixed' over a specified period of time or range of production. Variable costs are costs that vary directly with the number of products produced. For the previous bakery

example, the fixed costs would include rent and maintenance of the shop, wages paid to employees, advertising costs, telephone, and any other costs that are not related to the number of loaves baked. The variable cost is primarily the cost of ingredients, which rises in relation to the number of loaves baked.

EXAMPLE 5.21

LO7

Estimating fixed and variable costs

XM05-21 A tool and die maker operates out of a small shop making specialised tools. He is considering increasing the size of his business and needs to know more about his costs. One such cost is electricity, which he needs to operate his machines and lights. (Some jobs require that he turn on extra-bright lights to illuminate his work.) He keeps track of his daily electricity costs and the number of tools he made that day. These data are listed below. Determine the fixed and variable costs of electricity.

Day	1	2	3	4	5	6	7	8	9	10
Number of tools	7	3	2	5	8	11	5	15	3	6
Electricity cost (\$)	23.80	11.89	15.98	26.11	31.79	39.93	12.27	40.06	21.38	18.65

Solution

The dependent variable (y) is the daily cost of electricity and the independent variable (x) is the number of tools made in a day. To calculate the coefficients of the least squares line and other statistics (calculated below) we need the sum of x , y , xy , x^2 and y^2 .

Day	x	y	xy	x^2	y^2
1	7	23.80	166.6	49	566.44
2	3	11.89	35.67	9	141.37
3	2	15.98	31.96	4	255.36
4	5	26.11	130.55	25	681.73
5	8	31.79	254.32	64	1010.60
6	11	39.93	439.23	121	1594.40
7	5	12.27	61.35	25	150.55
8	15	40.06	600.90	225	1604.80
9	3	21.38	64.14	9	457.10
10	6	18.65	111.90	36	347.82
Total	65	241.86	1896.62	567	6810.20

$$\sum x_i = 65; \sum y_i = 241.86; \sum x_i y_i = 1896.62; \sum y_i^2 = 6810.20$$

Covariance

$$s_{xy} = \frac{1}{n-1} \left[\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n} \right] = \frac{1}{10-1} \left[1896.62 - \frac{(65)(241.86)}{10} \right] = 36.06$$





Variance of x

$$s_x^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \right] = \frac{1}{10-1} \left[567 - \frac{(65)^2}{10} \right] = 16.06$$

Sample means

$$\bar{x} = \frac{\sum x_i}{n} = \frac{65}{10} = 6.5$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{241.86}{10} = 24.19$$

The coefficients of the least squares line are:

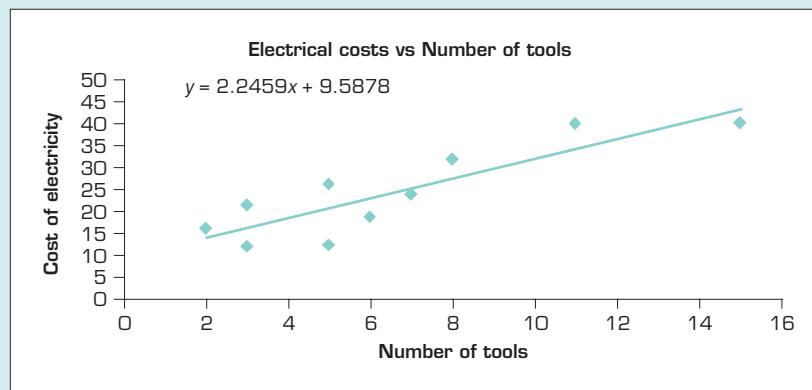
$$\text{Slope: } \hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{36.06}{16.06} = 2.25$$

$$y\text{-intercept: } \bar{y} = \hat{\beta}_1 \bar{x} = 24.19 - (2.25)(6.5) = 9.57$$

The least squares line: $\hat{y} = 9.57 + 2.25x$

Using the computer

Excel output for Example 5.21



COMMANDS

- 1 Type the data in two columns where the first column stores X and the second stores Y, or open the data file (**XMO5-21**).
- 2 Highlight the columns containing the data (including column titles) (**A1:B11**) and follow the commands to draw a **Scatter diagram** (page 113).
- 3 Right click on any of the data points in the scatter diagram, click Add **Trendline...** and then select **Linear** from the **Trendline Options** menu that appears on the right-hand side of the screen.
- 4 At the bottom of the **Trendline Options** menu, click **Display equation on Chart**. Click the cross at the top of the menu to close it.

Interpreting the results

The slope is defined as rise over run, which means that it is the change in y (rise) for a 1-unit increase in x (run). Put less mathematically, the slope measures the *marginal* rate of change in the dependent variable. The marginal rate of change refers to the effect of increasing the independent variable by one additional unit. In this example



the slope is 2.25, which means that in this sample, for each 1-unit increase in the number of tools made in a day, the marginal increase in the electricity cost is \$2.25. Thus, the estimated variable cost is \$2.25 per tool.

The y -intercept is 9.59. That is, the line strikes the y -axis at 9.59. This is simply the value of y when $x = 0$. However, when $x = 0$, we are producing no tools and hence the estimated fixed cost of electricity is \$9.59 per day.

Because the costs are estimates based on a straight line, we often need to know how well the line fits the data.

EXAMPLE 5.22

LO7

Measuring the strength of the linear relationship

Calculate the coefficient of correlation for Example 5.21.

Solution

To calculate the coefficient of correlation, we need the covariance between X and Y and the standard deviations of both variables. The covariance between X and Y and the variance of X were calculated in Example 5.21.

The covariance is

$$s_{xy} = 36.06$$

Variance of X is

$$s_x^2 = 16.06$$

Standard deviation of X is

$$s_x = \sqrt{s_x^2} = \sqrt{16.06} = 4.01$$

All we need is the standard deviation of Y .

$$s_y^2 = \frac{1}{n-1} \left[\sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i \right)^2}{n} \right] = \frac{1}{10-1} \left[6810.20 - \frac{(241.86)^2}{10} \right] = 106.73$$

$$s_y = \sqrt{s_y^2} = \sqrt{106.73} = 10.33$$

The coefficient of correlation is

$$r = \frac{s_{xy}}{s_x s_y} = \frac{36.06}{(4.01)(10.33)} = 0.871$$

Interpreting the results

The coefficient of correlation is 0.871, which tells us that there is a positive linear relationship between the number of tools made and the electricity cost. The coefficient of correlation tells us that the linear relationship is quite strong and thus the estimates of fixed and variable costs should be reasonably accurate.

Using the computer

Excel output for Example 5.22

Variance–covariance matrix (sample)

	A	B	C
1		Number of tools	Electrical costs
2			
3	Number of tools	16.056	
4	Electrical costs	36.059	106.730





Correlation matrix

	A	B	C
1		Number of tools	Electrical costs
2			
3	Number of tools		1
4	Electrical costs	0.8711	1

COMMANDS

Follow the commands used in Example 5.20 (page 183). Note that Excel generates the covariance and variances as population parameters. You need to multiply by $n/(n - 1)$ to obtain the sample statistics. Alternatively, we can use the COVARIANCE.S and VAR.S functions to obtain the sample covariance and variances individually.

5.4e Coefficient of determination

When we introduced the coefficient of correlation (page 181) we pointed out that except for -1 , 0 and $+1$ we cannot precisely interpret its meaning. We can judge the coefficient of correlation in relation to its proximity to only -1 , 0 and $+1$. We also pointed out that we have another measure that can be precisely interpreted. It is the *coefficient of determination*, which is calculated by squaring the coefficient of correlation. For this reason we denote it R^2 .

The **coefficient of determination** measures the amount of variation in the dependent variable that is explained by the independent variable. For example, if the coefficient of correlation is -1 or $+1$, a scatter diagram would display all the points lining up in a straight line. The coefficient of determination is 1 , which we interpret to mean that 100% of the variation in the dependent variable Y is explained by the variation in the independent variable X . If the coefficient of correlation is 0 , then there is no linear relationship between the two variables, and $R^2 = 0$: none of the variation in Y is explained by the variation in X .

In Example 5.22 the coefficient of correlation was calculated to be $r = 0.871$. Thus, the coefficient of determination is

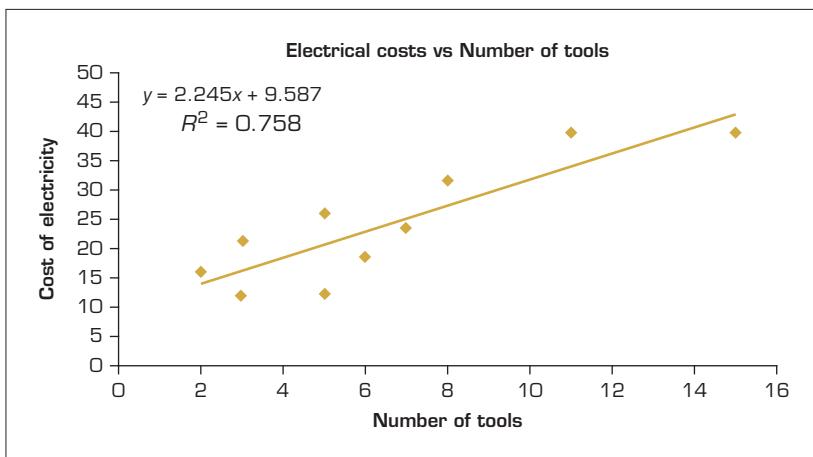
$$R^2 = (0.871)^2 = 0.7588$$

This tells us that 75.88% of the variation in electrical costs is explained by the number of tools made in a day. The remaining 24.12% is unexplained.

coefficient of determination

The proportion of the amount of variation in the dependent variable that is explained by the independent variable.

Using the computer



COMMANDS

You can use Excel to calculate the coefficient of correlation and then square the result. Alternatively, use Excel to draw the least squares line, called the trendline. After doing so, click on the chart area. Right click on any data point and select **Add Trendline...**. Then select **Linear** under **TRENDLINE OPTIONS** in the menu that appears on the right-hand side of the screen. Tick the boxes for **Display equation on chart** and **Display R-squared value on chart**. Click the cross to close the menu.

The concept of explained variation is an extremely important one in statistics. We return to this idea repeatedly in Chapters 15–16. In Chapter 15, we explain why we interpret the coefficient of determination in the way that we do.

We complete this section with a review of when to use the techniques introduced in this section.

IN SUMMARY

Factors that identify when to calculate covariance and coefficient of correlation

- 1 *Objective:* to describe the relationship between two variables
- 2 *Type of data:* numerical (quantitative)

EXERCISES

Applying the techniques

- 5.81 XR05-81** Consider the following sample of five observations of variables X and Y .

X	8	14	12	20	10
Y	20	26	28	34	24

- a Draw a scatter diagram of Y against X and comment on the resulting plot.
- b Manually calculate the covariance and the coefficient of correlation.
- c Determine the least squares line and interpret the coefficients.

- 5.82 XR05-82** Are the marks received in a course related to the amount of time spent studying the subject? To analyse the mysterious possibility, a student took a random sample of 10 students who had enrolled in an accounting class last semester. He asked each to report his or her mark in the course and the total number of hours spent studying accounting. These data are listed here.

Marks	79	65	81	88	53	80	85	92	67	49
Time spent studying	45	47	42	52	30	49	46	53	40	33

- a Plot the sample observations in a scatter diagram and comment on the pattern.
- b Calculate the covariance.

- c Calculate the coefficient of correlation.
- d Calculate the coefficient of determination.
- e Determine the least squares line.
- f What do the statistics calculated above tell you about the relationship between marks and study time?

- 5.83** Refer to Exercise 4.41, which considered the relationship between the All Ordinaries Index (Y) and inflation (X). The annual All Ordinaries Index and the annual inflation rates from 1995 to 2018 are stored in columns 1 and 2 respectively of file **XR04-41**.

- a Calculate covariance (s_{xy}) between X and Y and the coefficient of correlation r .
- b What do these statistics tell you about the relationship between the All Ordinaries Index and inflation?
- c Determine the least squares line.
- d Determine the coefficient of determination R^2 .

- 5.84** Refer to Exercise 4.42, which considered the relationship between a manufacturing firm's cost of electricity (Y) and hours of machine time (X). Data for X and Y were recorded for each of the 12 months (**XR04-42**).

- a Graphically describe the relationship between the firm's cost of electricity and hours of machine time.
- b Calculate the covariance and the coefficient of correlation between X and Y .

- c What do these statistics tell you about the relationship between machine time and the cost of electricity?
- d Conduct an analysis of the relationship between the cost of electricity and hours of machine time. What does this analysis tell you?

5.85 XR05-85 The owner of a furniture store was attempting to analyse the relationship between advertising and sales, and recorded the monthly advertising budget (\$'000) and the sales (\$m) for a sample of 12 months. The data are listed here.

Advertising	23	46	60	54	28	33
	25	31	36	88	90	99
Sales	9.6	11.3	12.8	9.8	8.9	12.5
	12.0	11.4	12.6	13.7	14.4	15.9

- a Use a graphical technique to present these numbers.
- b Calculate the covariance of the two variables.
- c Determine the coefficient of correlation.
- d What do these statistics tell you about the relationship between advertising and sales?
- e Determine the least squares line.
- f Determine the coefficient of determination R^2 .

Computer applications

5.86 XR05-86 The unemployment rate is an important measure of the economic health of a country. The unemployment rate measures the percentage of people who are looking for work and who are without jobs. Another way of measuring this economic variable is to calculate the employment rate, which is the percentage of adults who are employed. The unemployment rates and employment rates of 19 countries published in the *National Post Business* were recorded. Calculate the

coefficient of determination and describe what you have learnt.

5.87 Refer to Exercise 4.46, which considered the relationship between the Perth Mint's annual (end of financial year) Australian gold and silver spot prices using data over a 16-year period (2000–18). The year is stored in column 1, and gold price and silver price are stored in columns 2 and 3 respectively of file **XR04-46**.

- a Calculate covariance and the correlation (r) between the two variables.
- b What do these statistics tell you about the relationship between the daily gold price and the daily silver price?

5.88 XR05-88 Besides the known long-term effects of smoking, do cigarettes also cause short-term illnesses such as colds? To help answer this question, a sample of smokers was drawn. Each person was asked to report the average number of cigarettes smoked per day and the number of sick days taken due to colds last year. The data are stored in the file.

- a Calculate the covariance of the two variables.
- b Determine the coefficient of correlation.
- c What do these statistics tell you about the relationship between smoking cigarettes and the incidence of colds?

5.89 XR05-89 A manufacturing firm produces its products in batches using sophisticated machines and equipment. The general manager wanted to investigate the relationship between direct labour costs and the number of units produced per batch. He recorded the data from the last 30 batches.

- a Determine the strength and direction of the relationship.
- b Determine the fixed and variable labour costs.

5.5 General guidelines on the exploration of data

The purpose of applying graphical and numerical techniques is to describe and summarise data. Statisticians usually apply graphical techniques as a first step because we need to know the shape of the distribution. The shape of the distribution helps to answer the following questions.

- 1 Where is the approximate centre of the distribution?
- 2 Are the observations close to one another, or are they widely dispersed?
- 3 Is the distribution unimodal, bimodal or multimodal? If there is more than one mode, where are the peaks and where are the valleys?
- 4 Is the distribution symmetric? If not, is it skewed? If symmetric, is it bell shaped?

Histograms, stem-and-leaf displays and box plots provide most of the answers. We can frequently make several inferences about the nature of the data from the shape. For example, we can assess the relative risk of investments by noting their spreads, or improve the teaching of a course by examining whether the distribution of final grades is bimodal or skewed.

The shape can also provide some guidance on which numerical techniques to use. As we noted in this chapter, the central location of highly skewed data may be more appropriately measured by the median. We may also choose to use the interquartile range instead of the standard deviation to describe the spread of skewed data.

When we have an understanding of the structure of the data, we may proceed to further analysis. For example, we often want to determine how one variable (or several variables) affects another. Scatter diagrams, covariance and the coefficient of correlation are useful techniques for detecting relationships between variables. A number of techniques to be introduced later in this book will help to uncover the nature of these associations.

Study Tools

CHAPTER SUMMARY

This chapter extended our discussion of descriptive statistics, which deals with methods of summarising and presenting the essential information contained in a set of data, whether the set is a population or a sample taken from a population. After constructing a frequency distribution to obtain a general idea about the distribution of a data set, we can use numerical measures to describe the *central location* and the *dispersion* of the data. Three popular measures of central location, or averages, are the *mean*, the *median* and the *mode*. Taken by themselves, these measures provide an inadequate description of the data because they say nothing about the extent to which the data are dispersed about their central value. Information regarding the dispersion, or variability, of the data is conveyed by such numerical measures as *range*, *variance*, *standard deviation* and *coefficient of variation*.

Chebyshev's theorem describes how the mean and the standard deviation of a set of observations can be combined to determine the minimum proportion of observations that lie within various intervals centred at the mean. For the special case in which a sample of observations has a *mound-shaped* distribution, the *empirical rule* provides a good approximation of the percentages of observations that fall within one, two or three standard deviations of the mean. Beginning in Chapter 10, we will discuss how these two important descriptive measures (mean and standard deviation), calculated for a sample of observations, can be combined to support inferences about the mean and the standard deviation of the population from which the sample was taken.

This chapter also included an introduction to *box plots*, which can be used to identify outliers in a data set. *Covariance* and *coefficient of correlation*, which are used to measure the strength of a linear relationship between two variables, and a brief introduction to the estimation of a *linear relationship* between two variables using the *least squares method*, are also provided in this chapter.

The data files for Examples, Exercises and Cases are provided on the companion website (accessible through <https://login.cengagebrain.com/>).

SYMBOLS

Symbol	Pronounced	Represents
μ	<i>Mu</i>	Population mean
σ	<i>Sigma</i>	Population standard deviation
σ^2	<i>Sigma squared</i>	Population variance
ρ	<i>Rho</i>	Population coefficient of correlation
Σ	<i>Sum of</i>	Summation
$\sum_{i=1}^n x_i$	<i>Sum of x_i from $i = 1$ to n</i>	Summation of n numbers

SUMMARY OF FORMULAS

Population mean	$\mu = \frac{\sum_{i=1}^N x_i}{N}$
Sample mean	$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$

Population variance	$\sigma^2 = \frac{1}{N} \left[\sum_{i=1}^N x_i^2 - N\mu^2 \right]$
Sample variance	$s^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right]$
Population standard deviation	$\sigma = \sqrt{\sigma^2}$
Sample standard deviation	$s = \sqrt{s^2}$
Population coefficient of variation	$CV = \frac{\sigma}{\mu}$
Sample coefficient of variation	$cv = \frac{s}{\bar{x}}$
Approximate sample standard deviation	$s \approx \frac{\text{Range}}{4}$
Population covariance	$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$
Sample covariance	$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$
Population coefficient of correlation	$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$
Sample coefficient of correlation	$r = \frac{s_{xy}}{s_x s_y}$
Least squares line	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ where $\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
Coefficient of determination	$R^2 = r^2$ (r = coefficient of correlation)

SUPPLEMENTARY EXERCISES

5.90 XR05-90 Osteoporosis is a condition in which bone density decreases, often resulting in broken bones. Bone density usually peaks at age 30 and decreases thereafter. To understand more about the condition, a random sample of women aged 50 and over was recruited. Each woman's bone density loss was recorded.

- a Compute the mean and the median of these data.
- b Compute the standard deviation of the bone density losses.
- c Describe what you have learned from the statistics.

5.91 XR05-91 The underemployment rates of the Australian states and territories during September 2018 for males and females aged 15 and over are presented below. Calculate the mean, standard deviation and coefficient of variation of the underemployment rates for males and females. Discuss the variability in

underemployment between the Australian states and territories for males and females during September 2018.

State/Territory	Male	Female
New South Wales	5.7	9.9
Victoria	5.6	9.4
Queensland	6.4	10.4
South Australia	6.3	11.8
Western Australia	7.1	12.0
Tasmania	7.5	12.9
Northern Territory	3.8	4.4
Australian Capital Territory	6.2	7.2

Source: Australian Bureau of Statistics, *Labour Force Australia*, September 2018, cat. no. 6202.0, ABS, Canberra, Labour Underutilisation (aged 15 and over).

5.92 Find the sample mean (\bar{x}), the median, the mode, the sample variance (s^2) and the sample standard deviation (s) for the data in Exercise 4.55.

5.93 Refer to Exercises 4.55 and 5.92.

- a What proportion of items falls in the interval $\bar{x} \pm 2s$?
- b Does it appear that the population from which this sample was taken has a mound-shaped distribution?
- c Compare the actual proportion of items falling into the intervals $(\bar{x} - s, \bar{x} + s)$ and $(\bar{x} - 2s, \bar{x} + 2s)$ with the proportions suggested by the empirical rule.

5.94 XR05-94 The dividend yields for 15 listed property trusts in Australia are recorded as percentages below.

7.34	7.53	8.62	5.95	9.20
7.75	6.40	6.80	5.50	9.75
8.35	6.84	7.43	7.77	8.92

- a Calculate the mean and the standard deviation of the 15 yields.
- b Construct a relative frequency histogram for the yields.
- c Find the median of the 15 yields, and locate the mean and the median on your histogram.

5.95 In Exercise 4.9, the ages of a sample of 25 stockbrokers were recorded (XR04-09) in the table.

50	64	32	55	41	44	24	46	58
47	36	52	54	44	66	47	59	51
61	57	49	28	42	38	45		

- a Find the median age for the sample data.
- b Find the lower quartile of the ages.
- c Find the upper quartile of the ages.
- d Find the 80th percentile of the ages.
- e Does the firm have reason to be concerned about the distribution of the ages of its brokers?

5.96 Refer to Exercise 5.95.

- a Calculate the mean of the sample data.
- b Calculate the variance of the sample data.
- c Calculate the standard deviation of the sample data.
- d Calculate the range of the data.
- e Calculate the range approximation for the standard deviation of the data.

5.97 Refer to Exercise 5.95.

- a Construct a frequency distribution for the data, using five class intervals and the value 20 as the lower limit of the first class. (If you did Exercise 4.9b, you have already constructed this histogram.)
- b Approximate the mean and the variance of the ages, based on the frequency distribution constructed in part (a). (See Appendix 5.A)
- c Construct a relative frequency histogram for the data, using five class intervals and the value 20 as the lower limit of the first class. (If you did Exercise 4.9c, you have already constructed this histogram.)
- d Locate the interval $\bar{x} \pm s$ on the histogram, and find the proportion of ages that fall in this interval. How does this proportion compare with the empirical rule approximation?
- e Construct a box plot for the brokers' ages.
- f Does the distribution of the ages appear to be symmetric or skewed? Explain.

5.98 XR05-98 In the past few years, car makers have significantly improved the quality of their products. To determine the degree of improvement, a survey of 200 new cars was undertaken. The number of minor flaws (for example, slightly misaligned doors, improperly directed headlights) was recorded. These data, and the data from a similar survey undertaken five years earlier, are stored in column 1 and column 2 respectively.

- a Use a graphical technique of your choosing to compare the two sets of data.
- b Use numerical techniques of your choosing to compare the two sets of data.
- c Briefly describe what you have learned from your statistical analysis.

5.99 XR05-99 The following is a sample of duplex house prices (in '\$000):

474	629	429	635	460
422	492	619	442	402
435	415	590	559	609
575	409	465	640	565
519	538	614	449	479

- a Depict graphically the distribution of the sample data.
- b Calculate the variance and the standard deviation of this sample of prices.

- c** Compare the range approximation of s to the true value of s . Explain why you would or would not expect the approximation to be a good one for this sample.

5.100 XR05-100 The three-month returns (in percentages) for 30 shares are given in the table below.

4.73	4.28	1.03	6.27	5.76	5.28
4.75	3.42	5.75	5.3	3.19	9.07
6.99	8.18	6.91	4.39	12.78	0.84
4.67	4.18	5.93	5.4	6.07	5.69
5.78	4.77	3.80	0.79	6.51	4.38

- a** Calculate the mean and the standard deviation of the returns.
b Construct a relative frequency histogram for the returns.
c Determine the median returns and locate the mean and the median on your histogram.
d Use your relative frequency distribution to estimate the mean and the standard deviation of the 30 returns, and compare your estimates with the values obtained in part (a).

5.101 XR05-101 The closing prices of 30 top shares are shown in the table below.

14.23	29.41	1.96	4.0	3.12	25.05
4.43	34.81	8.67	4.48	17.50	26.10
7.87	7.10	47.40	3.50	8.54	23.36
4.49	57.48	9.80	4.61	1.48	11.06
25.25	13.28	1.0	9.85	31.22	25.24

- a** Calculate the mean and the standard deviation of the sample of 30 closing prices.
b Construct a relative frequency histogram for the returns.

5.102 XR05-102 The volatility measure for the top 20 Australian income-distributing managed investments are expressed as percentages in the table below.

4.61	4.50	2.63	4.13
3.62	3.49	1.44	1.82
3.35	2.70	3.17	3.35
5.33	5.66	2.76	1.72
5.14	2.87	1.66	3.40

- a** Calculate the mean and the standard deviation of the sample of 20 volatility measures.

- b** Construct a relative frequency histogram for the volatility measures.
c Find the median of the 20 volatility measures, and locate the mean and the median on your histogram.

5.103 XR05-103 A company that supplies temporary workers sponsored a survey of 100 executives. Each was asked to report the number of minutes they spend screening each job résumé they receive.

- a** Determine the quartiles.
b What information did you derive from the quartiles? What does this suggest about writing your résumé?

5.104 XR05-104 How much do pets cost? A random sample of dog and cat owners was asked to calculate the amount of money spent on their pets (exclusive of pet food). Draw a box plot for each data set and describe your findings.

5.105 XR05-105 Consider the relationship between the number of houses sold (Y) and the mortgage rate (X). The values of Y and X for 12 months are stored in columns 1 and 2 respectively.

- a** Calculate the covariance and coefficient of correlation between X and Y .
b What do these statistics tell you about the relationship between the level of house sales and mortgage rates based on these data?

5.106 XR05-106 The number of regular users of the internet is growing rapidly. However, internet use by people older than 60 is still relatively low. To learn more about this issue, a sample of 250 men and women older than 60 who had used the internet at least once were selected. The number of hours they spent on the internet during the past month was recorded.

- a** Calculate the mean and the median.
b Calculate the variance and the standard deviation.
c Draw a box plot.
d Briefly describe what you have learned from the statistics you calculated.

In addition to internet use, we have also recorded the number of years of education undertaken by each participant.

- e** Calculate the covariance and the coefficient of correlation between internet use and level of education.
f Describe what these statistics tell you about the relationship between internet use and level of education.

Case Studies

CASE 5.1 Return to the global warming question

C05-01a, C05-01b Now that we have presented techniques that allow us to conduct more precise analyses we'll return to Case 4.1. Recall that there are two issues in this discussion. First, is there global warming? Second, if so, is carbon dioxide the cause? The only tools available at the end of Chapter 4 were graphical techniques such as line charts and scatter diagrams. You are now invited to apply the more precise techniques in this chapter to answer the same questions.

- a Use the least squares method to estimate average monthly changes in temperature anomalies.
- b Calculate the least squares line and the coefficient of correlation between CO₂ levels and temperature anomalies and describe your findings.

CASE 5.2 Another return to the global warming question

C05-02a–C05-02d Did you conclude in Case 5.1 that Earth has warmed since 1880? If so, here is another look at the same data. **C05-02a** lists the temperature anomalies from 1880 to 1940; **C05-02b** lists the data from 1941 to 1975; **C05-02c** stores temperature anomalies from 1976 to 1997; and **C05-02d** contains the data from 1998 to 2016. For each set of data, calculate the least squares line and the coefficient of determination. Report your findings.

CASE 5.3 GDP versus consumption

C05-03 Although Australia is experiencing low rates of inflation, the expenditure on various consumer items has increased by a significant proportion. The following table presents the per capita gross domestic product (GDP) and per capita consumption for the Australian states and territories and for Australia as a whole for 2015–16. Present a summary report about the relationship between per capita GDP and consumption.

Per capita GDP and per capita consumption, Australia, 2015–16

State/Territory	GDP per capita (\$)	Per capita consumption (\$)
New South Wales (NSW)	67841	41 033
Victoria (Vic.)	61 258	38 322
Queensland (QLD)	64 280	37 543
South Australia (SA)	58 306	35 328
Western Australia (WA)	96 475	39 043
Tasmania (Tas.)	49 837	32 851
Northern Territory (NT)	94 932	41 572
Australian Capital Territory (ACT)	89 737	40 995
Australia	68 171	38 723

Source: Australian Bureau of Statistics, *Household Expenditure Survey, Australia: Summary of Results, 2015–2016*, cat. no. 6530.0, ABS, Canberra.

CASE 5.4 The gulf between the rich and the poor

C05-04 The human development index (HDI) is a composite measure of life expectancy, education and per capita income of a country. The following table presents the HDI and male

life expectancy (in years) of a group of the 50 richest countries and a group of the 50 poorest countries (in human development index rank order in 2017) in the world. Use descriptive summary measures to present an analysis of the disparities between the rich and the poor.

HDI rank	Rich country	Human Development Index	Life expectancy at birth (years)
1	Norway	0.953	82.3
2	Switzerland	0.944	83.5
3	Australia	0.939	83.1
4	Ireland	0.938	81.6
.	.	.	.
49	Russia	0.816	71.2
50	Montenegro	0.814	77.3
HDI rank	Poor country	Human Development Index	Life expectancy at birth (years)
1	Ghana	0.592	63.0
2	Equatorial Guinea	0.591	57.9
.	.	.	.
47	Chad	0.404	53.2
48	South Sudan	0.388	57.3
49	Central African Republic	0.367	52.9
50	Niger	0.354	60.4

Source: © Copyright 2019 United Nations Development Programme (UNDP).

CASE 5.5 Sydney and Melbourne leading the way in the growth in house prices

C05-05 A number of recent reports, including those published by the Commonwealth Bank, and two other national housing reports indicated that there is some improvement in home sales activity in Australia combined with a sharp increase in house prices. Some economists attributed this housing market recovery to the low interest rates and first-home owner grants and exemption from stamp duties in some states. Data for median (established) house prices in the capital cities of the six states and two territories in Australia for the March quarter of 2002–19 were collected from the Australian Bureau of Statistics website. Prepare a summary report using these data, analysing the movement in house prices in Australia as a whole and the six states and two territories in particular.

Source: Australian Bureau of Statistics, Residential Property Price Indexes: Eight Capital Cities, Tables 4 and 5, Dec 2019, cat. no. 6416.0 – March Quarter 2019.

CASE 5.6 Performance of managed funds in Australia: 3-star, 4-star and 5-star rated funds

C05-06 Average annual return is one of the key indicators in assessing the performance of a managed fund. The following table shows the returns of three-star, four-star and five-star rated funds over one- and three-year periods. Prepare a descriptive statistical summary report for a firm that provides information to clients who are keen to invest in one-year or three-year managed funds considering the star rating of the fund.

Fund name	STARS (*)	1-Year	3-Year
BT Japanese Share Retail	3	35.5	24.6
Colonial FirstChoice Investments – Goldman Sachs Global Small Comp	3	29.4	29.6
OnePath OA IP – Magellan Global Trust EF/Sel	3	34.9	25.2
OnePath OA IP – Magellan Global Trust NEF	3	33.8	24.1
OnePath OA IP – OP Global Share EF	3	31.4	27.5
Perpetual WFIA – BlackRock Global Sml Capl	3	31.2	31.6
PM Capital Global Companies	3	38.4	35.4
BlackRock Scientific International Equity Fund	4	30.0	27.7
Colonial FirstChoice Investments – Generation Global Share	4	32.6	27.9
Colonial FirstChoice Investments – Magellan Global	4	35.5	26.2
Colonial FirstChoice – W Inv – Goldman Sachs W Global Sm Co	4	30.5	30.7
Colonial FirstChoice – W Inv – Magellan W Global	4	36.5	27.2
Fidelity China	4	34.7	26.5
Generation Wholesale Global Share	4	32.7	28.5
Platinum Japan	4	42.2	39.0
Acadian Wholesale Global Equity Long Short	5	36.8	30.5
Arrowstreet Global Equity	5	31.0	30.1
BlackRock W Global Small Capital	5	34.3	34.5
Colonial FirstChoice Investments – Acadian Global Equity LS	5	35.8	29.6
Colonial FirstChoice Investments – PM Capital Global Companies	5	44.1	38.5
Colonial FirstChoice W Inv – PM Capital W Global Companies	5	42.7	38.6
IFP Global Franchise	5	35.0	25.7
Macquarie Asia New Stars No.1	5	30.5	34.5
Magellan Global	5	36.1	27.3
Zurich Investments Global Growth	5	30.1	25.8

Source: © Copyright 2020 Morningstar, Inc. All rights reserved

CASE 5.7 Life in suburbs drives emissions higher

C05-07 A research study by Parsons Brinckerhoff and Curtin University in Perth reported that people living in the outer suburbs in most Australian states and territories are contributing significantly to increases in greenhouse gases by driving their cars long distances to the city centre. It is reported that transport accounts for about half of an average Australian household's greenhouse footprint, so inner city residents could be generating just half the emissions of those living in the outer suburbs. The following table reports the greenhouse gas emissions from cars in kilograms per capita per day. Use numerical descriptive measures to present an analysis that compares the greenhouse gas emissions from cars in New South Wales and Victoria.

Greenhouse gases from cars (kg per capita per day)

NSW	Greenhouse gases (kg)	Victoria	Greenhouse gases (kg)
Sydney CBD	0.88	Melbourne CBD	1.07
Leichhardt	3.46	Moonee Valley	3.56
Woollahra	4.51	Darebin	2.53
Marrickville	3.80	Glen Eira	3.36
Ku-ring-gai	5.82	Monash	4.30
Pittwater	7.10	Knox	5.25
Baulkham Hills	8.10	Frankston	6.82
Penrith	8.18	Yarra Ranges	9.73
Hawkesbury	10.37		

Source: Reproduced from *The Australian*, 19 September 2009.

CASE 5.8 Aussies and Kiwis are leading in education

C05-08 According to 2016 published statistics on the education index (EI) by the UN, Australians and New Zealanders are leading the world. The EI is calculated using two indices: the Mean Years of Schooling Index and the Expected Years of Schooling Index for each country. The EI data for the top 20 and the bottom 20 of the 188 countries listed in the UN report for 2015 are recorded. Some of the data are given below. Use suitable numerical summary (central location and variability) measures to analyse the data.

Top 20 countries	EI	Bottom 20 countries	EI
1 Australia	0.939	1 Afghanistan	0.398
2 Denmark	0.923	.	.
3 New Zealand	0.917	14 Mali	0.312
4 Norway	0.916	15 Djibouti	0.310
5 Germany	0.914	16 South Sudan	0.297
6 Ireland	0.910	17 Chad	0.280
7 Iceland	0.906	18 Eritrea	0.267
.	.	19 Burkina Faso	0.262
20 Poland	0.852	20 Niger	0.206

Source: © Copyright 2016 United Nations Development Programme (UNDP).

CASE 5.9 Growth in consumer prices and consumption in Australian states

C05-09 Suppose the commodities consumed by Australian households can be grouped into 10 commodity groups: food, alcoholic beverages, clothing, housing, durables, health care, transport and communication, recreation, education and all others. Data for the rate of growth in prices and consumption of the 10 commodities and the share of total income Australian households allocate to each of the commodities for the six states of Australia are given. Calculate the overall growth in consumer prices and consumption in the six states.

Appendix 5.A

Summation notation

This appendix offers an introduction to the use of summation notation. Because summation notation is used extensively throughout statistics, you should review this appendix even if you have had previous exposure to summation notation. Our coverage of the topic begins with an introduction to the necessary terminology and notation, follows with some examples, and concludes with four rules that are useful in applying summation notation.

Consider n numbers x_1, x_2, \dots, x_n . A concise way of representing their sum is:

$$\sum_{i=1}^n x_i$$

That is:

$$\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n$$

Terminology and notation

- 1 The symbol Σ is the capital Greek letter sigma, and means 'the sum of'.
 - 2 The letter i is called the *index of summation*. The letter chosen to represent the index of summation is arbitrary.
 - 3 The expression $\sum_{i=1}^n x_i$ is read as 'the sum of the terms x_i where i assumes the values from 1 to n inclusive'.
 - 4 The numbers 1 and n are called the lower and upper limits of summation respectively.
- Summation notation is best illustrated by means of examples.

Examples

- 1 Suppose that $x_1 = 5$, $x_2 = 6$, $x_3 = 8$ and $x_4 = 10$. Then:

a
$$\begin{aligned}\sum_{i=1}^4 x_i &= x_1 + x_2 + x_3 + x_4 \\ &= 5 + 6 + 8 + 10 \\ &= 29\end{aligned}$$

b
$$\begin{aligned}\sum_{i=3}^4 x_i &= x_3 + x_4 \\ &= 8 + 10 \\ &= 18\end{aligned}$$

c
$$\begin{aligned}\sum_{i=1}^2 x_i(x_i - 1) &= x_1(x_1 - 1) + x_2(x_2 - 1) \\ &= 5(5 - 1) + 6(6 - 1) \\ &= 50\end{aligned}$$

d
$$\begin{aligned}\sum_{i=1}^3 f(x_i) &= f(x_1) + f(x_2) + f(x_3) \\ &= f(5) + f(6) + f(8)\end{aligned}$$

- 2 Suppose that $x_1 = 2$, $x_2 = 3$, $x_3 = 4$, and $y_1 = 8$, $y_2 = 9$, $y_3 = 13$. Then:

$$\begin{aligned}\mathbf{a} \quad \sum_{i=1}^3 x_i y_i &= x_1 y_1 + x_2 y_2 + x_3 y_3 \\ &= 2(8) + 3(9) + 4(13) \\ &= 95\end{aligned}$$

$$\begin{aligned}\mathbf{b} \quad \sum_{i=2}^3 x_i y_i^2 &= x_2 y_2^2 + x_3 y_3^2 \\ &= 3(9^2) + 4(13^2) \\ &= 919\end{aligned}$$

$$\begin{aligned}\mathbf{c} \quad \sum_{i=1}^2 (x_i - y_i) &= (x_1 - y_1) + (x_2 - y_2) \\ &= (2 - 8) + (3 - 9) \\ &= -12\end{aligned}$$

Remark

It is not necessary that the index of summation be a subscript, as the following examples demonstrate.

Examples

$$\mathbf{1} \quad \sum_{x=0}^4 x = 0 + 1 + 2 + 3 + 4 = 10$$

$$\mathbf{2} \quad \sum_{x=1}^3 (x^2 - x) = (1^2 - 1) + (2^2 - 2) + (3^2 - 3) = 8$$

$$\mathbf{3} \quad \sum_{x=1}^2 5x = 5(1) + 5(2) = 15$$

$$\mathbf{4} \quad \sum_{x=0}^3 f(x) = f(0) + f(1) + f(2) + f(3)$$

$$\mathbf{5} \quad \sum_{x=1}^2 f(x, y) = f(1, y) + f(2, y)$$

$$\mathbf{6} \quad \sum_{y=3}^5 f(x, y^2) = f(x, 3^2) + f(x, 4^2) + f(x, 5^2)$$

Rules of summation notation

1 If c is a constant, then

$$\sum_{i=1}^n c x_i = c \sum_{i=1}^n x_i$$

2 If c is a constant, then

$$\sum_{i=1}^n c = nc$$

3 If a and b are constants, then

$$\sum_{i=1}^n (ax_i + by_i) = a \sum_{i=1}^n x_i + b \sum_{i=1}^n y_i$$

4 If c is a constant, then

$$\sum_{i=1}^n (x_i + c) = \sum_{i=1}^n x_i + nc$$

Remark

Notice that

$$\sum_{i=1}^n x_i^2 \neq \left(\sum_{i=1}^n x_i \right)^2$$

To verify this, observe that

$$\sum_{i=1}^n x_i^2 = x_1^2 + x_2^2 + \cdots + x_n^2$$

and

$$\left(\sum_{i=1}^n x_i \right)^2 = (x_1 + x_2 + \cdots + x_n)^2$$

Exercises

A5.1 Evaluate $\sum_{i=1}^5 (i^2 + 2i)$.

A5.2 Evaluate $\sum_{x=0}^2 (x^3 + 2x)$.

A5.3 Using the set of observations below, evaluate the following sums.

a $\sum_{i=1}^{13} x_i$

b $\sum_{i=1}^{13} (2x_i + 5)$

c $\sum_{i=1}^6 (x_i - 5)^2$

i	1	2	3	4	5	6	7	8	9	10	11	12	13
x_i	3	12	10	-6	0	11	2	-9	-5	8	-7	4	-5

Appendix 5.B

Descriptive measures for grouped data

The two most important descriptive measures are the mean and the variance (or, alternatively, the standard deviation). This appendix looks briefly at how to approximate these two measures for data that have been grouped into a frequency distribution.

Frequently in practice you may need to rely on secondary data sources such as government publications. Data collected by others is usually presented in the form of a frequency distribution, and you do not have access to the ungrouped raw data. In this case, you will find that the approximations given here are useful.

Consider a sample of n observations that have been grouped into k classes. If f_i denotes the frequency of class i (for $i = 1, 2, \dots, k$), then $n = f_1 + f_2 + \dots + f_k$. A good approximation of the sample mean \bar{x} can be obtained by making the assumption that the midpoint m_i of each class closely approximates the mean of the observations in class i . This assumption is reasonable whenever the observations in a class are dispersed fairly symmetrically around the midpoint. The sum of the observations in class i is then approximately equal to $f_i m_i$.

Approximate mean and variance for grouped data

$$\text{Group mean} = \bar{x} \approx \frac{\sum_{i=1}^k f_i m_i}{n}$$

$$\begin{aligned}\text{Group variance} = s^2 &\approx \frac{1}{n-1} \left[\sum_{i=1}^k f_i m_i^2 - \frac{\left(\sum_{i=1}^k f_i m_i \right)^2}{n} \right] \\ &\approx \frac{1}{n-1} \left[\sum_{i=1}^k f_i m_i^2 - n \bar{x}^2 \right]\end{aligned}$$

$$\text{Group standard deviations} = \sqrt{s^2}$$

EXAMPLE A5.1

L05

Telephone-call durations – revisited

Consider Example 5.13, in which 30 long-distance telephone call durations were recorded and stored in **XM05-13**. Group the data into six class intervals and obtain the approximate mean and standard deviation for this sample of long-distance call durations. Compare these with the true values.





Solution

The frequency distribution for the duration of telephone calls, introduced in Example 5.13, is presented in the first three columns of **Table A5.1**. The last three columns have been included in the table to record the information required by the formulas for approximating the mean and the variance of the durations from these grouped data. We are now treating the 30 telephone-call durations as a sample, and the sample mean and variance are approximated as shown below.

TABLE A5.1 Extended frequency distribution of telephone-call durations of a Melbourne firm

Class <i>i</i>	Class limits	Frequency <i>f_i</i>	Midpoint <i>m_i</i>	<i>f_im_i</i>	<i>f_im_i²</i>
1	2 to 5	3	3.5	10.5	36.75
2	5 to 8	6	6.5	39.0	253.50
3	8 to 11	8	9.5	76.0	722.00
4	11 to 14	7	12.5	87.5	1093.75
5	14 to 17	4	15.5	62.0	961.00
6	17 to 20	2	18.5	37.0	684.50
Total		<i>n = 30</i>		312.0	3751.50

$$\bar{x} \approx \frac{\sum_{i=1}^6 f_i m_i}{30} = \frac{312.0}{30} = 10.4$$

$$s^2 \approx \frac{1}{29} \left[\sum_{i=1}^6 f_i m_i^2 - \frac{\left(\sum_{i=1}^6 f_i m_i \right)^2}{30} \right]$$

$$\approx \frac{1}{29} \left[3751.5 - \frac{(312)^2}{30} \right]$$

$$= 17.47$$

$$s \approx \sqrt{s^2} = \sqrt{17.47} = 4.29$$

These approximations match the true values obtained using the raw data – $\bar{x} = 10.26$ and $s = 4.29$ – reasonably well.

EXERCISES

A5.4 Self-correcting exercise.

- a Approximate the mean and the variance of the sample data presented in the following frequency distribution.

Class	Frequency
-20 to -10	8
-10 to 0	21
0 to 10	43
10 to 20	48
20 to 30	25
30 to 40	15

- b Use the range approximation of s to check your approximation of the variance in part (a).

- A5.5** The gross hourly earnings of a group of blue-collar workers randomly selected from the payroll list of a large Melbourne company were organised into the following frequency distribution.

Hourly earnings (\$)	Number of workers
8 to 10	11
10 to 12	17
12 to 14	32
14 to 16	27
16 to 18	13

- a Approximate the mean and the standard deviation of hourly earnings for this sample of workers.
- b Your answers to part (a) are only approximations of the true values of \bar{x} and s for this group's earnings. Explain why this is so.

A5.6 An Australia-wide car-rental agency recently bought 1000 identical new small cars from a major car manufacturer. After the customary 1000 km break-in period, it selected 100 cars at random and obtained the fuel consumption data listed in the following table.

Fuel consumption (km per litre)	Number of cars
9.0 to 10.5	9
10.5 to 12.0	13
12.0 to 13.5	24
13.5 to 15.0	38
15.0 to 16.5	16

Approximate the average fuel consumption and the standard deviation of consumption for this sample.

A5.7 The number of Australian males by age at first marriage in 1990, 2000 and 2010 are summarised by age group in the following table.

Number of male Australians by age group at first marriage – 1990, 2000 and 2010

Age	1990	2000	2010
14 to 24	31 761	17 454	14 119
24 to 34	58 060	61 111	64 325
34 to 44	16 170	20 485	24 793
44 to 54	6 371	9 024	10 801
54 to 64	2 865	3 583	5 189
64 to 74	1 314	1 300	1 488
74 to 94	417	472	461

Source: Australian Bureau of Statistics, *Divorces, Australia, 2010*, cat. no. 3307.0, ABS, Canberra, CC BY 2.5 AU
<http://creativecommons.org/licenses/by/2.5/au/legalcode>.

Approximate the mean and the standard deviation of the age at first marriage of Australian males (a) in 1990, (b) in 2000 and (c) in 2010. (The width of the last class is different from that of the others.)

Review of descriptive techniques

Here is a list of the statistical techniques introduced in Chapters 3, 4 and 5. This is followed by a flowchart designed to help you select the most appropriate method to use to address any problem requiring a descriptive method.

Graphical techniques

- Histogram**
- Stem-and-leaf display**
- Ogive**
- Bar chart**
- Pie chart**
- Scatter diagram**
- Bar chart of a contingency table**
- Line chart (time series)**
- Box plot**

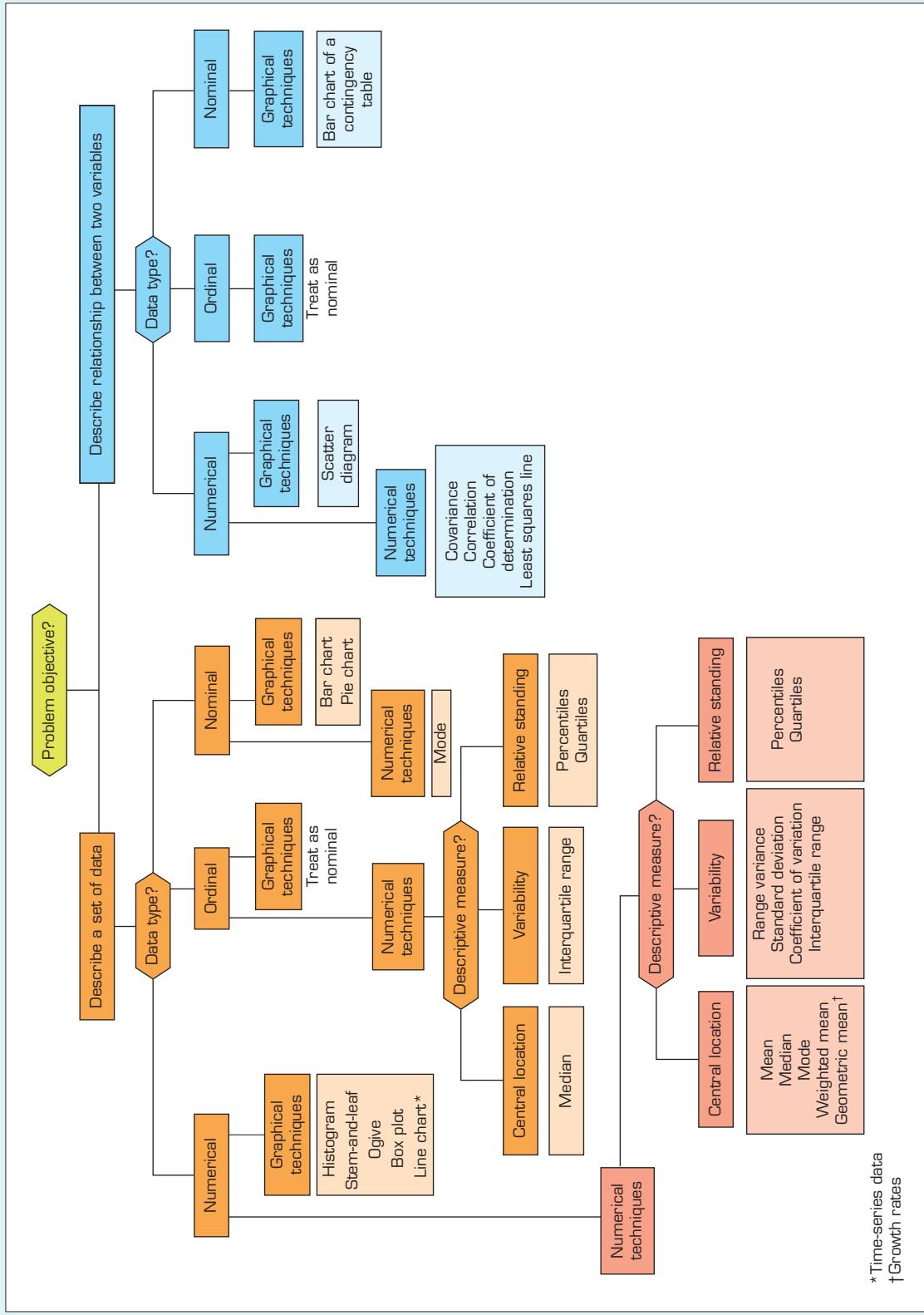
Numerical techniques

Measures of central location

- Mean**
- Median**
- Mode**

Measures of variability

- Range**
- Variance**
- Standard deviation**
- Coefficient of variation**
- Interquartile range**
- Measures of relative standing*
- Percentiles**
- Quartiles**
- Measure of linear relationship*
- Covariance**
- Coefficient of correlation**
- Coefficient of determination**
- Least squares line**

FIGURE A5.1 Flowchart: Graphical and numerical techniques

Probability

Learning objectives

This chapter introduces the basic concepts of probability and outlines the rules and techniques for assigning probabilities to events.

At the completion of this chapter, you should be able to:

- L01** explain the importance of probability theory to statistical inference
- L02** define the terms *random experiment*, *sample space*, *simple event*, *event*, *union of events*, *intersection of events*, *complement of an event* and *mutually exclusive events*
- L03** explain the different approaches to assigning probabilities to events
- L04** define the terms *joint*, *marginal* and *conditional probabilities* and *independent events*
- L05** calculate probabilities using the basic probability rules – complement rule, addition rule and multiplication rule of probabilities
- L06** understand the use of probability trees in calculating complicated probabilities
- L07** understand Bayes' law as an alternative way of calculating conditional probabilities.

CHAPTER OUTLINE

Introduction

- 6.1** Assigning probabilities to events
- 6.2** Joint, marginal and conditional probability
- 6.3** Rules of probability
- 6.4** Probability trees
- 6.5** Bayes' law
- 6.6** Identifying the correct method

SPOTLIGHT ON STATISTICS

Auditing tax returns

The Australian Taxation Office's auditors routinely check tax returns to determine whether calculation errors have been made and to detect fraudulent returns. There are several methods used by dishonest taxpayers to evade income tax. One method is not to declare various sources of income. Auditors have several methods of detecting this, including analysing spending patterns. Another form of tax fraud is to claim fraudulent deductions.

After analysing the returns of thousands of self-employed taxpayers, an auditor has determined that 45% of fraudulent returns contain two suspicious deductions, 28% contain one suspicious deduction, and the rest no suspicious deductions. Among honest returns the rates are 11% for two suspicious deductions, 18% for one suspicious deduction, and 71% for no suspicious deductions.



Source: iStock.com/Faberrink

The auditor believes that 5% of the returns of self-employed individuals contain significant fraud. The auditor has just received a tax return for a self-employed individual that contains one suspicious expense deduction. What is the probability that this tax return contains significant fraud? (See page 247 for the answer.)

Introduction

In Chapters 3, 4 and 5, we introduced graphical and numerical descriptive methods. Although these methods are useful on their own, we are particularly interested in developing statistical inference. Decision makers make inferences about population parameters (e.g. the *average* income of blue-collar workers in Australia, or the *proportion* of Australian voters supporting the federal government's policy on boat people) based on sample statistics. The topic of probability, which we present in this chapter, and which is an integral part of all statistics, provides a link between the sample statistics and the inferences made about the population parameters. Statistical inference provides decision makers, such as business people and scientists, with a body of methods that aid in decision making in situations of uncertainty. Such uncertainty arises because in real-life situations we rarely have perfect information about the various conditions affecting our decisions. Whether the uncertainty relates to the future demand for a product, future level of interest rates, the possibility of a labour strike or the proportion of defective widgets in the next production run, probability theory can be used to measure the degree of uncertainty involved.

In this chapter, we provide a brief introduction to the basics of probability.

6.1 Assigning probabilities to events

random experiment

A process that results in one of a number of possible different outcomes.

Random experiment

A random experiment is a process that results in one of a number of possible outcomes that cannot be predicted with certainty.

Here are six illustrations of random experiments and their possible outcomes.

	Experiment	Possible outcomes
1	Flip a coin	Head, tail
2	Roll a die	1, 2, 3, 4, 5, 6
3	Record the quality of service at the customer service desk of a supermarket	Very poor, poor, unsure, reasonable, good, very good
4	Observe changes in the Qantas share price over one week	Increase, decrease, no change
5	Record marks on a statistics test (out of 100)	Numbers between 0 and 100
6	Record grade in a marketing test	Fail (F), Pass (P), Credit (C), Distinction (D), High Distinction (HD)

probability

The likelihood of the occurrence of an outcome or a collection of outcomes.

An important feature of a random experiment is that *the actual outcome cannot be determined in advance*. The best we can do is to determine the **probability** (chance or likelihood) that a particular outcome will occur.

To determine the probability of an outcome of a random experiment, the first step is to list all such possible outcomes of the experiment. Such a listing of all possible outcomes is called

a **sample space**. The listed outcomes must be **exhaustive**, which means that all possible outcomes must be included. In addition, the outcomes must be **mutually exclusive**, which means that no two outcomes can occur at the same time.

To illustrate the concept of exhaustive outcomes consider this list of the outcomes in Experiment 2 above – roll a die:

1 2 3 4 5

This list is not exhaustive, because we have omitted 6.

The concept of mutual exclusiveness can be seen by listing the following outcomes in Experiment 5 above: Suppose we group the student marks into the following five groups to match the standard grades: F, P, C, D and HD ranges.

F: 0–50, P: 50–65, C: 65–75, D: 75–85, HD: 85–100

If these intervals include both the lower and upper limits (for example, $65 \leq X \leq 75$ and $75 \leq X \leq 85$), then these outcomes are not mutually exclusive because two outcomes can occur for a student. For example, if a student receives a mark of 75, both the third and fourth outcomes occur. Therefore, the grades should be defined such that they include the lower limit, but not the upper limit of the marks.

Note that we could produce more than one list of exhaustive and mutually exclusive outcomes. For example, here is another list of outcomes for Experiment 5:

Pass, Fail

A list of exhaustive and mutually exclusive outcomes is called a *sample space* and is denoted by S . The outcomes are denoted by O_1, O_2, \dots, O_n .

Sample space

The sample space of a random experiment is a list of all possible outcomes of the experiment. The outcomes must be exhaustive and mutually exclusive.

Using the set notation, where { } is read as 'the set consisting of', we represent the sample space and its outcomes as

$$S = \{O_1, O_2, \dots, O_n\}$$

For example, in Experiment 1 above, the sample space would be $S = \{\text{Head, Tail}\}$. For Experiment 2, the sample space would be $S = \{1, 2, 3, 4, 5, 6\}$. The individual outcomes in a sample space are also called elementary (or simple) events.

Probability of an outcome

The probability of an outcome is simply the chance that an outcome O_i will occur, denoted by $P(O_i)$, where $i = 1, 2, \dots, n$.

Once a sample space has been prepared, we begin the task of assigning probabilities to the outcomes. There are three ways to assign probability to outcomes. However it is done, there are two basic rules governing probabilities.

Requirements of probabilities

Given a sample space of $S = \{O_1, O_2, \dots, O_n\}$, the probabilities assigned to the outcomes O_i must satisfy two basic requirements:

1 The probability of an outcome must lie between 0 and 1,

$$0 \leq P(O_i) \leq 1 \text{ for each } i, \text{ where } i = 1, 2, \dots, n.$$

2 The sum of the probabilities of all possible outcomes in a sample space must be equal to 1. That is:

$$P(O_1) + P(O_2) + \dots + P(O_n) = 1 \text{ or } \sum_{i=1}^n P(O_i) = 1$$

sample space

The set of all possible simple events or outcomes.

exhaustive

Covers all possible outcomes.

mutually exclusive

Outcomes that cannot occur at the same time.

6.1a Three approaches to assigning probabilities

There are three distinct ways of assigning probability to outcomes: the *classical approach*, the *relative frequency approach* and the *subjective approach*.

classical approach

Assigning equal probabilities to all the elementary events or outcomes.

relative frequency approach

Expresses the probability of an outcome as the relative frequency of its occurrence based on past experience.

subjective approach

Assigns probability to an outcome based on personal judgement.

The **classical approach** is used by mathematicians to help determine the probability logically for outcomes associated with games of chance. If a perfectly balanced (or unbiased) coin is flipped, for example, it is logical to expect that the outcome *heads* and the outcome *tails* are equally likely. Hence we assert that the probability of observing an occurrence of heads is $1/2$. More generally, if an experiment has n possible outcomes (each of which is equally likely), under the classical approach, the probability of any particular outcome occurring is $1/n$.

The **relative frequency approach** expresses an outcome's probability as the long-run relative frequency with which the outcome occurs. For example, if 600 of the last 1000 customers entering a shop have made a purchase, the probability that any given customer entering the shop will make a purchase is approximately $600/1000 = 0.6$. Suppose that a random experiment is repeated n times, where n is a large number. In general, if x represents the number of times a particular outcome occurred in those n trials, under the relative frequency approach, the proportion x/n provides an estimate of the probability that that particular outcome will occur. The larger n is, the better the estimate of the required probability will be.

In many practical situations, the experimental outcomes are not equally likely, and there is no history of repetitions of the experiment. This might be the case, for example, if you wished to estimate the probability of striking oil at a new offshore drilling site, or the likelihood of your firm's sales reaching \$1 million this year. In such situations, we resort to the **subjective approach**, under which the probability assigned to an outcome simply reflects the degree to which we believe that the outcome will occur. The probability assigned to a particular outcome under the subjective approach therefore reflects a personal evaluation of the situation and may be based simply on intuition.

6.1b Defining events

An individual outcome of a sample space is called a *simple event* or elementary event. All other events are composed of the simple events in a sample space. An **event** is a collection of simple events or outcomes and is denoted by a capital letter. For example, in the tossing a die experiment, the event 'an even number is observed' can be described as $A = \{2, 4, 6\}$.

An event

An event is a collection or set of one or more simple events in a sample space.

In Experiment 6 we can define the event 'achieve a grade of HD' as the set of numbers that lie between 85 and 100, inclusive. Using set notation, we have:

$$HD = \{85, 86, \dots, 99, 100\}$$

Similarly:

$$F = \{0, 1, 2, \dots, 48, 49\}$$

Probability of an event

The **probability of an event** is the sum of the probabilities of the simple events that constitute the event.

probability of an event

Sum of the probabilities of the simple events in the event.

For example, in Experiment 2, rolling a die, the probability of the event $E = \{\text{an even number is observed}\}$ is:

$$P(E) = P\{2, 4, 6\} = P(2) + P(4) + P(6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

Suppose that in Experiment 6, we employed the relative frequency approach to assign probabilities to the simple events as follows:

$$\begin{aligned}P(F) &= 0.10 \\P(P) &= 0.15 \\P(C) &= 0.25 \\P(D) &= 0.30 \\P(HD) &= 0.20\end{aligned}$$

The probability of the event ‘pass the course’ is:

$$P(\text{Pass}) = P(P) + P(C) + P(D) + P(HD) = 0.15 + 0.25 + 0.30 + 0.20 = 0.90$$

It follows from the two basic requirements that the probability of an event that is certain to occur is 1, since such an event must contain all the outcomes in the sample space, and the sum of all possible outcome probabilities must be 1. However, the probability of an event that cannot possibly occur is 0.

We now state the basic rule governing the assignment of probabilities under the classical approach.

Probability of an event – equally likely outcomes

If each outcome in a finite sample space S has the same chance of occurring, the probability that an event E will occur is:

$$P(E) = \frac{\text{number of outcomes in } E}{\text{number of outcomes in } S}$$

6.1c Venn diagram

A useful graphical representation of a sample space S , the Venn diagram, is presented in **Figure 6.1**. In a Venn diagram, the entire sample space S is represented by a rectangle; points inside the rectangle represent the individual outcomes in S . An event A in the sample space S is represented by a circle, and points inside the circle represent the individual outcomes in event A .

Now, let’s look at how to combine events to form new ones. In arithmetic, new numbers can be created from existing ones by means of operations such as addition and multiplication. Similarly, in probability, new events can be created from events already defined by means of operations called complement, intersection and union.

The **complement of an event** A , denoted as \bar{A} , is the set of all outcomes in the sample space S that do not belong to A .

Thus, \bar{A} is an event in which the event A *does not occur*, as shown in **Figure 6.2**.

complement of an event

The set of all outcomes not in the event but in the sample space.

FIGURE 6.1 Venn diagram depicting an event A in a sample space

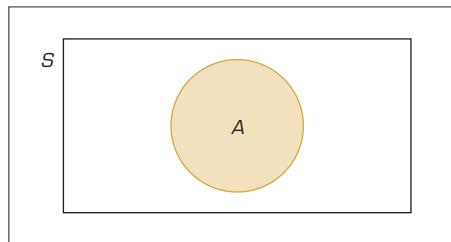
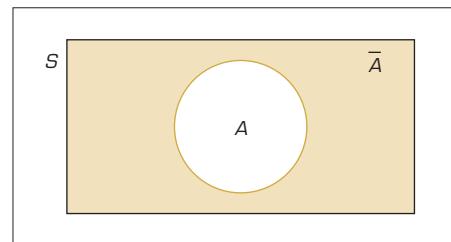


FIGURE 6.2 Venn diagram depicting \bar{A} , the complement of any event A , in a sample space



6.1d Intersection

Intersection

A joint event consisting of common outcomes from two events.

Intersection of events A and B

The intersection of events A and B is the event that occurs when both A and B occur. It is denoted as $A \cap B$ or A and B .

The probability of the intersection is called the joint probability.

Thus, the event $A \cap B$ occurs only if both event A and event B occur together. The event $A \cap B$ is sometimes denoted as A and B , and is as shown in **Figure 6.3**.

If there are no outcomes common to both A and B , then the event $A \cap B$ is an empty event (i.e. $A \cap B = \{\}$). Such events A and B are called mutually exclusive events or *disjoint events*, and $P(A \cap B) = 0$. This is illustrated in **Figure 6.4**.

FIGURE 6.3 Venn diagram depicting the event $A \cap B$ in a sample space

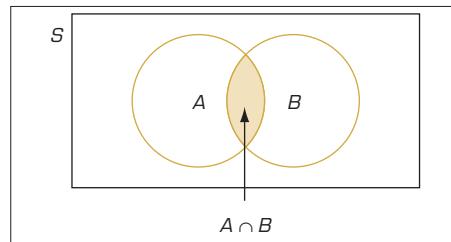
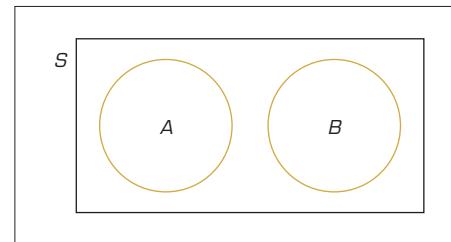


FIGURE 6.4 Venn diagram depicting the event $P(A \cap B) = 0$ in a sample space



6.1e Union

union

An event consisting of all outcomes from two events.

Union of events A and B

The union of events A and B , denoted as $A \cup B$, is the event consisting of all outcomes in A or in B or in both. It is denoted as $A \cup B$ or A or B .

Thus, the event $A \cup B$ occurs if only A occurs, if only B occurs or if both A and B occur. The event $A \cup B$ is sometimes denoted as A or B . This is illustrated in **Figure 6.5**. **Figure 6.6** illustrates the event A only or B only.

FIGURE 6.5 Venn diagram depicting the event $A \cup B$ in a sample space

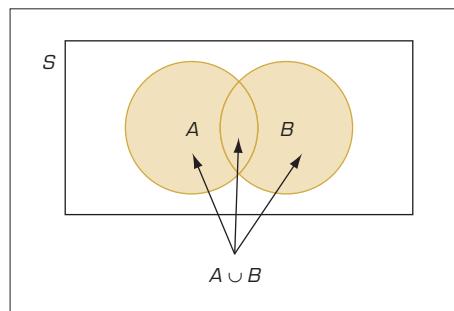
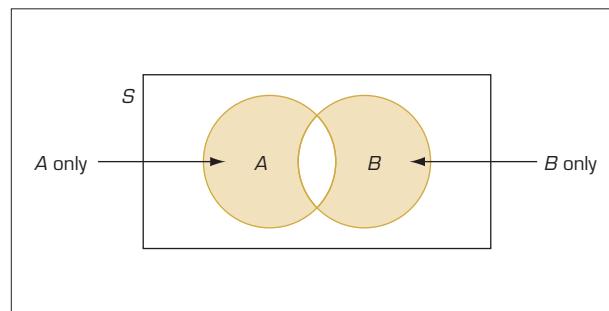


FIGURE 6.6 Venn diagram depicting the event A only or B only in a sample space



We will refer frequently to the probability of occurrence of the complement, the union or the intersection of events. The notation for these probabilities is as follows:

$$P(\bar{A}) = P(A \text{ does not occur})$$

$$P(A \cup B) = P(\text{only } A \text{ occurs or only } B \text{ occurs or both } A \text{ and } B \text{ occur})$$

$$P(A \cap B) = P(\text{both } A \text{ and } B \text{ occur})$$

EXAMPLE 6.1

LO2

Tossing a six-sided die

The number of spots turning up when a six-sided die is tossed is observed. Consider the following events:

A: The number observed is an even number.

B: The number observed is greater than 4.

C: The number observed is less than 4.

D: The number observed is 4.

a Define a sample space for this random experiment, and assign probabilities to the outcomes.

b Find $P(A)$, $P(B)$, $P(C)$ and $P(D)$.

c Find $P(\bar{A})$.

d Find $P(A \cap B)$.

e Find $P(A \cup B)$.

f Are events *B* and *C* mutually exclusive?

Solution

a The sample space is $S = \{1, 2, 3, 4, 5, 6\}$, see **Figure 6.7**. As each of the six possible outcomes is equally likely to occur:

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$$

b The events *A*, *B*, *C* and *D* are as follows: $A = \{2, 4, 6\}$, $B = \{5, 6\}$, $C = \{1, 2, 3\}$, $D = \{4\}$. Since the probability of an event *E* is equal to the sum of the probabilities assigned to the outcomes contained in *E*:

$$P(A) = P(2) + P(4) + P(6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 0.5$$

Similarly,

$$P(B) = P(5) + P(6) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3} = 0.333$$

$$P(C) = P(1) + P(2) + P(3) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2} = 0.5$$

$$P(D) = P(4) = \frac{1}{6} = 0.167$$

- c The three outcomes in \bar{A} are represented in **Figure 6.8** by the points lying outside the region describing event A. As can be seen, the complement of event A is $\bar{A} = \{1, 3, 5\}$. Therefore,

$$P(\bar{A}) = P(1) + P(3) + P(5) = \frac{3}{6} = 0.5$$

- d The intersection of the two events A and B is $(A \cap B) = \{6\}$, which is shown in **Figure 6.9**. Therefore,

$$P(A \cap B) = P(6) = \frac{1}{6} = 0.167$$

- e The union of the two events A and B is $(A \cup B) = \{2, 4, 5, 6\}$, which is depicted by the shaded area in **Figure 6.10**. Therefore,

$$P(A \cup B) = P(2) + P(4) + P(5) + P(6) = \frac{4}{6} = 0.67$$

- f Events $B = \{5, 6\}$ and $C = \{1, 2, 3\}$ defined in this example are mutually exclusive because B and C have no outcomes in common. (Note that the regions representing B and C in **Figure 6.9** do not overlap.) That is, the event $(B \cap C) = \{\}$ is an empty set.

FIGURE 6.7 Venn diagram for sample space S

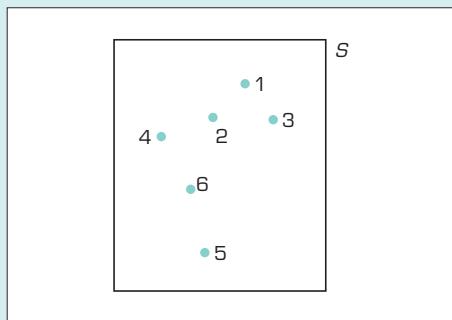


FIGURE 6.8 Venn diagram depicting event A

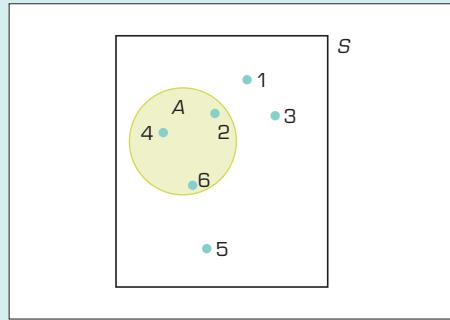


FIGURE 6.9 Venn diagram depicting event $A \cup B$ and event C

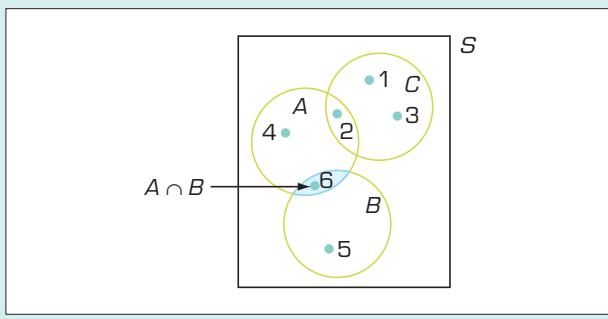
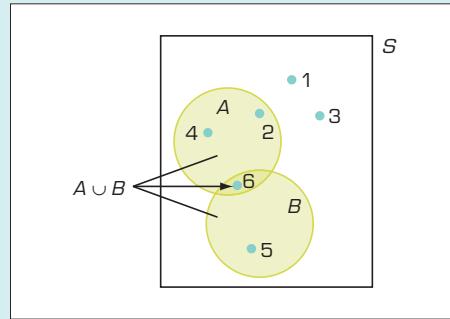


FIGURE 6.10 Venn diagram depicting event $A \cup B$



EXAMPLE 6.2

LO2

Selecting shares to sell

An investor who has \$3000 invested in each of four shares must sell two to help finance his daughter's wedding. Since he feels that all four shares are of comparable quality and have the same likelihood of appreciating in price over the coming year, he simply chooses at random the two shares to be retained and sells the other two. Suppose that, one year later, two of the original four shares have increased in value and two have decreased.

- Find the probability that both of the retained shares have increased in value.
- Find the probability that at least one of the retained shares has increased in value.
- Find the probability that only one of the retained shares has increased in value.

Solution

Let I_1 and I_2 represent the two shares that increased in value, and let D_1 and D_2 represent the two shares that decreased in value. Since the random experiment consists of choosing two shares from these four, a sample space for the experiment is

$$S = \{I_1 I_2, I_1 D_1, I_1 D_2, I_2 D_1, I_2 D_2, D_1 D_2\}$$

where each outcome represents a possible pair of retained shares. Each pair of shares had the same chance of being selected, so the probability that any particular one of the six pairs of shares was retained is $1/6$.

- Let A be the event in which both of the retained shares increased in value. Then $A = \{I_1 I_2\}$ and $P(A) = 1/6 = 0.17$.
- Let B be the event in which at least one of the retained shares increased in value. Then the event B consists of all outcomes for which either I_1 or I_2 is retained. That is, $B = \{I_1 I_2, I_1 D_1, I_1 D_2, I_2 D_1, I_2 D_2\}$. Since the probability of an event is the sum of the probabilities of the outcomes contained in that event, $P(B) = 5/6 = 0.83$.
- Let C be the event in which only one of the retained shares increased in value. That is, $C = \{I_1 D_1, I_1 D_2, I_2 D_1, I_2 D_2\}$. Therefore, $P(C) = 4/6 = 0.67$.

EXAMPLE 6.3

LO2

Deluxe versus standard air-conditioners

Keep Kool Pty Ltd manufactures window air-conditioners in both a deluxe model and a standard model. An auditor undertaking a compliance audit of the firm is validating the sales account for April. She has collected 200 invoices for the month, some of which were sent to wholesalers and the remainder to retailers. Of the 140 retail invoices, 28 are for the standard model. Of the wholesale invoices, 24 are for the standard model. If the auditor selects one invoice at random, find the probability that:

- the invoice selected is for the deluxe model
- the invoice selected is a wholesale invoice for the deluxe model
- the invoice selected is either a wholesale invoice or an invoice for the standard model.

Solution

The sample space S here consists of the 200 invoices. Whenever the outcomes can be classified according to two relevant characteristics – in this case, according to model sold and type of purchaser – it is worthwhile displaying the pertinent information in a cross-classification table, such as **Table 6.1**. The data (number of sales) given in this example have been circled in **Table 6.1**; you should check to confirm that you can fill in the remaining numbers yourself from the given information.

TABLE 6.1 Classification of invoices

Model	Wholesale (W)	Retail (\bar{W})	Total
Deluxe (D)	36	112	148
Standard (\bar{D})	(24)	(28)	52
Total	60	(140)	(200)

The events of interest are as follows:

W : A wholesale invoice is selected.

\bar{W} : A retail invoice is selected.

D : An invoice for deluxe model is selected.

\bar{D} : An invoice for standard model is selected.

- a There are 200 invoices in total, so the number of possible outcomes in the sample space S is 200. As there are 148 invoices for the deluxe model, event D contains 148 outcomes. Therefore, the probability that the invoice selected was for the deluxe model is:

$$P(D) = \frac{\text{number of outcomes in } D}{\text{number of outcomes in } S} = \frac{148}{200} = 0.74$$

- b As there are 36 wholesale invoices for the deluxe model, the event $(D \cap W)$ contains 36 outcomes. Hence, the probability that the invoice selected was a wholesale invoice for the deluxe model is:

$$P(D \cap W) = \frac{36}{200} = 0.18$$

- c The number of invoices that are either wholesale invoices (W) or invoices for the standard model (\bar{D}) is $60 + 52 - 24 = 88$. (Use a Venn diagram to see this.) Thus, the event $(W \cup \bar{D})$ contains 88 outcomes. Therefore:

$$P(W \cup \bar{D}) = \frac{88}{200} = 0.44$$

6.1f Interpreting probability

No matter what method was used to assign probability, we interpret it using the relative frequency approach for an infinite number of experiments. For example, an investor might have used the subjective approach to determine that there is a 65% probability that the price of a particular stock will increase over the next month. However, we interpret the 65% figure to mean that if we had an infinite number of stocks with exactly the same economic and market characteristics as the one the investor will buy, 65% of them will increase in price over the next month. Similarly, we can determine that the probability of throwing a 5 with a balanced die is 1/6. We might have used the classical approach to determine this probability; however, we interpret the number as the proportion of times that a 5 is observed on a balanced die thrown an infinite number of times. This relative frequency approach is useful when interpreting probability statements such as those heard from weather forecasters or scientists. You will also discover that this is the way we link the population and the sample in statistical inference.

EXERCISES

Learning the techniques

- 6.1** The weather forecaster reports that the probability of rain tomorrow is 10%.
- Which approach was used to assign this probability?
 - How would you interpret this probability?
- 6.2** A quiz contains a multiple-choice question with five possible answers, only one of which is correct. A student plans to guess the answer because he knows absolutely nothing about the subject.
- Produce the sample space for the outcomes of a question.
 - Assign probabilities to the outcomes in the sample space you produced.
 - Which approach did you use to answer part (b)?
 - Interpret the probabilities you assigned in part (b).
- 6.3** An investor tells you that in her estimation there is a 60% probability that the Australian All Ordinaries Index will increase tomorrow.
- Which approach was used to produce this figure?
 - Interpret the 60% probability.
- 6.4** A manager must decide which two of two male applicants (Bill and David) and two female applicants (Anne and Cynthia) should receive job offers.
- What is the random experiment?
 - List all possible outcomes in S .
 - List the outcomes in the following events:
 L : Cynthia receives an offer.
 M : Bill doesn't receive an offer.
 N : At least one woman receives an offer.
- 6.5** The number of spots turning up when a six-sided die is tossed is observed. The sample space for this experiment is $S = \{1, 2, 3, 4, 5, 6\}$. Answer each of the following questions, and use a Venn diagram to depict the situation graphically.
- What is the union of $A = \{2, 3\}$ and $B = \{2, 6\}$?
 - What is the intersection of $A = \{2, 3, 4, 5\}$ and $B = \{3, 5, 6\}$?
 - What is the complement of $A = \{2, 3, 4, 5\}$?
 - Are $A = \{3\}$ and $B = \{1, 2\}$ mutually exclusive events? Explain.
- 6.6** A sample space for the experiment consisting of flipping a coin twice is $S = \{HH, HT, TH, TT\}$.

Consider the following events:

$$A = \{HT, TH\}$$

$$B = \{HH, HT, TH\}$$

$$C = \{TT\}$$

- Describe each of the events A , B and C in words.
- List the outcomes in $A \cup B$, and use a Venn diagram to depict the union graphically.
- List the outcomes in $(A \cap B)$, and use a Venn diagram to depict the intersection graphically.
- List the outcomes in \bar{A} , and use a Venn diagram to depict the complement graphically.
- Is there a pair of mutually exclusive events among A , B and C ? Explain.

- 6.7** During a recent promotion, a bank offered mortgages with terms of one, two and three years at a reduced interest rate. Customers could also choose between open and closed mortgages. From the file of approved mortgage applications, the manager selects one application and notes both the term of the mortgage and whether it is open or closed. A sample space for this experiment is $\{O_1, O_2, O_3, C_1, C_2, C_3\}$, where, for example, O_2 represents selection of an open two-year mortgage. Consider the following events:

$$A = \{O_1, O_2, O_3\}$$

$$B = \{O_2, C_2\}$$

- Describe each of the events A and B in words.
- List the outcomes in $A \cup B$, and use a Venn diagram to depict the union graphically.
- List the outcomes $(A \cap B)$, and use a Venn diagram to depict the intersection graphically.
- List the outcomes in \bar{B} , and use a Venn diagram to depict the complement graphically.
- Are A and B mutually exclusive events? Explain.

- 6.8** The number of spots turning up when a six-sided die is tossed is observed. List the outcomes in each of the following events, and then find the probability of each event occurring.
- S : An event in the sample space is observed.
 A : A 6 is observed.
 B : The number observed is less than 4.
 C : An odd number is observed.
 D : An even number greater than 2 is observed.

- 6.9** The result of flipping two fair coins is observed.
- Define the sample space.
 - Assign probabilities to the individual outcomes.
 - Find the probability of observing one head and one tail.
 - Find the probability of observing at least one head.

Applying the techniques

- 6.10** In Exercise 6.4, we considered a manager who was deciding which two of four applicants (Anne, Bill, Cynthia and David) should receive job offers. Suppose that the manager, having deemed the applicants to be equally qualified, chooses at random the two who will receive job offers.
- Assign probabilities to the outcomes in the sample space.
 - Find the probability that Cynthia will receive an offer.
 - Find the probability that one man and one woman will receive an offer.
 - Find the probability that Bill will not receive an offer.
 - Find the probability that at least one woman will receive an offer.

- 6.11** **Self-correcting exercise.** A computer supplies store that sells personal computers and related supplies is concerned that it may be overstocking printers. The store has tabulated the number of printers sold weekly for each of the past 80 weeks. The store records show that the maximum number of printers sold in any given week is four. The results are summarised in the following table:

Number of printers sold	Number of weeks
0	36
1	28
2	12
3	2
4	2

The store intends to use the tabulated data as a basis for forecasting printer sales in any given week.

- Define the random experiment of interest to the store.
- List all possible outcomes in the sample space.
- Assign probabilities to each of the individual outcomes.

- What approach did you use in determining the probabilities in part (c)?
- Find the probability of selling at least three printers in any given week.

- 6.12** The trustee of a mining company's accident compensation plan has solicited the employees' feelings towards a proposed revision in the plan. A breakdown of the responses is shown in the following table. Suppose that an employee is selected at random, with the relevant events defined as follows:

- B*: The employee selected is a blue-collar worker.
W: The employee selected is a white-collar worker.
M: The employee selected is a manager.
F: The employee selected favours the revision.

Decision	Blue-collar workers (<i>B</i>)	White-collar workers (<i>W</i>)	Managers (<i>M</i>)
For (<i>F</i>)	67	32	11
Against (\bar{F})	63	18	9

- Define a sample space for this experiment.
- List all the outcomes belonging to the event *F*.
- Find the probability that the employee selected is:
 - a blue-collar worker
 - a white-collar worker
 - a manager
 - in favour of the decision
 - against the decision.
- Find the probability that the employee selected is not a manager.

- 6.13** Refer to Exercise 6.12. Express each of the following events in words, and find its probability:
- $B \cup W$ (*B* or *W*)
 - $F \cup M$ (*F* or *M*)
 - $\bar{F} \cap W$ (\bar{F} and *W*)
 - $F \cap \bar{M}$ (*F* and \bar{M})

- 6.14** Referring to Exercise 6.7, suppose that 300 mortgage applications were approved and that the numbers of mortgages of each type were as shown in the table below. The manager selects one mortgage application at random. The relevant events are defined as follows:

- L*: The application selected is for a one-year mortgage.
M: The application selected is for a two-year mortgage.

- N:** The application selected is for a three-year mortgage.
- C:** The application selected is for a closed mortgage.

Type of mortgage	Term of mortgage (years)		
	One (L)	Two (M)	Three (N)
Open (\bar{C})	32	36	60
Closed (C)	80	48	44

- a** Find the probability that the mortgage application selected is for:
- i** a one-year mortgage
 - ii** a two-year mortgage
 - iii** a three-year mortgage
 - iv** a closed mortgage
 - v** an open mortgage.
- b** Find the probability that the term of the mortgage selected is longer than one year.
- 6.15** Refer to Exercise 6.14. Express each of the following events in words, and find its probability.
- a** $L \cup M$
- b** $L \cup C$
- c** $M \cap \bar{C}$
- d** $\bar{N} \cap C$
- 6.16 XR06-16** According to a share ownership survey carried out by the Australian Securities Exchange in 2014, about 8 million Australians owned shares. The following table lists the estimated numbers of share owners by state and territory.

State/Territory	Number of share owners ('000)
NSW	2716
Victoria	2053
Queensland	1422
SA	456
WA	854
Tasmania	93
ACT	124
Northern Territory	59
Total	7777

Source: Calculated based on Australian Securities Exchange 2015, 2014 Australian Share Ownership Study, Australian Securities Exchange Limited, Sydney

Suppose that a share owner is selected at random. We are interested in finding the probabilities of selecting a share owner from each of the various states and territories.

- a** Define the random experiment.
- b** List all possible outcomes in the sample space.
- c** Assign probabilities to each of the individual outcomes.
- d** What approach did you use in determining the probabilities in part (c)?
- e** What is the probability of selecting a share owner from a state or territory in which the number of share owners is greater than 1000000?

- 6.17** Four candidates are running for local council. The four candidates are Soorley, Hayes, Quinn and Atkinson. Determine the sample space of the results of the election. A political scientist has used the subjective approach and assigned the following probabilities:

$$P(\text{Soorley wins}) = 0.42$$

$$P(\text{Hayes wins}) = 0.09$$

$$P(\text{Quinn wins}) = 0.27$$

$$P(\text{Atkinson wins}) = 0.22$$

Determine the probabilities of the following events:

- a** Soorley loses
- b** either Hayes or Atkinson wins
- c** one of Soorley, Hayes or Quinn wins.

6.2 Joint, marginal and conditional probability

In Section 6.1, we considered the intersection of two events known as a joint event. In this section, we consider methods of obtaining probabilities of the occurrence of a joint event, as well as the occurrence of an event given that another event has already occurred.

6.2a Joint and marginal probabilities

joint probability

The likelihood of occurrence of a joint event.

marginal probability

The probability of an event irrespective of any other event.

The probability of the intersection of two events is called the **joint probability**, since it expresses the likelihood of occurrence of a joint event. For an illustration of the notion of *joint* and *marginal probabilities*, consider once again Example 6.3, involving the audit of Keep Kool Pty Ltd. The information displayed in **Table 6.1** can be expressed alternatively as probabilities (relative frequencies), by dividing the number in each category by the total of 200, as shown in **Table 6.2**. The four probabilities in the interior of **Table 6.2** are joint probabilities. As can be seen, these four joint probabilities sum to one. For example, 0.18 (= 36/200) is the probability that the joint event ($D \cap W$) will occur. The two sets of probabilities (0.30, 0.70) and (0.74, 0.26) that appear at the margins of the table are called marginal probabilities. They are, respectively, the marginal probabilities that (W, \bar{W}) and (D, \bar{D}) will occur. **Marginal probabilities** are simply unconditional probabilities. Each set of marginal probabilities also sums to one.

TABLE 6.2 Probabilities for invoice classifications

Model	Wholesale (W)	Retail	Total
Deluxe (D)	0.18	0.56	0.74
Standard	0.12	0.14	0.26
Total	0.30	0.70	1.00

Marginal probabilities of sales types **Marginal probabilities of model types** **Joint probabilities**

6.2b Conditional probability

conditional probability

The probability an event will occur, given that another has occurred or will occur.

When calculating the probability of an event, we can sometimes make use of partial knowledge about the outcome of the experiment. For example, suppose that we are interested in the probability that the share price of Telstra increased today. If we hear on the radio that the All Ordinaries Index rose by 20 points today (but no individual share prices are given), that information will surely affect the probability that the price of Telstra shares has gone up. In light of this new information, we may wish to calculate the **conditional probability** that the price of Telstra shares increased, given that the All Ordinaries Index has risen by 20 points.

For an illustration of *conditional probability*, consider again Example 6.3. Suppose that we are told that the invoice selected by the auditor is a wholesale invoice (W). We may then determine the probability that this invoice is for the deluxe model (D), by making use of our knowledge that it is a wholesale invoice. In other words, we are seeking the conditional probability that D will occur, given that W has occurred; this is expressed as $P(D|W)$. (The vertical stroke $|$ is read ‘given that’, and is followed by the event that has occurred.) In attempting to calculate this conditional probability, we first note that knowing that a wholesale invoice was selected restricts our enquiry to the first column of **Table 6.1**. That is, the new information has reduced the size of the sample space to 60 possible outcomes. Of these 60 possible outcomes in the *reduced* sample space, 36 belong to the event D . Hence, the desired conditional probability is

$$P(D|W) = \frac{36}{60} = 0.6$$

which, as you can see, differs from the (*unconditional*) probability:

$$P(D) = \frac{36+112}{200} = \frac{148}{200} = 0.74$$

We now present a slightly different but equivalent way of calculating $P(D|W)$.

Our calculation of the conditional probability $P(D|W)$ as the ratio of two probabilities follows from our previous calculation of $P(D|W)$:

$$P(D|W) = \frac{36}{60} = \frac{36/200}{60/200} = \frac{P(D \cap W)}{P(W)}$$

Similarly, we can calculate the conditional probability that the invoice selected is to a wholesaler, given that it is for a deluxe model:

$$P(W|D) = \frac{P(D \cap W)}{P(D)} = \frac{0.18}{0.74}$$

Having worked through the calculation of a particular conditional probability, we now present the general formula for conditional probability.

Conditional probability

Let A and B be two events such that $P(B) > 0$. The conditional probability that A occurs, given that B has occurred, is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

The conditional probability that B occurs, given that A has occurred is:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

6.2c Independent events

In the preceding example, we saw that $P(D) = 0.74$ and that $P(D|W) = 0.6$, so that $P(D) \neq P(D|W)$. In other words, knowing that event W occurred changes the probability that D occurred. Such events, D and W , are called **dependent events**. However, if the occurrence of one event does not change the probability of occurrence of the other event, the two events are said to be **independent events**.

If one equality in the preceding definition holds, so does the other. The concept of independence is illustrated in the following example.

dependent events

Events in which the occurrence of one does affect the probability the other will occur.

independent events

Events in which the occurrence of one does not affect the probability the other will occur.

Independent and dependent events

Two events A and B are said to be independent if

$$P(A|B) = P(A) \text{ or } P(B|A) = P(B)$$

Otherwise, the events are dependent.

EXAMPLE 6.4

LO4

Equity in company promotions

A group of female managers working for an insurance company has lodged a complaint with the personnel department. Although the women agree that the company has increased its number of female managers, they assert that women tend to remain in lower-level management positions when promotions are handed out. They have supported their argument by noting that, over the past three years, only eight of the 54 promotions awarded went to women. The personnel department has responded by claiming that these numbers are misleading on two counts: first, there are far fewer female managers than male managers; second, many of the female managers have been hired during the past year, and employees are almost never promoted during their first year at the managerial level. The personnel department has compiled the data shown in **Table 6.3**, which classifies those employed as managers for at least one year, according to gender and promotion record. The department claims that the decision to promote a manager (or not) is independent of the manager's gender. Would you agree?

TABLE 6.3 Classification of managers

Manager	Promoted (A)	Not promoted (\bar{A})	Total
Male (M)	46	184	230
Female (\bar{M})	8	32	40
Total	54	216	270

Solution

The events of interest are as follows:

M : a manager is male

\bar{M} : a manager is female

A : a manager is promoted

\bar{A} : a manager is not promoted

In order to show that the decision about whether or not to promote a manager is independent of the manager's gender, we must verify that:

$$P(A|M) = P(A) \text{ and } P(M|A) = P(M)$$

If this equality holds, the probability that a male manager is promoted is not different from the probability that any manager is promoted. Given no information other than the data in **Table 6.3**, the probability that a manager is promoted is:

$$P(A) = \frac{54}{270} = 0.20$$

If we now consider only male managers, we restrict our attention to the first row of **Table 6.3**. Given that a manager is male, the probability that he is promoted is

$$P(A|M) = \frac{46}{230} = 0.20$$

Notice the distinction between this conditional probability and the joint probability that a manager is both male and promoted, which is $P(A \cap M) = 46/270 = 0.17$. In any case, we have verified that $P(A) = P(A|M)$, so the events A and M are independent. Based on the data in **Table 6.3**, we must agree with the personnel department's contention that there is no discrimination in awarding promotions.

As indicated in the definition of independent events, an alternative way of showing that A and M are independent events is to verify that $P(M|A) = P(M)$. The probability that a manager is male is $P(M) = 230/270 = 46/54$, which equals $P(M|A)$, the probability that a manager who is promoted is male. Thus, we again conclude that the events A and M are independent.

REAL-LIFE APPLICATIONS

Mutual funds

A mutual fund is a pool of investments made on behalf of people who share similar objectives. In most cases, a professional manager who has been educated in finance and statistics manages the fund. He or she makes decisions to buy and sell individual stocks and bonds in accordance with a specified investment philosophy. For example, there are funds that concentrate on other publicly traded mutual fund companies. Other mutual funds specialise in internet stocks (so-called dot-coms), whereas others buy stocks of biotechnology firms. Surprisingly, most mutual funds do not outperform the market; that is,

the increase in the net asset value (NAV) of the mutual fund is often less than the increase in the value of stock indexes that represent their stock markets. One reason for this is the management expense ratio (MER) which is a measure of the costs charged to the fund by the manager to cover expenses, including the salary and bonus of the managers. The MERs for most funds range from 0.5% to more than 4%. The ultimate success of the fund depends on the skill and knowledge of the fund manager. This raises the question, which managers do best?

EXAMPLE 6.5

LO4

Determinants of success among mutual fund managers

Why are some mutual fund managers more successful than others? One possible factor is the university at which the manager earned his or her Master of Business Administration (MBA). Suppose that a potential investor examined the relationship between how well the mutual fund performs and where the fund manager earned his or her MBA. After the analysis, **Table 6.4**, a table of joint probabilities, was developed.

TABLE 6.4 Determinants of success among mutual fund managers

Manager's qualification	Fund performance	
	Mutual fund outperforms market (B_1)	Mutual fund does not outperform market (B_2)
Top-20 MBA programs (A_1)	0.11	0.29
Not top-20 MBA programs (A_2)	0.06	0.54

Source: Adapted from 'Are Some Mutual Fund Managers Better than Others? Cross-Sectional Patterns in Behavior and Performance' by Judith Chevalier and Glenn Ellison, Working paper 5852, National Bureau of Economic Research

- a Analyse the probabilities shown in **Table 6.4** and interpret the results.
- b Suppose that we select one mutual fund at random and discover that it did not outperform the market. What is the probability that a graduate of a top-20 MBA program manages it?
- c Determine whether the event that the manager graduated from a top-20 MBA program and the event the fund outperforms the market are independent events.
- d Determine the probability that a randomly selected fund outperforms the market or the manager graduated from a top-20 MBA program.

Solution

- a **Table 6.4** tells us that the joint probability that a mutual fund outperforms the market and that its manager graduated from a top-20 MBA program is 0.11; that is, 11% of all mutual funds outperform the market and their managers graduated from a top-20 MBA program. The other three joint probabilities are defined similarly. The probability that a mutual fund:

- outperforms the market and its manager did not graduate from a top-20 MBA program is 0.06
- does not outperform the market and its manager graduated from a top-20 MBA program is 0.29
- does not outperform the market and its manager did not graduate from a top-20 MBA program is 0.54.

To help make our task easier, we'll use notation to represent the events. Let:

- A_1 = Fund manager graduated from a top-20 MBA program
- A_2 = Fund manager did not graduate from a top-20 MBA program
- B_1 = Fund outperforms the market
- B_2 = Fund does not outperform the market

Thus, the joint probabilities are

$$\begin{aligned}P(A_1 \cap B_1) &= 0.11 \\P(A_2 \cap B_1) &= 0.06 \\P(A_1 \cap B_2) &= 0.29 \\P(A_2 \cap B_2) &= 0.54\end{aligned}$$

The joint probabilities in **Table 6.4** allow us to compute various probabilities. The marginal probabilities are computed by adding across rows or down columns.

The marginal probabilities that correspond to managers' qualifications are $P(A_1)$ and $P(A_2)$ and can be calculated as:

$$\begin{aligned}P(A_1) &= P(A_1 \cap B_1) + P(A_1 \cap B_2) = 0.11 + 0.29 = 0.40 \\P(A_2) &= P(A_2 \cap B_1) + P(A_2 \cap B_2) = 0.06 + 0.54 = 0.60\end{aligned}$$

Thus, when randomly selecting mutual funds, the probability that its manager graduated from a top-20 MBA program, $P(A_1)$, is 0.40. Expressed as relative frequency, 40% of all mutual fund managers graduated from a top-20 MBA program. The probability $P(A_2)$ tells us that 60% of all mutual fund managers did not graduate from a top-20 MBA program. Notice that the probability that a mutual fund manager graduated from a top-20 MBA program and the probability that the manager did not graduate from a top-20 MBA program add to 1.

The marginal probabilities that correspond to fund performance are $P(B_1)$ and $P(B_2)$ and can be calculated as:

$$\begin{aligned}P(B_1) &= P(A_1 \cap B_1) + P(A_2 \cap B_1) = 0.11 + 0.06 = 0.17 \\P(B_2) &= P(A_1 \cap B_2) + P(A_2 \cap B_2) = 0.29 + 0.54 = 0.83\end{aligned}$$

These marginal probabilities tell us that 17% of all mutual funds outperform the market and that 83% of mutual funds do not outperform the market.

TABLE 6.5 Joint and marginal probabilities

Manager's qualification	Fund performance		Total
	Mutual fund outperforms market (B_1)	Mutual fund does not outperform market (B_2)	
Top-20 MBA programs (A_1)	$P(A_1 \cap B_1) = 0.11$	$P(A_1 \cap B_2) = 0.29$	$P(A_1) = 0.40$
Not top-20 MBA programs (A_2)	$P(A_2 \cap B_1) = 0.06$	$P(A_2 \cap B_2) = 0.54$	$P(A_2) = 0.60$
Total	$P(B_1) = 0.17$	$P(B_2) = 0.83$	1.00

- b We wish to find a conditional probability. The condition is that the fund did not outperform the market (event B_2), and the event whose probability we seek is that the fund is managed by a graduate of a top-20 MBA program (event A_1). Thus, we want to compute the following probability:

$$P(A_1|B_2)$$

Using the conditional probability formula, we find

$$P(A_1|B_2) = \frac{P(A_1 \cap B_2)}{P(B_2)} = \frac{0.29}{0.83} = 0.349$$

Thus, 34.9% of all mutual funds that do not outperform the market are managed by top-20 MBA program graduates.

- c We wish to determine whether A_1 and B_1 are independent. By definition, events A_1 and B_1 are independent if $P(A_1|B_1) = P(A_1)$. We must calculate the probability of A_1 given B_1 ; that is:

$$P(A_1|B_1) = \frac{P(A_1 \cap B_1)}{P(B_1)} = \frac{0.11}{0.17} = 0.647$$

The marginal probability that a manager graduated from a top-20 MBA program is

$$P(A_1) = 0.40$$

As the two probabilities are not equal, we conclude that the two events are dependent.

Incidentally, we could have made the decision by calculating $P(B_1|A_1)$ and observing that it is not equal to $P(B_1)$.

Note that there are three other combinations of events in this problem. They are $(A_1$ and $B_2)$, $(A_2$ and $B_1)$, $(A_2$ and $B_2)$ (ignoring mutually exclusive combinations $(A_1$ and $A_2)$ and $(B_1$ and $B_2)$, which are dependent). In each combination, the two events are dependent. In this type of problem, where there are only four combinations, if one combination is dependent, then all four will be dependent. Similarly, if one combination is independent, then all four will be independent. This rule does not apply to any other situation.

- d We want to compute the probability of the union of two events:

$$P(A_1 \cup B_1)$$

The union A_1 or B_1 (i.e. $A_1 \cup B_1$) consists of three events; that is, the union occurs whenever any of the following joint events occurs:

- 1 The fund outperforms the market and the manager graduated from a top-20 MBA program.
- 2 The fund outperforms the market and the manager did not graduate from a top-20 MBA program.
- 3 The fund does not outperform the market and the manager graduated from a top-20 MBA program.

Their probabilities are given:

$$P(A_1 \cap B_1) = 0.11$$

$$P(A_2 \cap B_1) = 0.06$$

$$P(A_1 \cap B_2) = 0.29$$

Thus, the probability of the union – the fund outperforms the market or the manager graduated from a top-20 MBA program – is the sum of the three probabilities:

$$\begin{aligned} P(A_1 \cup B_1) &= P(A_1 \cap B_1) + P(A_2 \cap B_1) + P(A_1 \cap B_2) \\ &= 0.11 + 0.06 + 0.29 = 0.46 \end{aligned}$$

Notice that there is another way to produce this probability. Of the four probabilities in **Table 6.4**, the only one representing an event that is not part of the union is the probability of the event the fund does not outperform the market and the manager did not graduate from a top-20 MBA program. That probability is

$$P(A_2 \cap B_2) = 0.54$$

which is the probability that the union does not occur. Thus, the probability of the union is

$$P(A_1 \cup B_1) = 1 - P(A_2 \cap B_2) = 1 - 0.54 = 0.46$$

In the next section we introduce the addition rule, which can also be used to produce the required probability $P(A_1 \cup B_1)$. Thus, we determined that 46% of mutual funds either outperform the market or are managed by a top-20 MBA program graduate or have both characteristics.

Before concluding this section, we draw your attention to a common misconception. Students often think that independent events and mutually exclusive events are the same thing. They are not. For example, events A and M in Example 6.4 are independent events but they are not mutually exclusive, since the event $(A \cap M)$ contains 46 outcomes. In fact, it can be shown that *any two independent events A and B that occur with non-zero probabilities cannot be mutually exclusive*. If A and B were mutually exclusive, we would have $P(A \cap B) = 0$ and $P(A|B) = 0$; but since A occurs with non-zero probability, $P(A) \neq P(A|B)$, and so A and B cannot be independent events.

EXERCISES

Learning the techniques

- 6.18** Suppose we have the following joint probabilities:

	A_1	A_2	A_3
B_1	0.1	0.3	0.2
B_2	0.2	0.1	0.1

- a Calculate the marginal probabilities.
- b Calculate $P(A_1|B_1)$.
- c Calculate $P(A_2|B_1)$.
- d Calculate $P(A_3|B_1)$.
- e Did your answers to parts (b), (c) and (d) sum to one? Is this a coincidence? Explain.

- 6.19** Consider a sample space $S = \{E_1, E_2, E_3, E_4\}$ where $P(E_1) = 0.1$, $P(E_2) = 0.1$, $P(E_3) = 0.3$ and $P(E_4) = 0.5$. The events are defined as:

$$\begin{aligned} A &= \{E_1, E_2, E_4\} \\ B &= \{E_1, E_4\} \\ C &= \{E_1, E_2, E_3\} \end{aligned}$$

Calculate the following probabilities:

- a $P(A|B)$
- b $P(B|A)$
- c $P(A|C)$
- d $P(C|A)$
- e $P(B|C)$
- f $P(C|B)$

- 6.20** Suppose that you have been given the following joint probabilities. Are the pairs of events independent? Explain.

	A_1	A_2
B_1	0.20	0.60
B_2	0.05	0.15

- 6.21** Consider a sample space $S = \{E_1, E_2, E_3, E_4\}$ where $P(E_1) = P(E_4) = 0.3$ and $P(E_2) = P(E_3) = 0.2$.

Define these events:

$$\begin{aligned} A &= \{E_1, E_2\} \\ B &= \{E_2, E_3\} \\ C &= \{E_3, E_4\} \end{aligned}$$

Which of the following pairs of events are independent? Explain.

- a A and B
- b B and C
- c A and C

- 6.22** Suppose we have the following joint probabilities.

	A_1	A_2	A_3
B_1	0.15	0.20	0.10
B_2	0.25	0.25	0.05

- a Calculate the marginal probabilities.
- b Calculate $P(A_2|B_2)$.
- c Calculate $P(B_2|A_2)$.
- d Calculate $P(B_1|A_2)$.
- e Calculate $P(A_1 \text{ or } A_2)$.
- f Calculate $P(A_2 \text{ or } B_2)$.
- g Calculate $P(A_3 \text{ or } B_1)$.

Applying the techniques

- 6.23 Self-correcting exercise.** An ordinary deck of playing cards has 13 cards of each suit. Suppose that a card is selected at random from the deck.
- a What is the probability that the card selected is an ace?
 - b Given that the card selected is a spade, what is the probability that the card is an ace?
 - c Are 'an ace is selected' and 'a spade is selected' independent events?

- 6.24** The female lecturers at a large university recently lodged a complaint about the most recent round of promotions from lecturer to senior lecturer. An analysis of the relationship between gender and promotion was undertaken, with the joint probabilities in the following table being produced:

	Promoted	Not promoted
Female	0.03	0.12
Male	0.17	0.68

- a If a female lecturer is selected, what is the probability of being promoted?
- b If a male lecturer is selected, what is the probability of being promoted?
- c Is it reasonable to accuse the university of gender bias?

6.25 A department store analysed its most recent sales and determined the relationship between the way the customer paid for the item and the price category of the item. The joint probabilities in the following table were calculated.

Size of purchase	Method of payment		
	Cash	Credit card	Debit card
Under \$20	0.09	0.03	0.04
\$20–\$100	0.05	0.21	0.18
Over \$100	0.03	0.23	0.14

- a Find the proportion of purchases paid for by debit card.
- b Find the probability of an 'over \$100' purchase, given that it is a credit card purchase.
- c Determine the proportion of purchases made by credit card or by debit card.
- d Are the events 'payment by cash' and 'purchase of under \$20' mutually exclusive? Explain.
- e Are the events 'payment by cash' and 'purchase of under \$20' independent? Explain.

6.26 The following table lists the probabilities of employment for males and females and their industry of employment.

Proportion of employment in various industries by gender, New Zealand, September quarter, 2014

Industry	Male	Female
Agriculture	0.078	0.035
Manufacturing	0.146	0.058
Electricity, gas and water	0.011	0.006
Construction	0.150	0.024
Wholesale and retail trade	0.053	0.025
Retail trade	0.122	0.165
Transport and communication	0.060	0.023
Information and media services	0.021	0.016
Mining	0.006	0.001
Total industry	0.647	0.353

Source: Stats NZ and licensed by Stats NZ for reuse under the Creative Commons Attribution 4.0 International licence. CC BY 4.0 <https://creativecommons.org/licenses/by/4.0/>

- a If one person is selected at random, what is the probability that he or she works for the construction industry?

- b If a female is selected at random, what is the probability that she works for the transport and communication industry?
- c If a person who works in information and media services is selected at random, what is the probability that this person is a male?

6.27 Of a company's employees, 30% are women and 6% are married women. Suppose that an employee is selected at random. If the employee who is selected is a woman, what is the probability that she is married?

6.28 A firm classifies its customers' accounts according to the balance outstanding and according to whether or not the account is overdue. The table below gives the proportion of accounts falling into various categories.

Account balance	Overdue	Not overdue
Under \$100	0.08	0.42
\$100–\$500	0.08	0.22
Over \$500	0.04	0.16

One account is selected at random.

- a Given that the account selected is overdue, what is the probability that its balance is under \$100?
- b If the balance of the account selected is over \$500, what is the probability that it is overdue?
- c If the balance of the account selected is \$500 or less, what is the probability that it is overdue?

6.29 A personnel manager of a manufacturing company has cross-classified the 400 employees of his company according to their record of absenteeism last year and whether or not they were smokers, as shown in the following table:

Number of days absent	Smoker	Non-smoker
Less than 10	34	260
10 or more	78	28

One of these employees is selected at random.

- a What is the probability that the employee selected was a non-smoker?
- b What is the probability that the employee selected was absent for 10 or more days?
- c Are the events 'non-smoker' and 'absent fewer than 10 days' mutually exclusive? Explain.
- d Determine whether an employee's absence for 10 or more days last year was independent of whether the employee was a smoker.

- e What is the probability that the employee selected was absent for fewer than 10 days given that the employee is a smoker?
- f What is the probability that the employee selected was absent for fewer than 10 days given that the employee does not smoke?

6.30 Refer to Exercise 6.12 (on page 222).

- a Determine whether F and M are independent or dependent events. Explain.
- b Repeat part (a) for events F and B .

6.31 A firm has classified its customers in two ways:

(1) according to whether the account is overdue, and (2) whether the account is new (less than 12 months) or old. An analysis of the firm's records provided the input for the following table of joint probabilities.

	Overdue	Not overdue
New	0.06	0.13
Old	0.52	0.29

One account is selected at random.

- a If the account is overdue, what is the probability that it is new?
- b If the account is new, what is the probability that it is overdue?
- c Is the age of the account related to whether it is overdue? Explain.

6.32 A restaurant chain routinely surveys customers and, among other questions, asks each customer to rate the quality of the food, and to state whether they would return. The following table summarises the survey responses in the form of a joint probability table.

Rating	Customer will return	Customer will not return
Poor	0.02	0.10
Fair	0.08	0.09
Good	0.35	0.14
Excellent	0.20	0.02

- a What proportion of customers rated the restaurant's food as good and said that they would return?
- b What proportion of customers who said that they would return rated the restaurant's food as good?
- c What proportion of customers who rated the restaurant's food as good said that they would return?
- d Discuss the differences in your answers to parts (a), (b) and (c).

6.33 Credit scorecards are used by financial institutions to help decide to whom loans should be granted. An analysis of the records of one bank produced the following probabilities.

Loan performance	Score	
	Under 400	400 or more
Fully repaid	0.19	0.64
Defaulted	0.13	0.04

- a What proportion of loans are fully repaid?
- b What proportion of loans given to scorers of less than 400 are fully repaid?
- c What proportion of loans given to scorers of 400 or more are fully repaid?
- d Are score and whether the loan is fully repaid independent? Explain.

6.34 According to labour market statistics published by the Australian Bureau of Statistics (ABS), the following data about the employment status of Australian men and women in May 2019 were obtained.

Type of person	Number ('000)	
	Male	Female
Employed	6869	6067
Unemployed	369	326
Not in the labour force	2860	4070

Source: Australian Bureau of Statistics, May 2019, *Labour Force*, cat. no. 6202.0, ABS, Canberra

- a What proportion of Australians are unemployed?
- b What is the probability that a person is unemployed, given that the person is a female?
- c If a male person in Australia is selected at random, what is the probability that he will be unemployed?

6.35 An analysis of the relationship between the gender of the shareholders and whether a vote was cast in the elections held during the company's last annual general meeting (AGM) produced the probabilities shown in the table.

- a What proportion of the shareholders voted in the last election?
- b Are gender and whether a vote was cast in the last AGM election independent? Explain how you arrived at your conclusion.

	Female	Male
Voted in last AGM election	0.25	0.18
Did not vote in last AGM election	0.33	0.24

- 6.36** The method of instruction in teaching statistics courses is changing. Historically, most courses were taught with an emphasis on manual calculation. The alternative is to use a computer and a software package to perform the calculations. An analysis of applied statistics courses and their lecturers investigated whether each lecturer's educational background was primarily statistics or some other field. The result of this analysis is the table of joint probabilities below.

	Manual calculations	Using the computer and software
Statistics education	0.23	0.36
Other education	0.11	0.30

- a** What is the probability that a randomly selected statistics course lecturer, whose education was in statistics, emphasises manual calculations?
- b** What proportion of statistics courses use a computer and software?
- c** Are the educational background of the lecturer and the way his or her course is taught independent?
- 6.37** To determine whether drinking alcoholic beverages has an effect on the bacteria that cause ulcers, researchers developed the following table of joint probabilities:

Number of alcoholic drinks per day	Ulcer	No ulcer
None	0.01	0.22
One	0.03	0.19
Two	0.03	0.32
More than two	0.04	0.16

- a** What proportion of people have ulcers?
- b** What is the probability that a teetotaller (no alcoholic beverages) develops an ulcer?
- c** What is the probability that someone who has an ulcer does not drink alcohol?
- d** What is the probability that someone who does not have an ulcer consumes two alcoholic drinks per day?

- 6.38** When drivers are lost, what do they do? After a thorough analysis, the following joint probabilities were developed:

Action	Male	Female
Consult a map	0.25	0.14
Ask for directions	0.12	0.28
Continue driving until direction or location is determined	0.13	0.08

- a** What is the probability that a driver would ask for directions given that the driver is a male?
- b** What proportion of drivers consult a map?
- c** Are gender and consulting a map independent? Explain.

- 6.39** Many critics of television claim that there is too much violence and that this has a negative impact on society. However, some say there may also be a negative effect on advertisers. To examine this issue, researchers developed two versions of a cops-and-robbers made-for-television movie. One version depicted several violent crimes, and the other removed these scenes. In the middle of the movie, a 60-second commercial was shown advertising a new product. At the end of the movie, viewers were asked to name the brand of the new product. Based on the reviewers' responses, the following table of joint probabilities was produced:

	Watched violent movie	Watched non-violent movie
Remember the brand name	0.15	0.18
Do not remember the brand name	0.35	0.32

- a** What proportion of viewers remembered the brand name?
- b** What proportion of viewers who watched the violent movie remembered the brand name?
- c** Does watching a violent movie affect whether the viewer will remember the brand name? Explain.

- 6.40** How are the size of a firm (measured in terms of number of employees) and the type of firm related? To help answer this question an analyst developed the following table of joint probabilities:

Number of employees	Industry		
	Construction	Manufacturing	Retail
Less than 20	0.231	0.099	0.501
20–99	0.019	0.035	0.088
100 or more	0.002	0.015	0.011

A firm is selected at random. Find the probability of the following events.

- a The firm employs fewer than 20 employees.
- b The firm is in the retail industry.
- c A firm in the construction industry employs between 20 and 99 workers.

6.41 A retail outlet wanted to know whether its weekly advertisement in the local newspaper was working.

To acquire this critical information, the store manager surveyed the people who entered the store and asked each individual whether he/she had seen the advertisement and recorded whether a purchase was made. From the information obtained, the manager produced the following table of joint probabilities:

	Purchase	No purchase
Saw advertisement	0.18	0.42
Did not see advertisement	0.12	0.28

Are the ads effective? Explain.

6.3 Rules of probability

In Section 6.1 we introduced intersection and union, and described how to determine the probability of the intersection and the union of two events. In this section we present three rules of probability that enable us to calculate the probability of more complex events.

6.3a Complement rule

The first rule of probability follows easily from the basic requirement that the sum of the probability assigned to each possible outcome in a sample space must equal one. Given any event A and its complement \bar{A} , each outcome in the sample space must belong to either A or \bar{A} . We must therefore have

$$P(A) + P(\bar{A}) = 1$$

The **complement rule** is obtained by subtracting $P(A)$ from each side of the equality.

Complement rule

For any event A :

$$P(\bar{A}) = 1 - P(A)$$

EXAMPLE 6.6

LO5

Flipping a coin to observe a head

Suppose that we intend to flip a coin until heads comes up for the first time. How would we determine the probability that at least two flips will be required to obtain a head?

Solution

A possible sample space for this experiment is $S = \{1, 2, 3, 4, \dots\}$, where each integer indicates a possible number of flips required to obtain a head for the first time. Let A be the event that the coin will show a head in



the first flip – that is, $A = \{1\}$. Also consider event B that at least two flips are required – that is, $B = \{2, 3, 4, \dots\}$. A direct approach to finding $P(B)$ would entail calculating and summing the probabilities $P(2), P(3), P(4), \dots$. A simpler approach, however, is to recognise the fact that $B = \bar{A}$ and the probability that event A will not occur (\bar{A}) is one minus the probability that A will occur $[P(\bar{A}) = 1 - P(A)]$. So we have:

$$P(A) = P(1) = \frac{1}{2}$$

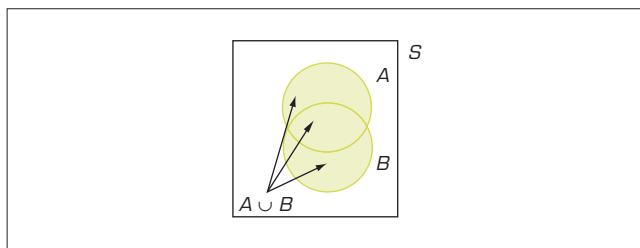
Therefore, the probability that at least two flips will be required before heads comes up for the first time is

$$P(B) = P(\bar{A}) = 1 - P(A) = \frac{1}{2}$$

6.3b Addition rule

The second rule of probability enables us to find the probability of the union of two events from the probability of other events. In the Venn diagram in **Figure 6.11**, the union of events A and B , $A \cup B$, is represented by the entire shaded area. When finding the probability $P(A \cup B)$ by summing $P(A)$ and $P(B)$, we must subtract $P(A \cap B)$ to avoid double-counting the probability of the event $A \cap B$, which belongs to both A and B .

FIGURE 6.11 Entire shaded area is $A \cup B$



Addition rule

For any two events A and B ,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If A and B are mutually exclusive, the event $A \cap B$ is an empty set and $P(A \cap B) = 0$. Therefore, for two mutually exclusive events A and B , the **addition rule** simplifies to:

$$P(A \cup B) = P(A) + P(B)$$

Addition rule for mutually exclusive events

If two events A and B are mutually exclusive,

$$P(A \cup B) = P(A) + P(B)$$

This is the **addition rule for mutually exclusive events**.

addition rule

The rule that allows us to calculate the probability of the union of two events.

addition rule for mutually exclusive events

The rule that allows us to calculate the probability of the union of two mutually exclusive events.

EXAMPLE 6.7

LO5

Applying the addition rule

Refer to Example 6.3, involving the audit of Keep Kool Pty Ltd. Find the probability that the invoice selected is either a wholesale invoice or an invoice for the deluxe model.

Solution

We need to find $P(W \cup D)$. For convenience, **Tables 6.1** and **6.2** are reproduced in **Table 6.6**.

TABLE 6.6 Classification of invoices

Model	(a) Invoice frequencies			Model	(b) Invoice probabilities		
	Wholesale (W)	Retail (\bar{W})	Total		Wholesale (W)	Retail (\bar{W})	Total
Deluxe (D)	36	112	148	Deluxe (D)	0.18	0.56	0.74
Standard (\bar{D})	24	28	52	Standard (\bar{D})	0.12	0.14	0.26
Total	60	140	200	Total	0.30	0.70	1.00

We know from **Table 6.6(b)** that $P(W) = 0.30$, $P(D) = 0.74$ and $P(W \cap D) = 0.18$. Using the addition rule,

$$P(W \cup D) = P(W) + P(D) - P(W \cap D) = 0.30 + 0.74 - 0.18 = 0.86$$

Thus, the probability that the invoice selected is either a wholesale invoice or an invoice for the deluxe model is 0.86.

To check the answer by means of the basic counting procedure, let us count the number of outcomes, using **Table 6.6(a)**. The number of outcomes in W is 60, and the number in D is 148. But the number of points in the event $(W \cap D)$ is $60 + 148 - 36$, as 36 outcomes are common to both W and D . We therefore have:

$$P(W \cup D) = \frac{172}{200} = 0.86$$

6.3c Multiplication rule

The third rule of probability, which is used to find the probability of a joint event, is simply a rearrangement of the definition of conditional probability. As

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

and

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

we obtain the following rule for computing the joint probability $P(A \cap B)$.

Multiplication rule

For any two events A and B ,

$$P(A \cap B) = P(B|A)P(A) \text{ or } P(A \cap B) = P(A|B)P(B)$$

Note that the two expressions for finding a joint probability using the **multiplication rule** are equivalent. The expression that should be used in a particular situation depends on the information that is given.

For the special case in which A and B are independent events, we have $P(B|A) = P(B)$, so we can simply write $P(A \cap B) = P(A)P(B)$.

Multiplication rule for independent events

If two events A and B are independent,

$$P(A \cap B) = P(A)P(B)$$

This is the **multiplication rule for independent events**.

multiplication rule

The rule that allows us to calculate the probability of the intersection of two events.

multiplication rule for independent events

The rule that allows us to calculate the probability of the intersection of two independent events.

EXAMPLE 6.8

LO5

Applying the basic probability rules

A computer software supplier has developed a new record-keeping package for use by hospitals. The company feels that the probability that the new package will show a profit in its first year is 0.6, unless a competitor introduces a product of comparable quality this year, in which case the probability of a first-year profit drops to 0.3. The supplier suggests that there is a 50/50 chance that a comparable product will be introduced this year. We define the following events:

- A : A competitor introduces a comparable product.
- B : The record-keeping package is profitable in its first year.

Calculate the following probabilities:

- What is the probability that both A and B will occur?
- What is the probability that either A or B will occur?
- What is the probability that neither A nor B will occur?

Solution

Summarising the given information, we know that

$$P(A) = 0.5$$

$$P(B) = 0.6$$

$$P(B|A) = 0.3$$

- According to the multiplication rule, the probability that a competitor will introduce a comparable product and that the first year will be profitable is

$$\begin{aligned} P(A \cap B) &= P(B|A)P(A) \\ &= (0.3)(0.5) \\ &= 0.15 \end{aligned}$$

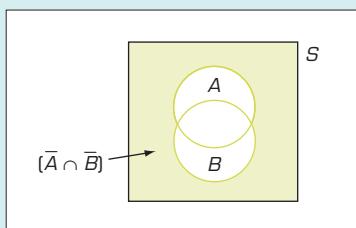
- Notice that $P(A \cup B)$ can be determined only after $P(A \cap B)$ has been calculated. The probability that either a competitor will introduce a comparable product or the record-keeping package will be profitable in its first year is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.5 + 0.6 - 0.15 = 0.95$$





FIGURE 6.12 Shaded area is $(\bar{A} \cap \bar{B}) = (\overline{A \cup B})$



- c** This part is somewhat more difficult, but it illustrates the effective use of event relations. The easiest way to find $P(\bar{A} \cap \bar{B})$ – the probability that neither A nor B will occur – is to recognise that $P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B})$ and to use the complement rule. You should convince yourself that events $(\bar{A} \cap \bar{B})$ and $(\overline{A \cup B})$ are the same, with the help of the Venn diagram in **Figure 6.12**. It follows that the probability that a competitor will not introduce a comparable product and that the first year will not be profitable is

$$P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - 0.95 = 0.05$$

EXERCISES

Learning the techniques

- 6.42** If $P(A) = 0.6$, $P(B) = 0.5$ and $P(A \cup B) = 0.9$, find $P(A \cap B)$ and $P(A|B)$.
- 6.43** If $P(A) = 0.2$, $P(B) = 0.4$ and $P(A \cup B) = 0.5$, find $P(A \cap B)$ and $P(B|A)$.
- 6.44** Given that $P(A) = 0.3$, $P(B) = 0.6$ and $P(B|A) = 0.4$, find:
- a $P(A \cap B)$
 - b $P(A \cup B)$
 - c $P(A|B)$
 - d $P(\bar{A} \cap \bar{B})$
- 6.45** Given that $P(A) = 0.4$, $P(B) = 0.5$ and $P(B|A) = 0.8$, find:
- a $P(A \cap B)$
 - b $P(A \cup B)$
 - c $P(A|B)$
 - d $P(\bar{A} \cap \bar{B})$
- 6.46** Let A and B be two mutually exclusive events for which $P(A) = 0.25$ and $P(B) = 0.6$. Find:
- a $P(A \cap B)$
 - b $P(A \cup B)$
 - c $P(A|B)$
- 6.47** Let A and B be two independent events for which $P(A) = 0.25$ and $P(B) = 0.6$. Find:
- a $P(A|B)$
 - b $P(B|A)$
 - c $P(A \cap B)$
 - d $P(A \cup B)$

- 6.48** Let A and B be two mutually exclusive events for which $P(A) = 0.15$ and $P(B) = 0.4$. Find:

- a $P(A \cap B)$
- b $P(A \cup B)$
- c $P(B|A)$

- 6.49** Let A and B be two independent events for which $P(A) = 0.15$ and $P(B) = 0.4$. Find:

- a $P(A|B)$
- b $P(B|A)$
- c $P(A \cap B)$
- d $P(A \cup B)$

Applying the techniques

- 6.50** An international aerospace company has submitted bids on two separate federal government defence contracts, A and B .

The company feels that it has a 60% chance of winning contract A and a 30% chance of winning contract B . If the company wins contract B , the company believes it would then have an 80% chance of winning contract A .

- a What is the probability that the company will win both contracts?
- b What is the probability that the company will win at least one of the two contracts?
- c If the company wins contract B , what is the probability that it will not win contract A ?

- 6.51** **Self-correcting exercise.** Suppose that the aerospace company in Exercise 6.50 feels that it has a 50% chance of winning contract A and a 40% chance of winning contract B . Furthermore, suppose

it believes that winning contract A is independent of winning contract B .

- a What is the probability that the company will win both contracts?
- b What is the probability that the company will win at least one of the two contracts?

6.52 A sporting goods store estimates that 10% of the students at a nearby university play tennis and 5% play cricket. Of those who play tennis, 40% also play cricket.

- a What percentage of the students play both tennis and cricket?
- b What percentage of the students do not play either of these two games?

6.53 The union executive of a mining company conducted a survey of union members to determine what the members felt were the important issues to be discussed during forthcoming negotiations with management. Results showed that 74% felt that job security was an important issue, while 65% felt that accident compensation/benefits were an important

issue. Of those who felt that accident compensation/benefits were an important issue, 60% also felt that job security was an important issue.

- a What percentage of the members felt that both job security and accident compensation/benefits were important?
- b What percentage of the members felt that at least one of these two issues was important?

6.54 Two six-sided dice are rolled, and the number of spots turning up on each is observed. Determine the probability of observing four spots on at least one of the dice. (*Hint:* Use the complement rule.)

6.55 A certain city has a milk and newspaper home delivery service. During the summer months, it is estimated that 20% of the households order milk and 60% order newspapers. Of those who order milk, 80% also order newspapers. What proportion of households:

- a order both milk and newspapers?
- b order at most one of the two?
- c order neither milk nor newspapers?

6.4 Probability trees

Another useful method of calculating probabilities is to use a probability tree in which the various possible events of an experiment are represented by lines or branches of the tree. When you want to construct a sample space for an experiment, a **probability tree** is a useful device for ensuring that you have identified all possible outcomes and have assigned the associated probabilities. The mechanics of using a probability tree will now be illustrated by reference to the simple coin example encountered earlier.

In Example 6.1, we stated that a sample space for the random experiment consisting of flipping a coin twice is $S = \{HH, HT, TH, TT\}$, where the first letter of each pair denotes the result of the first flip. An alternative representation of S , differing only in the notation used, is

$$S = \{H_1 \cap H_2, H_1 \cap T_2, T_1 \cap H_2, T_1 \cap T_2\}$$

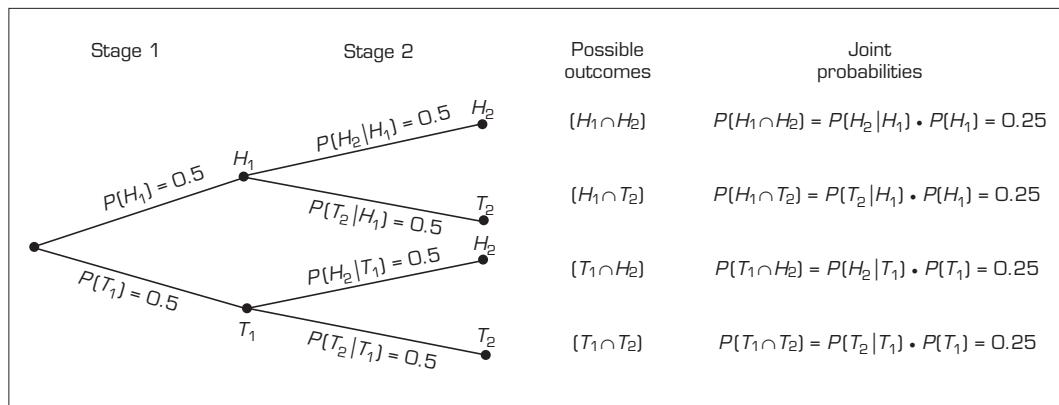
where the events are defined as:

- H_1 : Heads is observed on the first flip.
- H_2 : Heads is observed on the second flip.
- T_1 : Tails is observed on the first flip.
- T_2 : Tails is observed on the second flip.

A probability tree for this example is shown in **Figure 6.13**.

Whenever the process of observing the result of an experiment can be broken down into stages, with a different aspect of the result being observed at each stage, the various possible sequences of observations can be represented with a probability tree. In the coin example, stage 1 observes the outcome of the first flip, while stage 2 observes the outcome of the second flip. The heavy dots in **Figure 6.13** are called nodes, and the branches emanating from a particular node represent the alternative outcomes that may occur at the point. The probability attached to each branch is the conditional probability that a particular branch outcome will occur, given that the outcomes represented by preceding branches have all occurred. For example, the probability attached to the top branch at stage 2 is $P(H_2|H_1) = 0.5$.

probability tree
A depiction of events as branches of a tree.

FIGURE 6.13 Probability tree for coin sample

This is the probability of obtaining a result of heads on the second flip, given that a result of heads was obtained on the first flip. Since the branches emanating from any particular node represent all possible outcomes that may occur at that point, the sum of the probabilities on those branches must equal one.

origin

The initial node of the experiment.

The initial (unlabelled) node is called the **origin**. Any path through the tree from the origin to a terminal node corresponds to one possible outcome, the probability of which is the product of the probabilities attached to the branches forming that path. For example, if we follow along the top two branches of the tree, we observe the outcome H_1 and H_2 , which (according to the multiplication rule) has the probability:

$$P(H_1 \cap H_2) = P(H_2 | H_1) \cdot P(H_1) = (0.5)(0.5) = 0.25$$

EXAMPLE 6.9

LO6

Buying behaviour of men's clothing customers

The manager of a men's clothing store has recorded the buying behaviour of customers over a long period of time. He has established that the probability that a customer will buy a shirt is about 0.4. A customer buys a tie 50% of the time when a shirt is purchased, but only 10% of the time when a shirt is not purchased. Find the probability that a customer will buy:

- a a shirt and a tie
- b a tie
- c a shirt or a tie
- d a tie but not a shirt.

Solution

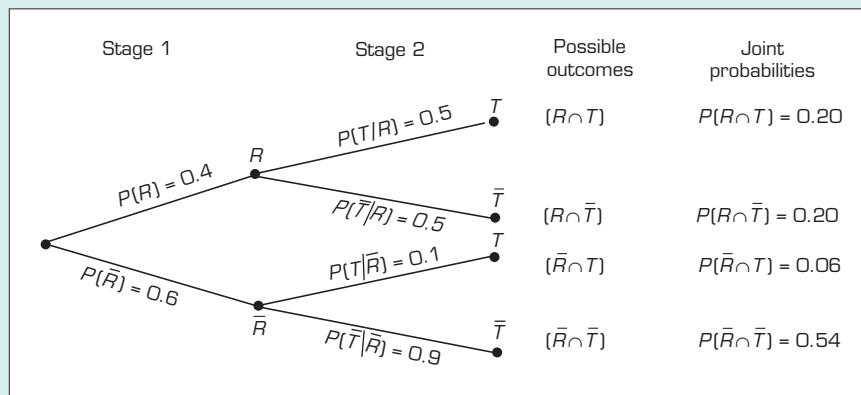
The random experiment consists of observing, for each customer, whether a shirt is purchased and whether a tie is purchased. The two basic events of interest are therefore:

R : A customer buys a shirt.

T : A customer buys a tie.

Summarising the given information, we have $P(R) = 0.4$, $P(T|R) = 0.5$ and $P(T|\bar{R}) = 0.1$. Before constructing a probability tree, we should search the given probabilities for an unconditional probability. In this example, we are given the unconditional probability that event R will occur, so event R will be considered at the first stage and event T at the second. We can now construct the probability tree for this experiment, as shown below in

Figure 6.14.

**FIGURE 6.14** Probability tree for shirts and ties

Collecting the outcomes at the end of the tree, we find that a sample space for the experiment is

$$S = \{(R \cap T), (R \cap \bar{T}), (\bar{R} \cap T), (\bar{R} \cap \bar{T})\}$$

Calculating the required probabilities is now quite simple, amounting to little more than reading off the joint (outcome) probabilities from the tree.

- a The probability that a customer will buy a shirt and a tie is

$$P(R \cap T) = 0.20$$

- b The probability that a customer will buy a tie is

$$\begin{aligned} P(T) &= P\{(R \cap T), (\bar{R} \cap T)\} \\ &= P(R \cap T) + P(\bar{R} \cap T) \\ &= 0.20 + 0.06 \\ &= 0.26 \end{aligned}$$

- c The probability that a customer will buy a shirt or a tie is

$$\begin{aligned} P(R \cup T) &= P\{(R \cap T), (R \cap \bar{T}), (\bar{R} \cap T)\} \\ &= P(R \cap T) + P(R \cap \bar{T}) + P(\bar{R} \cap T) \\ &= 0.20 + 0.20 + 0.06 \\ &= 0.46 \end{aligned}$$

Alternatively,

$$\begin{aligned} P(R \cup T) &= P(R) + P(T) - P(R \cap T) \\ &= 0.40 + 0.26 - 0.20 \\ &= 0.46 \end{aligned}$$

- d The probability that a customer will buy a tie but not a shirt is

$$P(\bar{R} \cap T) = 0.06$$

EXAMPLE 6.10

LOG

Selecting two coins from four

Suppose that you have four coins in a box: a one-dollar coin (D), two 50-cent coins (F) and a 20-cent coin (T). If you select two coins from the box without looking, what is the probability of you selecting a 50-cent coin and a 20-cent coin?

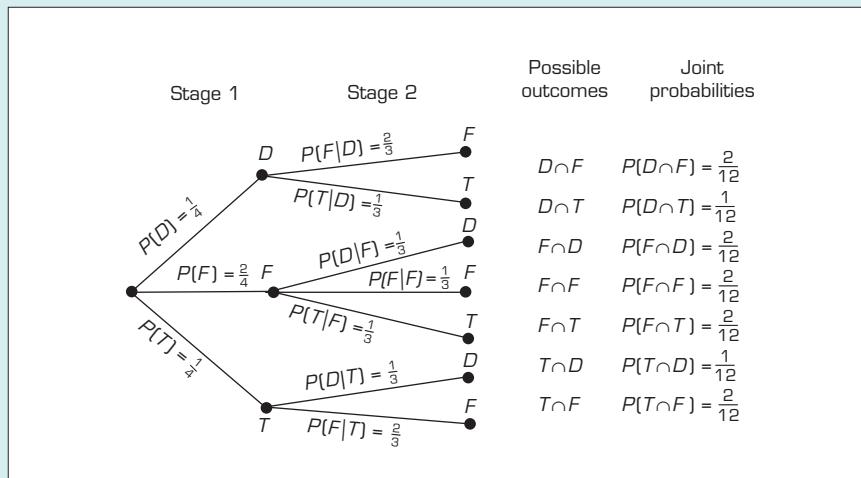
Solution

An organised procedure for solving this problem entails first constructing a relevant sample space, and then assigning probabilities to the outcomes.

For this example, the process of observing which two coins are selected can be broken down into two stages, as shown below in **Figure 6.15**. Once the branches are drawn, assigning the corresponding branch probabilities becomes relatively easy. For example, consider the top branch of the second stage. If the first coin selected is a one-dollar (D) coin, then two 50-cent (F) coins and a 20-cent (T) coin remain in the box. Thus $P(F|D) = 2/3$. Next, we identify the seven outcomes corresponding to the seven terminal nodes, and we obtain the sample space:

$$S = \{D \cap F, D \cap T, F \cap D, F \cap F, F \cap T, T \cap D, T \cap F\}$$

FIGURE 6.15 Probability tree for selecting two coins from four



Finally, we obtain the probability of each outcome in S by taking the product of the probabilities attached to the branches leading to that outcome. This is simply an application of the multiplication rule. For example:

$$P(D \cap F) = P(F|D) \cdot P(D) = (2/3)(1/4) = 2/12$$

Having constructed the probability tree, we can next determine the probability of selecting a 50-cent coin and a 20-cent coin. Let event E be the selection of a 50-cent coin and a 20-cent coin; then $E = \{F \cap T, T \cap F\}$, so:

$$P(E) = P(F \cap T) + P(T \cap F) = 2/12 + 2/12 = 1/3$$

EXERCISES**Learning the techniques**

- 6.56** Given the following probabilities, find the joint probability $P(A \cap B)$.

$$P(A) = 0.7$$

$$P(B|A) = 0.3$$

- 6.57** Given the following probabilities, calculate all the joint probabilities.

$$P(A) = 0.9$$

$$P(\bar{A}) = 0.1$$

$$P(B|A) = 0.4$$

$$P(B|\bar{A}) = 0.7$$

- 6.58** Determine all the joint probabilities from the following:

$$P(A) = 0.8$$

$$P(\bar{A}) = 0.2$$

$$P(B|A) = 0.4$$

$$P(B|\bar{A}) = 0.7$$

- 6.59** Let $P(A) = 0.6$, $P(B|A) = 0.1$ and $P(B|\bar{A}) = 0.3$.

- a Sketch a properly labelled probability tree to depict this situation.
- b Use the probability tree to find $P(A \cap \bar{B})$ and $P(\bar{B})$.

- 6.60** Let $P(\bar{A}) = 0.7$, $P(\bar{B}|A) = 0.8$ and $P(B|\bar{A}) = 0.4$.

- a Sketch a properly labelled probability tree to depict this situation.
- b Use the probability tree to find $P(A \cap B)$ and $P(\bar{B})$.

- 6.61** Let $P(A) = 0.3$, $P(B|A) = 0.4$ and $P(B|\bar{A}) = 0.8$.

- a Sketch a properly labelled probability tree to depict this situation.
- b Use the probability tree to find $P(A \cup B)$ and $P(\bar{A} \cap B)$.

- 6.62** Let $P(\bar{A}) = 0.4$, $P(\bar{B}|A) = 0.3$ and $P(B|\bar{A}) = 0.6$.

- a Sketch a properly labelled probability tree to depict this situation.
- b Use the probability tree to find $P(A \cup B)$ and $P(\bar{A} \cap B)$.

- 6.63** A fair coin is flipped three times. Use a probability tree to find the probability of observing:

- a no heads
- b exactly one head
- c exactly two heads
- d at least one tail.

Applying the techniques

- 6.64 Self-correcting exercise.** An assembler has been supplied with 10 electronic components, of which three are defective. If two components are selected at random, what is the probability that neither component is defective?

- 6.65** Approximately three out of every four Australians who filed a tax return received a refund in a particular year. If three individuals are chosen at random from among those who filed a tax return that year, find the probabilities of the following events:

- a All three received a refund.
- b None of the three received a refund.
- c Exactly one received a refund.

- 6.66** A door-to-door saleswoman sells carpet shampoo in three tube sizes: small, large and giant. The probability of finding a person at home is 0.6. If the saleswoman does find someone at home, the probabilities are 0.5 that she will make no sale, 0.2 that she will sell a small tube, 0.2 that she will sell a large tube, and 0.1 that she will sell a giant tube. The probability of selling more than one tube of shampoo at a house is 0.

- a Find the probability that, in one call, she will not sell any shampoo.
- b Find the probability that, in one call, she will sell either a large tube or a giant tube.

- 6.67** To determine who pays for coffee, three students each toss a coin and the odd person pays. If all coins show heads or all show tails, the students toss again. What is the probability that a decision will be reached in five or fewer tosses?

- 6.68** Of 20000 small businesses surveyed, about 82% said they employed women in some capacity. Of the businesses that employed women, 19.5% employed no female supervisors, 50% employed only one female supervisor, and the remainder employed more than one female supervisor.

- a How many of the businesses surveyed employed no women?
- b What proportion of businesses surveyed employed exactly one female supervisor?
- c What proportion of businesses surveyed employed no female supervisors?
- d Given that a particular firm employed women, what is the probability that it employed at least one female supervisor?

- 6.69** A telemarketer calls people and tries to sell them a subscription to a daily newspaper. On 20% of his calls, there is no answer or the line is busy. He sells subscriptions to 5% of the remaining calls. For what proportion of calls does he make a sale?

- 6.70** A financial analyst estimates that the probability that the economy will experience a recession within the next 12 months is 25%. She also believes that if the economy encounters a recession the probability that her mutual fund will increase in value is 20%. If there is no recession, the probability that the mutual fund will increase in value is 75%. Find the probability that the value of the mutual fund will increase.

6.5 Bayes' law

In Section 6.2 we learnt that the conditional probability that a particular event A will occur, given that B has occurred, is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Bayes' law

A method of revising probabilities after another event occurs.

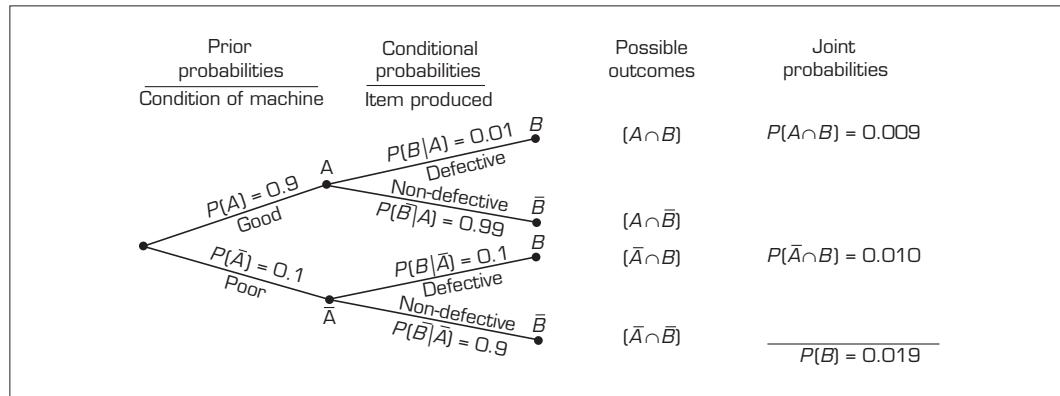
Bayes' law, named after the English clergyman and mathematician Thomas Bayes, provides us with an alternative formula for calculating this conditional probability. The alternative formula simplifies the calculation of $P(A|B)$ when $P(A \cap B)$ and $P(B)$ are not given directly.

Suppose that we are interested in the condition of a machine that produces a particular item. Let A be the event 'the machine is in good operating condition', then \bar{A} represents the event 'the machine is not in good operating condition'. We might know from experience that the machine is in good condition 90% of the time. That is, the initial or *prior probabilities* regarding the machine's condition are $P(A) = 0.9$ and $P(\bar{A}) = 0.1$. Given the machine's condition, we might also know the probability that a defective item will be produced (event B). Suppose that only 1% of the items produced are defective when the machine is in good condition, and 10% are defective when the machine is in poor condition. We therefore have the following conditional probabilities:

$$\begin{aligned} P(B|A) &= 0.01 \\ P(B|\bar{A}) &= 0.10 \end{aligned}$$

The situation just described can be treated as a two-stage experiment and represented by a probability tree, as shown below in **Figure 6.16**.

FIGURE 6.16 Probability tree for machine producing items



We are primarily concerned with the machine's condition, and we know from historical information that there is a 90% chance that it is in good condition. We can get a better idea of the likelihood that the machine is in good condition right now, however, by obtaining more information. Suppose that, without knowing the condition of the machine, we select an item from the current production run and observe that it is defective. It is then possible to revise the prior probability that the machine is in good condition (event A) in light of the new information that event B has occurred. That is, we can find the revised or *posterior probability*:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

The value of the numerator is obtained easily from the probability tree:

$$P(A \cap B) = P(B|A) \cdot P(A) = (0.01)(0.9) = 0.009$$

We next note that event B occurs only if one of two outcomes, $(A \cap B)$ or $(\bar{A} \cap B)$, occurs. The denominator is therefore:

$$\begin{aligned} P(B) &= P(A \cap B) + P(\bar{A} \cap B) \\ &= [P(B|A) \cdot P(A)] + [P(B|\bar{A}) \cdot P(\bar{A})] \\ &= (0.01)(0.9) + (0.1)(0.1) \\ &= 0.019 \end{aligned}$$

By using the rules of addition and multiplication, or simply by reading from the probability tree, we obtain:

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} = \frac{P(A \cap B)}{P(A \cap B) + P(\bar{A} \cap B)} \\ &= \frac{P(A) \cdot P(B|A)}{[P(A) \cdot P(B|A)] + [P(\bar{A}) \cdot P(B|\bar{A})]} = \frac{0.009}{0.019} = 0.47 \end{aligned}$$

In light of the sample information, we have drastically revised down the probability that the machine is currently in good condition from 0.9 to 0.47. Based on this posterior (after sampling) probability of 0.47, it is likely to be worth paying a mechanic to check and repair the machine.

The formula just developed, which we used to find the revised or posterior probability of A given that B has occurred, is the formulation of Bayes' law for the case in which the first stage consists of only two events. The two events in our example are A and \bar{A} , which could also have been denoted A_1 and A_2 , in order to fit the notation of the general formula for Bayes' law.

Bayes' law can be generalised to the situation in which the first stage consists of n mutually exclusive events A_1, A_2, \dots, A_n . The following statement of Bayes' law gives the formula for finding the posterior probabilities $P(A_i|B)$ where B is an event observed at the second stage.

Bayes' law

If A_i is one of the n mutually exclusive events A_1, A_2, \dots, A_n at the first stage, and B is an event at the second stage, then

$$P(A_i|B) = \frac{P(A_i) \cdot P(B|A_i)}{[P(A_1) \cdot P(B|A_1)] + [P(A_2) \cdot P(B|A_2)] + \dots + [P(A_n) \cdot P(B|A_n)]}$$

Before presenting a second example, we should note that using the probability tree approach is much easier than memorising Bayes' formula. Once you become familiar with the tree approach that incorporates Bayes' law, you may wish to use a tabular format to summarise the given prior and conditional probabilities, identify the probabilities to be calculated, and record the calculated joint probabilities and posterior probabilities. **Table 6.7** displays such a tabular format for the following example. For convenience, the notation has been changed slightly, using A_1 instead of A , and A_2 instead of \bar{A} . Notice that $P(B)$ is found easily by summing the joint probabilities, and that the posterior probabilities are obtained by dividing each of the joint probabilities by $P(B)$.

TABLE 6.7 Tabular alternative to Figure 6.16

Events	Prior $P(A_i)$	Conditional $P(B A_i)$	Joint $P(A_i \cap B)$	Posterior $P(A_i B)$
A_1	0.9	0.01	0.009	0.47
A_2	0.1	0.10	0.010	0.53
			$P(B) = 0.019$	

EXAMPLE 6.11

L07

How bitter would the morning coffee be?

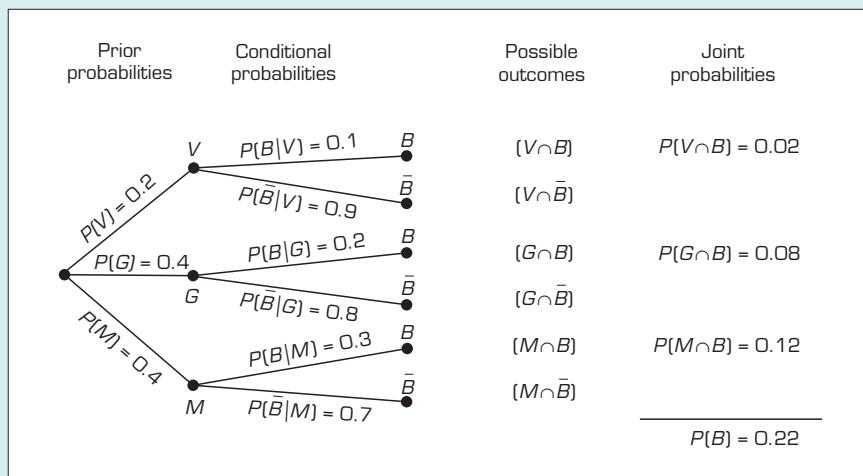
Each morning, coffee is prepared for the entire office staff by one of three employees, depending on who is first to arrive at work. Veronica arrives first 20% of the time; Gita and Michael are each the first to arrive on half of the remaining mornings. The probability that the coffee will be bitter when it is prepared by Veronica is 0.1, while the corresponding probabilities when it is prepared by Gita and by Michael are 0.2 and 0.3 respectively. If you arrive one morning and find the coffee bitter, what is the probability that it was prepared by Veronica? What is the probability that the coffee will not be bitter on any given morning?

Solution

Let B denote the event ‘the coffee is bitter’. Let V represent the event ‘the coffee is prepared by Veronica’, and let G and M represent the corresponding events when Gita and Michael respectively make the coffee. Since $P(V) = 0.2$, we know that $P(G \cup M) = 0.8$, so $P(G) = P(M) = 0.4$. The given prior and conditional probabilities are shown on the probability tree in [Figure 6.17](#). The probability tree allows us to quickly determine the outcomes belonging to the event B , namely, $B = \{V \cap B, G \cap B, M \cap B\}$. Hence:

$$P(B) = P(V \cap B) + P(G \cap B) + P(M \cap B) = 0.02 + 0.08 + 0.12 = 0.22$$

FIGURE 6.17 Probability tree for coffee preparation



Therefore, the probability that the coffee was prepared by Veronica, given that it is bitter, is

$$P(V|B) = \frac{P(V \cap B)}{P(B)}$$

$$= \frac{0.02}{0.22} = 0.09$$

Finally, the probability that the coffee will not be bitter on any given morning is

$$P(\bar{B}) = 1 - P(B) = 1 - 0.22 = 0.78$$

Now we will present the solution for the opening example to this chapter.

SPOTLIGHT ON STATISTICS

Auditing tax returns: Solution

We need to revise the prior probability that this return contains significant fraud.

The tree shown in **Figure 6.18** details the calculation.

F = Tax return is fraudulent.

\bar{F} = Tax return is honest.

E_0 = Tax return contains no suspicious deductions.

E_1 = Tax return contains one suspicious deduction.

E_2 = Tax return contains two suspicious deductions.

From auditor's determination,

$$P(F) = 0.05; P(\bar{F}) = 0.95$$

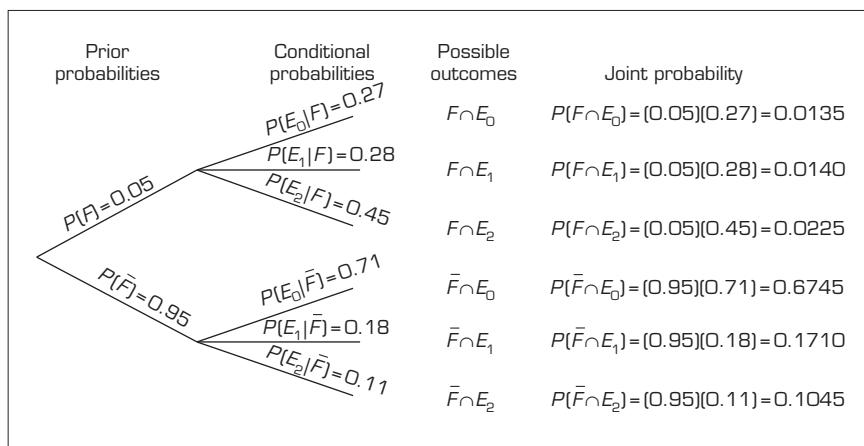
$$P(E_0|F) = 0.27; P(E_1|F) = 0.28; P(E_2|F) = 0.45;$$

$$P(E_0|\bar{F}) = 0.71; P(E_1|\bar{F}) = 0.18; P(E_2|\bar{F}) = 0.11.$$



Source: iStock.com/Fabertink

FIGURE 6.18 Probability tree for auditing tax returns example



$$\begin{aligned} P(E_1) &= P(F \cap E_1) + P(\bar{F} \cap E_1) \\ &= 0.0140 + 0.1710 = 0.1850 \end{aligned}$$

$$\begin{aligned} P(F|E_1) &= \frac{P(F \cap E_1)}{P(E_1)} \\ &= \frac{0.014}{0.185} = 0.0757 \end{aligned}$$

The probability that this return is fraudulent is 0.0757.

REAL-LIFE APPLICATIONS

Applications in medical screening and medical insurance

Medical practitioners routinely perform medical tests, called screenings, on their patients. Screening tests are conducted for all patients in a particular age and gender group, regardless of their symptoms. For example, men in their 50s are advised to take a PSA test to determine whether there is evidence of

prostate cancer. Women undergo a Pap smear for cervical cancer. Unfortunately, few of these tests are 100% accurate. Most can produce false-positive and false-negative results. A false-positive result is one in which the patient does not have the disease, but the test shows a positive result. A false-negative result is

► one in which the patient does have the disease, but the test produces a negative result.

The consequences of each test are serious and costly. A false-negative test results in a patient with a disease not detected and leads to postponing of treatment, perhaps indefinitely. A false-positive test leads to unnecessary apprehension and fear for the patient. In most such cases the patient is required to undergo further testing, such as a biopsy, and this unnecessary follow-up procedure can pose medical risks.

False-positive test results also have financial repercussions. That is, the cost of the follow-up procedure is usually far more expensive than the screening test itself. Medical insurance companies, as well as government-funded health insurance plans, are all adversely affected by false-positive test results. A correct analysis can, therefore, save both lives and money.

Bayes' law is the method we use to determine the true probabilities associated with screening

tests. Applying the complement rule to the rates of false-positive and false-negative results produces the conditional probabilities that represent correct conclusions. Prior probabilities are usually derived by looking at the overall proportion of people with the disease. In some cases, the prior probabilities may themselves have been revised because of heredity or demographic variables such as age or race. Bayes' law allows us to revise the prior probability after the test result is positive or negative.

The following example is based on actual rates of false-positive and false-negative results. Note, however, that different sources provide somewhat different probabilities. The differences may be due to the way positive and negative results are defined, or to the way technicians conduct the tests.

Note: Readers who are affected by the diseases described in the example and exercises should seek clarification from their doctor.

EXAMPLE 6.12

L07

Probability of prostate cancer

Prostate cancer is the most common form of cancer found in men. The probability of a man developing prostate cancer over his lifetime is 16%. (This figure may be higher since many prostate cancers go undetected.) Many physicians routinely perform a PSA test, particularly for men over age 50. Prostate specific antigen (PSA) is a protein produced only by the prostate gland and is fairly easy to detect. Normally, men have PSA levels between 0 and 4 mg/mL. Readings above 4 may be considered high and potentially indicative of cancer. However, PSA levels tend to rise with age, even among men who are cancer-free. Studies have shown that the test is not very accurate. In fact, the probability of having an elevated PSA level given that the man does not have cancer (false positive) is 0.135. If the man does have cancer, the probability of a normal PSA level (false negative) is almost 0.300. (This figure may vary by age and by the definition of high PSA level.) If a physician concludes that the PSA is high, a biopsy is performed. Besides the concerns and health needs of the patient, there are also financial costs. The cost of the blood test is low (approximately \$50). However, the cost of the biopsy is considerably higher (approximately \$1000). A false-positive PSA test will lead to an unnecessary biopsy. Because the PSA test is so inaccurate, some private and public medical plans do not pay for it.

Suppose you are a manager in a medical insurance company and must decide on guidelines for who should routinely be screened for prostate cancer. An analysis of prostate cancer incidence and age produces the following table of probabilities. (Note that the probability of a man under 40 developing prostate cancer is less than 0.0001, small enough to treat as 0.)

Age	Probability of developing prostate cancer
40 up to but not including 50	0.010
50 up to but not including 60	0.022
60 up to but not including 70	0.046
70 or over	0.079

▶ Assume that a man in each of the age categories undergoes a PSA test with a positive result. Calculate the probability that each man actually has prostate cancer and the probability that he does not. Perform a cost-benefit analysis to determine the cost per cancer detected.

Solution

As we did in the chapter-opening example, we'll draw a probability tree (Figure 6.19). The notation is

C = Has prostate cancer

PT = Positive test result

\bar{C} = Does not have prostate cancer

NT = Negative test result

Starting with a man between 40 and 50 years old, we have the following probabilities.

Prior

$$P(C) = 0.010$$

$$P(\bar{C}) = 1 - 0.010 = 0.990$$

Conditional (or likelihood) probabilities

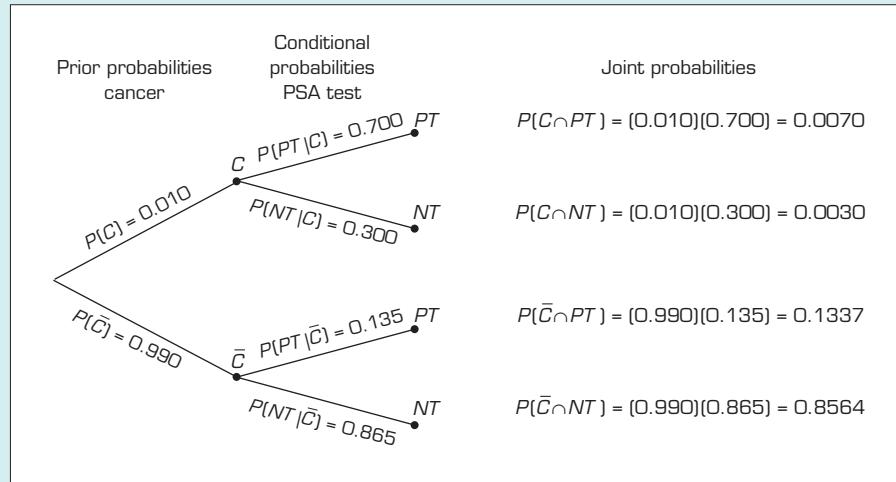
False negative: $P(NT|C) = 0.300$

True positive: $P(PT|C) = 1 - 0.300 = 0.700$

False positive: $P(PT|\bar{C}) = 0.135$

True negative: $P(NT|\bar{C}) = 1 - 0.135 = 0.865$

FIGURE 6.19 Probability tree for prostate cancer



The tree allows you to determine the probability of obtaining a positive test result.

$$P(PT) = P(C \cap PT) + P(\bar{C} \cap PT) = 0.0070 + 0.1337 = 0.1407$$

We can now calculate the probability that the man has prostate cancer given a positive test result:

$$P(C|PT) = \frac{P(C \cap PT)}{P(PT)} = \frac{0.0070}{0.1407} = 0.0498$$

The probability that he does not have prostate cancer given a positive test result can then be calculated:

$$P(\bar{C}|PT) = 1 - P(C|PT) = 1 - 0.0498 = 0.9502$$

We can repeat the process for the other age categories. The results are listed in the table.



Age	Probabilities given a positive PSA test	
	Has prostate cancer	Does not have prostate cancer
40–49	0.0498	0.9502
50–59	0.1045	0.8955
60–69	0.2000	0.8000
70 or over	0.3078	0.6922

The following table lists the proportion of each age category for which the PSA test is positive [$P(PT)$].

Age	Proportion of tests that are positive	Number of biopsies performed per million	Number of cancers detected	Number of biopsies per cancer detected
40–50	0.1407	140700	$0.0498(140700) = 7007$	20.10
50–60	0.1474	147400	$0.1045(147400) = 15403$	9.57
60–70	0.1610	161000	$0.2000(161000) = 32200$	5.00
70 or over	0.1796	179600	$0.3078(179600) = 55281$	3.25

If we assume a cost of \$1000 per biopsy, the cost per cancer detected is \$20 100 for 40–50, \$9570 for 50–60, \$5000 for 60–70, and \$3250 for 70 or over.

EXERCISES

Learning the techniques

- 6.71** Let $P(A) = 0.3$, $P(\bar{B}|A) = 0.1$ and $P(B|\bar{A}) = 0.2$.

Use a probability tree to find $P(A|B)$ and $P(A|\bar{B})$.

- 6.72** Let $P(A) = 0.6$, $P(B|A) = 0.2$ and $P(B|\bar{A}) = 0.3$. Use a probability tree to find $P(A|B)$ and $P(A|\bar{B})$.

- 6.73** Let $P(\bar{A}) = 0.4$, $P(\bar{B}|A) = 0.5$ and $P(\bar{B}|\bar{A}) = 0.6$. Use a probability tree to find $P(\bar{A}|B)$ and $P(\bar{A}|\bar{B})$.

- 6.74** Let $P(A) = 0.2$, $P(B|A) = 0.4$ and $P(\bar{B}|\bar{A}) = 0.3$. Use a probability tree to find $P(A|B)$ and $P(\bar{A}|\bar{B})$.

Applying the techniques

- 6.75 Self-correcting exercise.** Due to turnover and absenteeism at an assembly plant, 20% of the items are assembled by inexperienced employees. Management has determined that customers return 12% of the items assembled by inexperienced employees, whereas only 3% of the

items assembled by experienced employees are returned. Given that an item has been returned, what is the probability that it was assembled by an inexperienced employee?

- 6.76** A consumer goods company recruits several graduating students from universities each year. Concerned about the high cost of training new employees, the company instituted a review of attrition among new recruits. Over five years, 30% of new recruits came from a local university, and the balance came from more distant universities. Of the new recruits, 20% of those who were students from a local university resigned within two years, while 45% of other students resigned. Given that a student resigned within two years, what is the probability that they were hired from:

- a a local university?
- b a more distant university?

6.77 The finance director of a hardware wholesaler has asked the accountant to ring each customer five days before their account payment is due, as a means of reducing the number of late payments. As a result of time constraints, however, only 60% of customers receive such a call from the accountant. Of the customers called, 90% pay on time, while only 50% of those not called pay on time. The company has just received a payment on time from a customer. What is the probability that the accountant called this customer?

- 6.78** Bad gums may be evidence of a bad heart. Researchers have discovered that 85% of people who have suffered a heart attack had periodontal disease, an inflammation of the gums. Only 29% of healthy people have this disease.
- a Suppose that in a certain community, heart attacks are quite rare, occurring with only 10% probability. If someone has periodontal disease, what is the probability that he or she will have a heart attack?
 - b If 40% of the people in a community will have a heart attack, what is the probability that a person

with periodontal disease in this community will have a heart attack?

6.79 Your favourite team is in the grand final. You have assigned a probability of 60% that they will win the championship. Past records indicate that when teams win the championship, they have won the first game of the series 70% of the time. When they lose the championship, they have won the first game 25% of the time. Your team lost the first game of the series, what is the probability that they will win the championship?

6.80 The Pap smear is the standard test for cervical cancer. The false-positive rate is 0.636; the false-negative rate is 0.180. Family history and age are factors that must be considered when assigning the probability of cervical cancer. Suppose that after obtaining a medical history, a physician determines that the proportion of women who have cervical cancer is 2%. Determine the effects a positive and a negative Pap smear test have on the probability that the patient has cervical cancer.

6.6 Identifying the correct method

As we have previously pointed out, the emphasis in this book is on identifying the correct statistical technique to use in a given situation. In Chapters 2 and 4 we showed how to summarise data by first identifying the appropriate method to use. Although it is difficult to offer strict rules on which probability method to use, we can provide some general guidelines.

In the examples and exercises in this text (and most other introductory statistics books), the key issue is whether joint probabilities are provided or are required.

6.6a When joint probabilities are given

In Section 6.2 we addressed problems in which the joint probabilities were given. In these problems, we can calculate marginal probabilities by adding across rows and down columns. We can use the joint and marginal probabilities to calculate conditional probabilities, for which a formula is available. This allows us to determine whether the events described by the table are independent or dependent. We can also apply the addition rule to calculate the probability that either of two events occurs.

6.6b When joint probabilities are required

The previous sections introduced three probability rules and probability trees. We need to apply some or all of these rules in circumstances where one or more joint probabilities are required. We apply the multiplication rule (either by formula or using a probability tree) to calculate the probabilities of intersections. In some problems we're interested in adding these joint probabilities. We're actually applying the addition rule for mutually exclusive events here. We also frequently use the complement rule. In addition, we can calculate new conditional probabilities using Bayes' law.

Study Tools

CHAPTER SUMMARY

Gamblers, business people and scientists frequently find themselves in decision-making situations involving uncertain events. Probability is the basic tool they use to make rational judgements in such situations. The first step in assigning probabilities to uncertain events is to form a *sample space* – a listing of all possible outcomes that can result from a *random experiment*. A *probability* (number between zero and one) is then assigned to each outcome, measuring the likelihood of occurrence of that outcome.

The use of a *probability tree* often facilitates both the formation of a sample space and the assignment of probabilities to its outcomes. Probabilities may then be calculated for more complex events in accordance with rules of probability dealing with the *complement*, *union* and *intersection* of events. The notion of *conditional probability* allows us to express the probability that a particular event will occur when some partial knowledge of the experimental outcome is available. In particular, *Bayes' law* is used to revise the probability that an event will occur, in the light of newly acquired information.

The data files for Examples, Exercises and Cases are provided on the companion website (accessible through <https://login.cengagebrain.com/>).

SUMMARY OF FORMULAS

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(\bar{A}) = 1 - P(A)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$\begin{aligned} P(A \cap B) &= P(A) \cdot P(B|A) \\ &= P(B) \cdot P(A|B) \end{aligned}$$

$$P(A \cup B) = P(A) + P(B) \quad \text{if } A \text{ and } B \text{ are mutually exclusive.}$$

$$P(A \cap B) = P(A) \cdot P(B) \quad \text{if } A \text{ and } B \text{ are independent.}$$

If A and B are independent, then $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

SUPPLEMENTARY EXERCISES

- 6.81** The results of three flips of a fair coin are observed.

Consider the following events:

- A:** At least two heads are observed.
 - B:** Exactly one tail is observed.
 - C:** Exactly two tails are observed.
- a** Define an appropriate sample space S for this experiment.
- b** Assign probabilities to the outcomes in S .
- c** Find $P(A)$ by summing the probabilities of the outcomes in A .

- d** Find $P(\bar{B})$.
- e** Find $P(B \cap C)$.
- f** Find $P(A \cap B)$.
- g** Find $P(B|A)$.
- h** Is there a pair of independent events among A , B and C ?
- i** Is there a pair of mutually exclusive events among A , B and C ?

6.82 Two six-sided dice are rolled, and the number of spots turning up on each is observed. Find the probability of observing:

- a three spots on one die and five spots on the other
- b exactly one die that has two spots showing
- c that the sum of the spots showing is seven
- d that the sum of the spots showing is eight
- e that the sum of spots showing is an even number
- f that the sum of the spots showing is eight, given that the sum is an even number.

6.83 Referring to Exercise 6.82, define the following events:

- a A: The sum of the spots showing is two, three or 12.
- b B: The sum of the spots showing is an even number.
- c Are A and B independent events? Explain.

6.84 Exactly 100 employees of a Brisbane firm have each purchased one ticket in a lottery, with the draw to be held at the firm's annual party. Of the 80 men who purchased tickets, 25 are single. Only four of the women who purchased tickets are single.

- a Find the probability that the lottery winner is married.
- b Find the probability that the lottery winner is a married woman.
- c If the winner is a man, what is the probability that he is married?

6.85 A customer-service supervisor regularly conducts a survey of customer satisfaction. The results of the latest survey indicate that 8% of customers were not satisfied with the service they received at their last visit to the store. Of those who are not satisfied, only 22% return to the store within a year. Of those who are satisfied, 64% return within a year. A customer has just entered the store. In response to your question, he informs you that it is less than one year since his last visit to the store. What is the probability that he was satisfied with the service he received?

6.86 How does level of affluence affect health care? To address one dimension of the problem, a group of people who have had a heart attack was selected. Each person was categorised as a low-, medium- or high-income earner. Each was also categorised as

having survived or died. A demographer notes that in the society being considered, 21% fall into the low-income group, 49% are in the medium-income group and 30% are in the high-income group. Furthermore, an analysis of people who have had a heart attack reveals that 12% of low-income people, 9% of medium-income people, and 7% of high-income people die of heart attacks. Find the probability that a survivor of a heart attack is in the low-income group.

6.87 As input into pricing policy, the owner of an appliance store is interested in the relationship between the price at which an item is sold (retail or sale price) and the customer's decision on whether or not to purchase an extended warranty. The owner has constructed the accompanying table of probabilities, based on a study of 2000 sales invoices. Suppose that one sales invoice is selected at random, with the relevant events being defined as follows:

- A: The item is purchased at retail price.
- B: The item is purchased at sale price.
- C: An extended warranty is purchased.
- D: An extended warranty is not purchased.

Price	Extended warranty	
	Purchased (C)	Not purchased (D)
Retail price (A)	0.21	0.57
Sale price (B)	0.14	0.08

Express each of the following probabilities in words, and find its value:

- a $P(A)$
- b $P(\bar{A})$
- c $P(C)$
- d $P(C|A)$
- e $P(C|B)$
- f $P(D|B)$
- g $P(B|D)$
- h $P(C|D)$
- i $P(A \cup B)$
- j $P(A \cap D)$

6.88 A survey of students studying for an MBA at a particular university determined that 96% expect a starting salary in excess of \$35 000, while 92% expect a starting salary of under \$50 000. If one of these students is selected at random, what is the

probability that the student expects a starting salary of between \$35 000 and \$50 000?

- 6.89** The director of an insurance company's computing centre estimates that the company's mainframe computer has a 20% chance of 'catching' a computer virus. However, she feels that there is only a 6% chance of the computer catching a virus that will completely disable its operating system. If the company's computer should catch a virus, what is the probability that the operating system will be completely disabled?

- 6.90** It is known that 3% of the tickets in a certain scratch-and-win game are winners, in the sense that the purchaser of such a ticket will receive a prize. If three tickets are purchased at random, what is the probability of each of the following?
- All three tickets are winners.
 - Exactly one of the tickets is a winner.
 - At least one of the tickets is a winner.

- 6.91** A financial analyst estimates that a certain share has a 60% chance of rising in value by more than 15% over the coming year. She also predicts that the stock market in general, as measured by the All Ordinaries Index, has only a 20% chance of rising more than 15%. But if the Index does so, she feels that the share has a 95% chance of rising by more than 15%.
- Find the probability that both the All Ordinaries Index and the share will rise in value by more than 15%.
 - Find the probability that either the Index or the share, or both, will rise by more than 15%.

- 6.92** Many auditors have developed models based on financial ratios that predict whether or not a company

will go bankrupt within the next year. In a test of one such model, the model correctly predicted the bankruptcy of 85% of firms that in fact did fail, and it correctly predicted non-bankruptcy for 82% of firms that did not fail. Suppose that the model maintains the same reliability when applied to a new group of 100 firms, of which four fail in the year after the time at which the model makes its predictions.

- Determine the number of firms for which the model's prediction will prove to be correct.
- Find the probability that one of these firms will go bankrupt, given that the model has predicted it will do so.

- 6.93** Casino Windsor conducts surveys to determine the opinions of its customers. Among other questions, respondents are asked the question 'What is your overall impression of Casino Windsor?'. The possible responses are: Excellent, Good, Average and Poor. Additionally, the gender of the respondent is also noted. After analysing the results, the following table of joint probabilities was produced.

Rating	Women	Men
Excellent	0.27	0.22
Good	0.14	0.10
Average	0.06	0.12
Poor	0.03	0.06

- What proportion of customers rate Casino Windsor as excellent?
- Determine the probability that a male customer rates Casino Windsor as excellent.
- Find the probability that a customer who rates Casino Windsor as excellent is a man.
- Are gender and rating independent? Explain your answer.

Case Studies

CASE 6.1 Let's make a deal

A number of years ago there was a popular television game show on Channel 10 called 'Let's Make a Deal'. The host, Vince Sorrenti, would randomly select contestants from the audience and, as the title suggests, he would make deals for prizes. Contestants would be given relatively modest prizes and would then be offered the opportunity to risk the prize to win better ones.

Suppose that you are a contestant on this show. Vince has just given you a free trip worth \$500 to a place that is of little interest to you. He now offers you a trade: give up the trip in exchange for a gamble. On the stage there are three curtains: A, B and C. Behind one of them is a brand-new car worth \$20 000. Behind the other two curtains the stage is empty. You decide to gamble and select curtain A. In an attempt to make things more interesting, Vince exposes an empty stage by opening curtain C. (He knows that there is nothing behind it.) He then offers you the free trip again if you quit the game or, if you like, you can propose another deal (i.e. you can keep your choice of curtain A or switch to curtain B). What do you do? Explain why.

CASE 6.2 University admissions in Australia: Does gender matter?

C06-02 The following table gives the details of domestic students who are commencing at Australian universities in the six states and two territories by gender in 2008 and 2018. A newspaper article about the higher-education student population quoted that 'the likelihood of gaining admission to some Australian universities is higher for male than for female students'. Using the data for 2018, do you agree with the quote? Why? Would your response change for 2008?

Number of commencing domestic students in Australian universities by state and gender, 2008 and 2018

State	Number of male students		Number of female students	
	2008	2018	2008	2018
NSW	42 232	54 002	56 382	74 029
Victoria	26 682	39 775	37 231	56 602
Queensland	21 728	28 473	32 437	44 070
WA	11 148	16 266	17 000	24 202
SA	7 484	12 270	11 394	19 450
Tasmania	2 530	3 867	4 061	9 790
ACT	941	1 071	2 364	2 956
NT	3 787	4 809	4 731	6 327
Total	116 532	160 533	165 600	237 426

Source: © Commonwealth of Australia CC BY 4.0 International <https://creativecommons.org/licenses/by/4.0/>

CASE 6.3 Maternal serum screening test for Down syndrome

Pregnant women are screened for a genetic condition called Down syndrome. Down syndrome is characterised by varying degrees of intellectual and physical impairment. Some mothers choose to abort the foetus if they are certain that their baby will be born with the syndrome. The most common type of screening is called maternity serum screening, a blood test that

looks for markers in the blood to indicate the genetic condition. The false-positive and false-negative rates of this test vary according to the age of the mother, as shown in the table below.

Mother's age	False-positive rate	False-negative rate
Under 30	0.040	0.376
30–34	0.082	0.290
35–37	0.178	0.269
38 and over	0.343	0.029

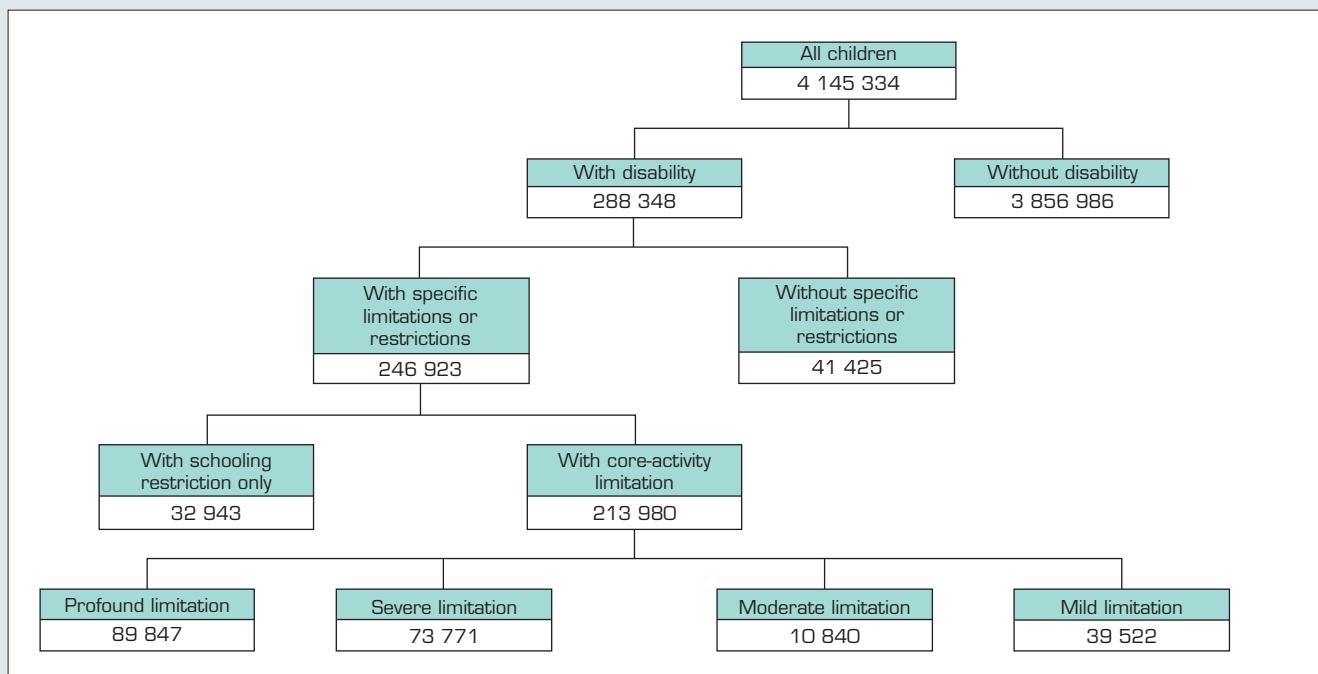
The probability that a baby will be born with Down syndrome is primarily dependent on the mother's age. The probabilities are listed below.

Mother's age	Probability of Down syndrome
25	1/1300
30	1/900
35	1/350
40	1/100
45	1/25
49	1/12

- a For each of the ages 25, 30, 35, 40, 45, and 49, determine the probability of Down syndrome if the maternity serum screening produces a positive result.
- b Repeat for a negative result.

CASE 6.4 Levels of disability among children in Australia

Among the children in Australia, 7% are assessed to have some form of disability. The level of disability can be ranked in various categories. The diagram below provides statistics on the number of children categorised by their level of disability. Use these data to determine the following probabilities.



- a Find the probability that a child selected at random will have a disability.
- b Find the probability that a child selected at random does not have any disability.
- c What proportion of children have a disability with specific limitations or restrictions?
- d Find the probability that a child with a disability who is selected at random will have specific limitations or restrictions.
- e Find the probability that a child with a disability who is selected at random will not have specific limitations or restrictions.
- f Find the probability that a child selected at random has a disability with schooling restrictions only.
- g What is the probability that a child selected at random has only a mild limitation, given that the child has core-activity limitation?

CASE 6.5 Probability that at least two people in the same room have the same birthday

Suppose that there are two people in a room. The probability that they share the same birthday (date, but not necessarily year) is $1/365$, and the probability that they have different birthdays is $364/365$. To illustrate, suppose that you're in a room with one other person and that your birthday is 1 July. The probability that the other person does not have the same birthday is $364/365$ because there are 364 days in the year that are not July 1. If a third person now enters the room, the probability that he or she has a different birthday from the first two people in the room is $363/365$. Thus, the probability that three people in a room have different birthdays is $(364/365)(363/365)$. You can continue this process for any number of people. Find the number of people in a room so that there is about a 50% probability that at least two have the same birthday.

Hint 1: Calculate the probability that they don't have the same birthday.

Hint 2: Excel users can use the product function to calculate joint probabilities.

CASE 6.6 Home ownership in Australia

For a significant proportion of Australians, owning a home in their lifetime is becoming a dream. Australians spend about one-fifth of their income on the cost of housing or renting a home. The Australian Bureau of Statistics has published the following flowchart giving information on home ownership based on data from a 2011–12 survey.

Based on the data provided:

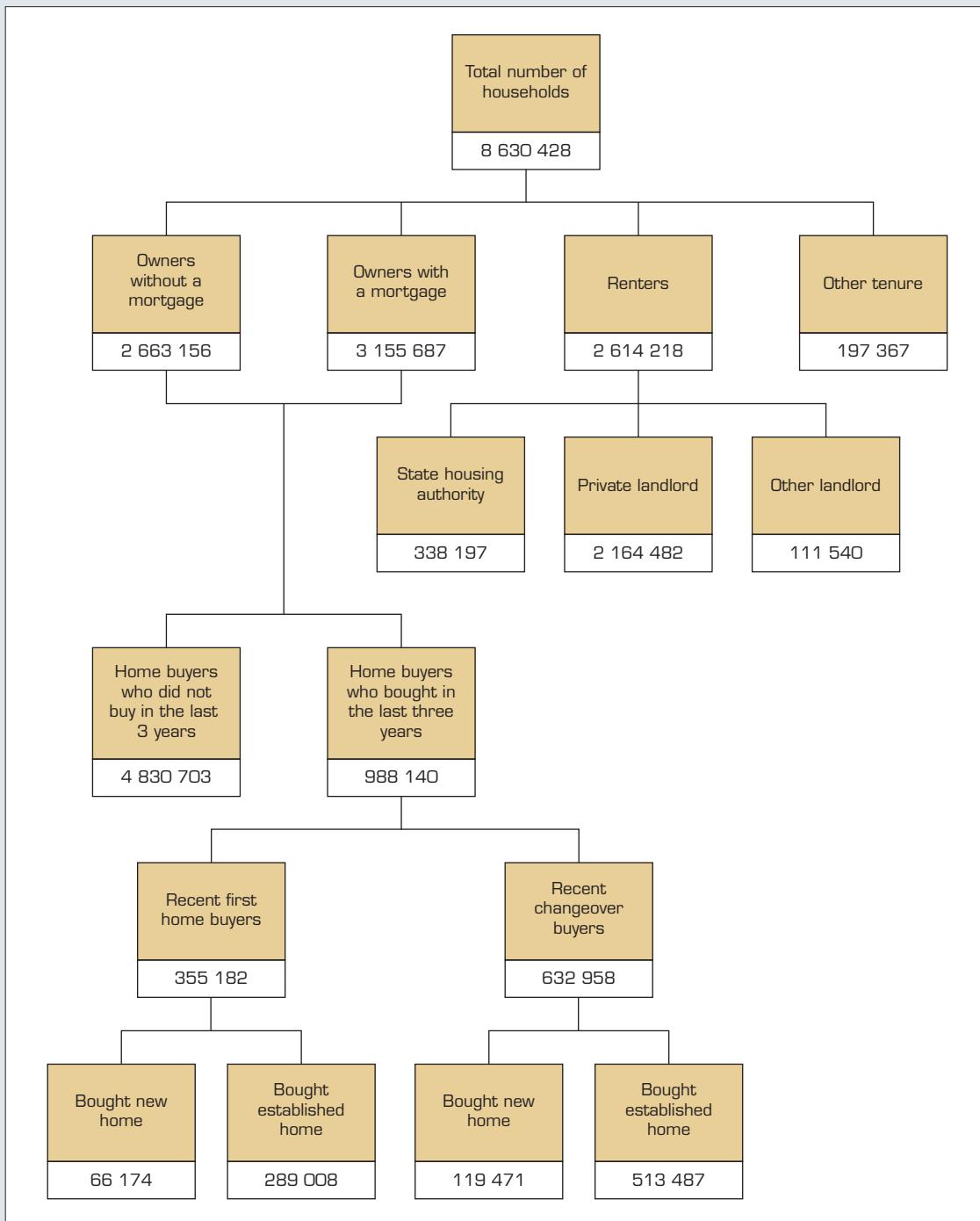
- a What proportion of households are renters?
- b What proportion of households are renters from the state housing authority?
- c Find the probability that a household selected at random will be living in a private rental property.
- d Find the probability that a household selected at random is a home owner without a mortgage.
- e What is the probability that a renter household selected at random is living in a house owned by the state housing authority?

Consider the homebuyers who bought their houses in the last three years.

- f What proportion of the recent homebuyers are changeover buyers?

- g** What proportion of the recent homebuyers are first home buyers?
h Given that the selected household bought a new home, find the probability that the household is a changeover buyer.
i Given that the selected household bought an established home, find the probability that the household is a first home buyer.

From these probabilities and the data, discuss the characteristics of the Australian homebuyers and renters.



Source: Australian Bureau of Statistics, *Housing Occupancy and Costs, Australia, 2011–12*, cat. no. 4130.0, ABS, Canberra

CASE 6.7 COVID-19 confirmed cases and deaths in Australia II

C06-07 Consider again Case 3.1, where data on the number of COVID-19 related deaths as at 30 June 2020, by states, age group and gender are presented.

- a If a person with confirmed coronavirus in Australia is selected at random, what is the probability that the person would be from NSW. Repeat for other states.
- b If a person who died due to coronavirus in Australia is selected at random, what is the probability that the person would be from NSW. Repeat for other states.

Number of COVID-19 confirmed cases and deaths by states, Australia, 30 June 2020

State	Confirmed cases	Deaths
New South Wales (NSW)	3211	48
Victoria (VIC)	2195	20
Queensland (QLD)	1079	6
Western Australia (WA)	594	9
South Australia (SA)	436	4
Tasmania (TAS)	225	13
Australian Capital Territory (ACT)	111	3
Northern Territory (NT)	30	0
Australia	7881	103

Source: covid19data.com.au. CC BY 4.0 <https://creativecommons.org/licenses/by/4.0/>.

Random variables and discrete probability distributions

Learning objectives

This chapter provides the characteristics of probability distributions and introduces binomial and other discrete probability distributions.

At the completion of this chapter, you should be able to:

- L01** explain the importance of probability distributions
- L02** discuss the concept of a random variable and its probability distribution
- L03** compute the mean and standard deviation of a discrete probability distribution, and apply the laws of expected value and variance
- L04** understand the Bernoulli trial and binomial distribution, and calculate the mean and the variance
- L05** recognise when it is appropriate to use a binomial distribution, and understand how to use the table of binomial probabilities
- L06** understand the Poisson distribution, recognise when it is appropriate to use a Poisson distribution and understand how to use the table of Poisson probabilities
- L07** explain the bivariate distribution and compute joint probabilities, covariances and correlations.

CHAPTER OUTLINE

- Introduction
- 7.1** Random variables and probability distributions
- 7.2** Expected value and variance
- 7.3** Binomial distribution
- 7.4** Poisson distribution
- 7.5** Bivariate distributions
- 7.6** Applications in finance: Portfolio diversification and asset allocation

SPOTLIGHT ON STATISTICS

Where to invest the super?

An employee received 2 million dollars from his superannuation fund upon retirement. He plans to develop a portfolio made up of four investments, namely residential rental properties, banking shares, commercial rental properties and resources shares. However, he doesn't know how much to invest in each one. He wants to maximise his return while minimising the risk. He has calculated the monthly returns for all four investments over a 6-year period from January 2014 to December 2019. These data are stored in file **CH07: XM07-00**. After some consideration, he narrowed his choices down to the following three portfolios. Which choice should he make?



Source: iStock.com/Faberrink

- 1 Equal amounts (\$500 000) in each of the four investments
- 2 Residential: \$500 000; banking: \$400 000; commercial: \$800 000; resources: \$300 000
- 3 Residential: \$200 000; banking: \$1 000 000; commercial: \$600 000; resources: \$200 000

We will provide our answer after we have developed the necessary tools in Section 7.6 (see pages 299–301).

Introduction

In this chapter, we extend the concepts and techniques of probability introduced in Chapter 6. We present random variables and probability distributions, which are essential in the development of statistical inference.

Here is a brief glimpse into the wonderful world of statistical inference. Suppose that you flip a coin 100 times and count the number of heads. The objective is to determine whether we can infer from the count that the coin is not balanced. It is reasonable to believe that observing a large number of heads (say, 90) or a small number (say, 15) would be a statistical indication of an unbalanced coin. However, where do we draw the line? At 75 or 65 or 55? Without knowing the probability of the frequency of the number of heads from a balanced coin, we cannot draw any conclusions from the sample of 100 coin flips.

The concepts and techniques of probability introduced in this chapter will allow us to calculate the probability we seek. As a first step, we introduce random variables and probability distributions.

7.1 Random variables and probability distributions

Consider the following examples.

- 1 Consider an experiment in which we flip two balanced coins and observe the results. We can represent the events as follows:
 - Heads on the first coin and heads on the second coin
 - Heads on the first coin and tails on the second coin
 - Tails on the first coin and heads on the second coin
 - Tails on the first coin and tails on the second coin

However, we can list the events in a different way. Instead of defining the events by describing the outcome of each coin, we can count the number of heads (or, if we wish, the number of tails). Thus, the events are now:

- 2 heads
- 1 head
- 1 head
- 0 heads

The number of heads in the experiment is called a random variable. We often label the random variable X , and we're interested in the probability of each value of X . Thus, in this illustration, the values that the random variable X can take are 0, 1 and 2.

- 2 In many parlour games as well as in the game of craps played in casinos, the player tosses two dice. One way of listing the events is to describe the number on the first die and the number on the second die as follows:

1, 1	1, 2	1, 3	1, 4	1, 5	1, 6
2, 1	2, 2	2, 3	2, 4	2, 5	2, 6
3, 1	3, 2	3, 3	3, 4	3, 5	3, 6
4, 1	4, 2	4, 3	4, 4	4, 5	4, 6
5, 1	5, 2	5, 3	5, 4	5, 5	5, 6
6, 1	6, 2	6, 3	6, 4	6, 5	6, 6

However, in almost all games the player is primarily interested in the total. Accordingly, we can list the totals of the two dice instead of the individual numbers:

2	3	4	5	6	7
3	4	5	6	7	8
4	5	6	7	8	9
5	6	7	8	9	10
6	7	8	9	10	11
7	8	9	10	11	12

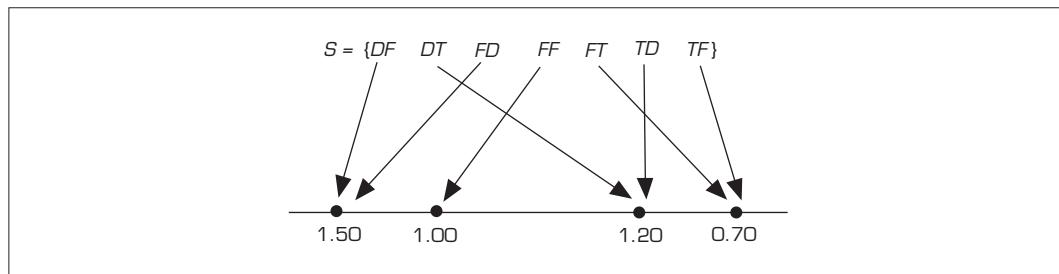
If we define the random variable X as the total of the two dice, X can take values 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12.

- 3 Consider once again the experiment described in Example 6.10, involving the selection of two coins from a box containing a one-dollar coin (D), two 50-cent coins (F) and a 20-cent coin (T). Recall that the sample space for this experiment was $S = \{DF, DT, FD, FF, FT, TD, TF\}$. Our primary interest in this experiment might be in the total sum of money selected. If X denotes the total sum of money observed, the possible values that the **random variable** X can take are \$0.70, \$1.00, \$1.20 and \$1.50, as shown in **Figure 7.1**.

random variable

A variable whose values are determined by outcomes of a random experiment.

FIGURE 7.1 Random variable X assigning values to simple events



Random variable

A random variable is a function or rule that assigns a numerical value to each outcome of an experiment.

In some experiments, the outcomes are numbers. For example, when we observe the return on an investment or measure the amount of time to assemble a computer, the experiment produces events that are numbers. Simply stated, the value of a random variable is a numerical event.

7.1a Discrete and continuous random variables

There are two types of random variables – discrete and continuous – distinguished from one another by the number of possible values that they can assume. A random variable is *discrete* if it can assume only a countable number of possible values. A random variable that can assume an uncountable number of values is *continuous*.

discrete random variable

A random variable that can assume only a countable number of values (finite or infinite).

A **discrete random variable** has a finite or countably infinite number of possible values. In most practical situations, a discrete random variable counts the number of times a particular attribute is observed. Examples of discrete random variables include the number of defective items in a production batch, the number of telephone calls received in a given hour, and the number of customers served at a hotel reception desk on a given day. If X denotes the number of customers served at a hotel reception desk on a particular day, then X can take any one of the values $x = 0, 1, 2, \dots$ (countably infinite). If X denotes the number of students in a statistics class of 200 who pass the subject, then X can take one of the values $x = 0, 1, 2, \dots, 200$ (finite). The random variables we considered in the examples 1, 2 and 3 above are discrete random variables.

A **continuous random variable** has an uncountably infinite number of possible values; that is, it can take on any value in one or more intervals of values. Continuous random variables typically record the value of a measurement such as time, weight or length. For example, let X = the time taken by a student to complete a statistics exam at a university, where the time limit is three hours and students cannot leave before 30 minutes. The smallest value of X is 30 minutes. If we attempt to count the number of values that X can take, we need to identify the next value. Is it 30.1 minutes? 30.01 minutes? 30.001 minutes? None of these is the second possible value of X because numbers exist that are larger than 30 and smaller than 30.001. It becomes clear that we cannot identify the second, or third, or any other values of X (except the largest value, 180 minutes). Random variable X can take any one of infinitely many possible values: $30 \leq x \leq 180$. Thus, we cannot count the number of values X can take and therefore X is continuous.

Having considered the values of a random variable, we now turn to the probabilities with which those values are assumed. Once we know the possible values of a random variable and the probabilities with which those values are assumed, we have the probability distribution of the random variable – our main object of interest.

A **probability distribution** is a table, formula, or graph that describes the values of a random variable and the probability associated with these values. We will address discrete probability distributions in the rest of this chapter and cover continuous probability distributions in Chapter 8.

As we noted earlier, an uppercase letter will represent the *name* of the random variable, usually X . Its lowercase counterpart will represent the *value* of the random variable, x .

7.1b Discrete probability distributions

The probability associated with a particular value of a discrete random variable is determined in a manner you can probably anticipate. If x is a value of a random variable X , then the probability that X assumes the value x , denoted either by

$$P(X = x) \text{ or simply } p(x)$$

is the sum of the probabilities associated with the simple events for which X assumes the value x .

The probabilities of the values of a discrete random variable may be derived by means of probability tools, such as probability trees, or by applying one of the definitions of probability. However, two fundamental requirements apply, as stated in the box.

Requirements of discrete probability distribution

If a random variable X can take values x_1, x_2, \dots, x_k , and $p(x_i)$ is the probability that the random variable is equal to x_i , then:

i $0 \leq p(x_i) \leq 1$ for $i = 1, 2, \dots, k$

ii $\sum_{i=1}^k p(x_i) = 1$

Let us apply this rule to the experiment involving the selection of two coins from a box of four coins (Example 6.10). If the random variable X represents the total sum of money selected, X can assume any one of the values \$0.70, \$1.00, \$1.20 or \$1.50. Probabilities can be assigned to the values of X with the help of **Table 7.1**, which records each simple event, the corresponding value of the random variable and the probability of each simple event. (Recall that the simple event probabilities shown in **Table 7.1** were calculated in Example 6.10, with the help of a probability tree.)

continuous random variable

A random variable that can assume an uncountable number of values in an interval.

probability distribution

A table or graph that assigns probability values to all possible values or range of values that a random variable can assume.

TABLE 7.1 Values of X corresponding to simple events

Simple event	X	$p(x)$
DF	\$1.50	2/12
DT	\$1.20	1/12
FD	\$1.50	2/12
FF	\$1.00	2/12
FT	\$0.70	2/12
TD	\$1.20	1/12
TF	\$0.70	2/12

For example, X takes the value \$0.70 if either the simple event FT or the simple event TF occurs, so:

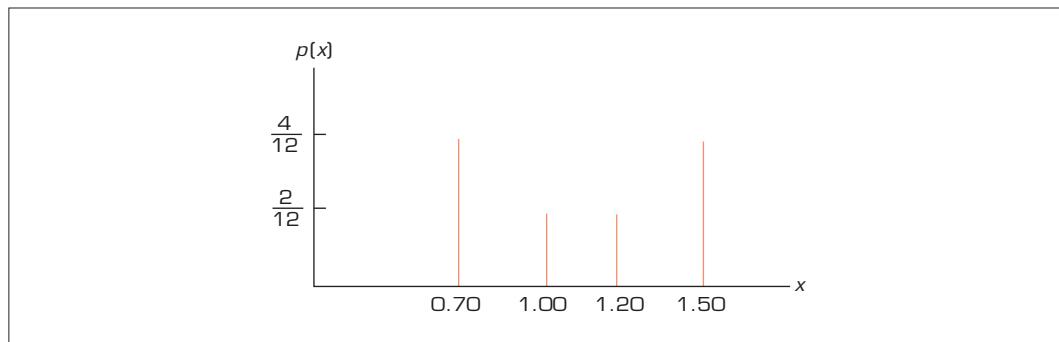
$$\begin{aligned} P(X = \$0.70) &= P(FT) + P(TF) \\ &= 2/12 + 2/12 = 4/12 \end{aligned}$$

The distinct values of X and their associated probabilities are summarised in **Table 7.2**, which gives the probability distribution of X . The probability distribution of X can be presented in the tabular form shown in **Table 7.2**, in the graphical form of **Figure 7.2**, or in terms of the following formula:

$$p(x) = \begin{cases} 4/12 & \text{if } x = \$0.70 \\ 2/12 & \text{if } x = \$1.00 \\ 2/12 & \text{if } x = \$1.20 \\ 4/12 & \text{if } x = \$1.50 \end{cases}$$

TABLE 7.2 Probability distribution of X

x	\$0.70	\$1.00	\$1.20	\$1.50
$p(x)$	4/12	2/12	2/12	4/12

FIGURE 7.2 Graphical presentation of probability distribution

In an example such as this one, where the formula is rather cumbersome, the tabular representation of the probability distribution of X is the most convenient. Whichever representation is used, a discrete probability distribution must satisfy the two conditions, which follow from the basic requirements for probabilities outlined in Chapter 6.

Once a probability distribution has been defined for a random variable X , we may talk about the probability that X takes a value in some range of values. The probability that X takes a value between a and b (both inclusive), denoted by $P(a \leq X \leq b)$, is obtained by summing the probabilities $p(x)$ for all values of x such that $(a \leq x \leq b)$.

In the preceding example, we would have:

$$\begin{aligned} P(0.70 \leq X \leq 1.20) &= p(0.70) + p(1.00) + p(1.20) \\ &= \frac{4}{12} + \frac{2}{12} + \frac{2}{12} = \frac{8}{12} = 0.67 \end{aligned}$$

In other words, the probability that the total sum of money selected is one of 70 cents or \$1.00 or \$1.20 is 0.67.

EXAMPLE 7.1

LO2

Probability distribution of the household size in Australia

The following table gives the distribution of the Australian population in 2016 according to household size.

Number of households by household size in Australia, 2016

Household size	Number of households ('000)
1	2024
2	2768
3	1338
4	1314
5	557
6 or more	285
Total	8286

Source: Australian Bureau of Statistics,
Census of Population and Housing 2016.

Develop the probability distribution of the random variable defined as the household size in Australia. Calculate the probability that a household selected at random will have a household size of 3. Also obtain the probability of selecting a household with size 4 or more.

Solution

The probability of each value of X , the household size, is computed as the relative frequency. We divide the frequency for each value of X by the total number of households, producing the following probability distribution.

Probability distribution of the household size in Australia

Household size (x)	$p(x)$
1	$2024/8286 = 0.244$
2	$2768/8286 = 0.334$
3	$1338/8286 = 0.161$
4	$1314/8286 = 0.159$
5	$557/8286 = 0.067$
6 or more	$285/8286 = 0.034$
Total	1.000

As you can see, the requirements are satisfied. Each probability lies between 0 and 1 and the total is 1. We interpret the probabilities in the same way we did in Chapter 6.

If we select one household at random, the probability that the household size is 3 is

$$P(X = 3) = p(3) = 0.161$$

We can also apply the addition rule for mutually exclusive events. (The values of X are mutually exclusive: a household size can be 1, 2, 3, 4, 5 or 6 or more.) The probability that a randomly selected household size is 4 or greater is

$$\begin{aligned} P(X \geq 4) &= P(X = 4) + P(X = 5) + P(X \geq 6) \\ &= 0.159 + 0.067 + 0.034 \\ &= 0.260 \end{aligned}$$

In Example 7.1, we calculated the probabilities using census information about the entire population. The next example illustrates the use of the techniques introduced in Chapter 6 to develop a probability distribution.

EXAMPLE 7.2

LO3

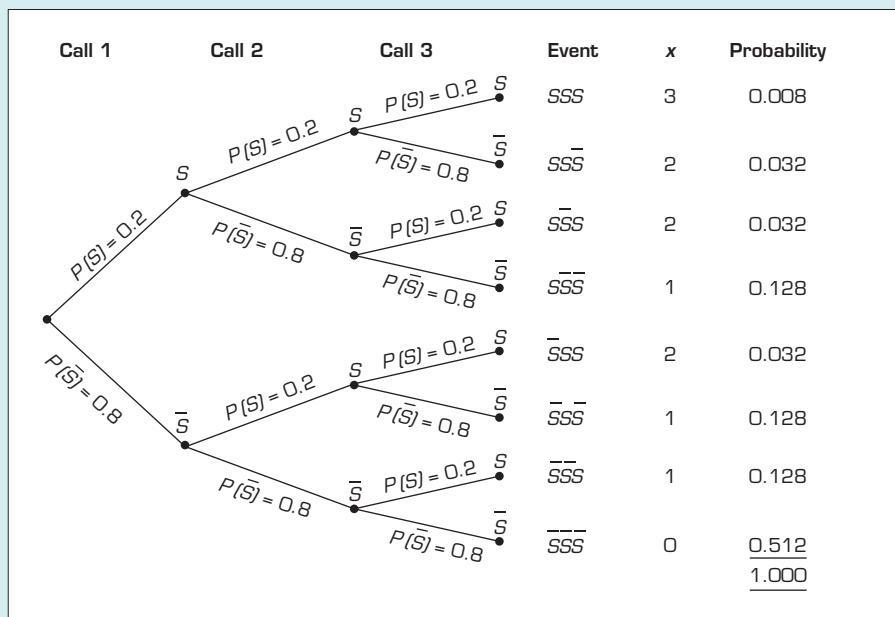
Probability distribution of the number of mutual fund sales

A mutual fund salesperson has arranged to call on three people tomorrow. Based on past experience, the salesperson knows that there is a 20% chance of closing a sale on each call. Determine the probability distribution of the number of sales they will make.

Solution

We can use the probability rules and trees introduced in Chapter 6. Let S denote a sale. **Figure 7.3** displays the probability tree for this example. Let X = the number of sales.

FIGURE 7.3 Probability tree for Example 7.2



The tree exhibits each of the eight possible outcomes and their probabilities. We see that there is one outcome that represents no sales and its probability is $p(0) = 0.512$. There are three outcomes representing one sale, each with probability 0.128, so we add these probabilities.

Thus,

$$p(1) = 0.128 + 0.128 + 0.128 = 3(0.128) = 0.384$$

The probability of two sales is computed similarly:

$$p(2) = 0.032 + 0.032 + 0.032 = 3(0.032) = 0.096$$

There is one outcome for which there are three sales:

$$p(3) = 0.008$$

The probability distribution of X is listed in **Table 7.3**.

TABLE 7.3 Probability distribution of the number of mutual fund sales in Example 7.2

$X = x$	0	1	2	3
$p(X = x)$	0.512	0.384	0.096	0.008

7.1c Probability distributions and populations

The importance of probability distributions derives from their use as representatives of populations. In Example 7.1 the distribution provided us with information about the population of household size. In Example 7.2 the population was the number of sales made in three calls by the mutual fund salesperson. And, as we noted before, statistical inference deals with inference about populations.

EXERCISES

Learning the techniques

- 7.1** The distance a car travels on one tank of petrol is a random variable.
- What are the possible values of this random variable?
 - Are the values countable? Explain.
 - Is there a finite number of values? Explain.
 - Is the random variable discrete or continuous? Explain.
- 7.2** The number of accidents that occur annually on a busy stretch of highway is a random variable.
- What are the possible values of this random variable?
 - Are the values countable? Explain.
 - Is there a finite number of values? Explain.
 - Is the random variable discrete or continuous? Explain.
- 7.3** The mark on a statistics exam that consists of 100 multiple-choice questions is a random variable.
- What are the possible values of this random variable?
 - Are the values countable? Explain.
 - Is there a finite number of values? Explain.
 - Is the random variable discrete or continuous? Explain.
- 7.4** Consider a random variable X with the following probability distribution:

x	-4	0	1	2
$p(x)$	0.2	0.3	0.4	0.1

Find the following probabilities:

- $P(X > 0)$
- $P(X \leq 0)$
- $P(0 \leq X \leq 1)$
- $P(X = -2)$
- $P(X = -4)$
- $P(X < 2)$

- 7.5** Determine whether each of the following are valid probability distributions and explain why.

a

x	1	2	3	4
$p(x)$	0.2	0.2	0.3	0.4

b

x	0	2	4	5
$p(x)$	0.1	0.2	0.3	0.4

c

x	-2	-1	1	2
$p(x)$	0.01	0.01	0.01	0.97

- 7.6** Let X be the number of spots turning up when a six-sided die is tossed.

- Express the probability distribution of X in tabular form.
- Express the probability distribution of X in graphical form.

- 7.7** Let X be the number of heads that are observed when a fair coin is flipped three times.

- Express the probability distribution of X in tabular form.
- Express the probability distribution of X in graphical form.

Applying the techniques

- 7.8** **Self-correcting exercise.** The number of persons in different living arrangements in Australia in 2016 are as follows.

Category (x)	Number of persons ('000)
Couple with children (1)	11 886
Couple without children (2)	5 119
One parent male (3)	194
One parent female (4)	878
Other (5)	6 114
Total	24 191

Source: Australian Bureau of Statistics, *Household and Family Projections, Australia, 2016–2026*, 2016, cat. no. 3236.0, ABS, Canberra

- a Construct the probability distribution of X and present it in graphical form.
- b What is the most likely category of living arrangement in Australia in 2016?
- c What is the probability that a randomly selected person is from a 'one parent female' living arrangement category?
- d If a person is randomly selected from families living as a 'couple', what is the probability that the person is from a 'couple with children' living arrangement?

7.9 XR07-09 In an attempt to collect information about the level of participation of children in cultural and other leisure activities, data on the participation of children aged 5 to 14 years in selected organised cultural activities outside of school hours (during the 12 months prior to interview in April 2018) were collected. The selected cultural activities include drama activities, singing, playing a musical instrument, dancing, art and craft, creative writing and creating digital content. The data are presented below.

Participation in selected cultural activities by 5- to 14-year-old children, Australia, 2017–18

Number of activities (x)	Number of children ('000)		
	Male	Female	Total
One (1)	12.2	10.5	22.7
Two (2)	99.4	80.0	179.4
Three (3)	358.3	332.4	690.7
Four (4)	245.9	314.1	560.0
Five (5)	112.0	197.6	309.6
Six or more (6)	46.3	129.8	176.1
Total population aged 5–14 years	874.1	1064.4	1938.5

Source: Australian Bureau of Statistics, 2017–18, *Children's Participation in Cultural and Leisure Activities, Australia*, Table 11, cat. no. 4901.0, ABS, Canberra.

- a Construct the probability distribution of X and present it in graphical form for:
 - i female children
 - ii male children
 - iii all children.
- b If a child is selected at random, what is the probability that the child would have participated in only one cultural and leisure activity?
- c What is the probability that a randomly selected male child would have participated in three or more cultural and leisure activities?

- d What is the probability that a randomly selected female child would have participated in four or more cultural and leisure activities?

7.10 Using historical data, the personnel manager of a car manufacturing company has determined that the probability distribution of X , the number of employees absent on any given day, is as follows:

x	0	1	2	3	4	5	6	7
$p(x)$	0.005	0.025	0.310	0.340	0.220	0.080	0.019	0.001

- a Express the probability distribution of X in graphical form.
- b Comment on the symmetry or skewness of the distribution.
- c Find the probability that $2 \leq X \leq 4$.
- d Find the probability that $X > 5$.
- e Find the probability that $X \leq 6$.
- f Find the probability of there being more than one absentee on a given day.

7.11 After watching a number of children playing games at a video arcade, a statistics practitioner estimated the following probability distribution of X , the number of games played per visit.

x	1	2	3	4	5	6	7
$p(x)$	0.05	0.15	0.15	0.25	0.20	0.10	0.10

- a What is the probability that a child will play more than four games?
- b What is the probability that a child will play at least two games?

7.12 A survey of Amazon.com shoppers reveals the following probability distribution of the number of books purchased per hit:

x	0	1	2	3	4	5	6	7
$p(x)$	0.35	0.25	0.20	0.08	0.06	0.03	0.02	0.01

- a What is the probability that an Amazon.com visitor will buy four books?
- b What is the probability that an Amazon.com visitor will buy eight books?
- c What is the probability that an Amazon.com visitor will not buy any books?
- d What is the probability that an Amazon.com visitor will buy at least one book?

7.13 An internet pharmacy advertises that it will deliver the over-the-counter products that customers purchase within 3 to 6 days. The manager of the company wanted to be more precise in her

advertising. Accordingly, she recorded the number of days (X) it took to deliver to customers. From the data the following probability distribution was developed:

x	0	1	2	3	4	5	6	7	8
p(x)	0.00	0.00	0.01	0.04	0.28	0.42	0.21	0.02	0.02

- a What is the probability that a delivery will be made within the advertised period of 3 to 6 days?
- b What is the probability that a delivery will be late?
- c What is the probability that a delivery will be early?

- 7.14** An expensive restaurant conducted an analysis of the number of people at tables, from which this probability distribution was developed.

x	1	2	3	4	5	6	7	8
p(x)	0.03	0.32	0.05	0.28	0.04	0.15	0.03	0.10

If one table is selected at random, determine the probability of the following events.

- a The table has more than 4 people.
- b The table has fewer than 5 people.
- c The table has between 4 and 6 people (inclusive).

- 7.15** When parking a car in a metered parking spot, drivers pay according to the number of hours or parts thereof. The probability distribution of the number of hours cars are parked has been estimated as follows.

x	1	2	3	4	5	6	7	8
p(x)	0.24	0.18	0.13	0.10	0.07	0.04	0.04	0.20

Determine the probability of the following events.

- a The car will be parked for 4 hours.
- b The car will be parked for less than 4 hours.
- c The car will be parked for more than 4 hours
- d The car will be parked for 7 hours or more.

7.2 Expected value and variance

In Chapter 5, we showed how to calculate the mean, the variance and the standard deviation of a population. The formulas we provided were based on knowing the value of the random variable for each member of the population. For example, if we want to know the mean and variance of the annual income of all Australian blue-collar workers, we would record each of their incomes and use the formulas for a population introduced in Chapter 5:

$$\mu = \frac{\sum_{i=1}^N x_i}{N} = \sum_{i=1}^N x_i \cdot \left(\frac{1}{N} \right)$$

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} = \sum_{i=1}^N (x_i - \mu)^2 \cdot \left(\frac{1}{N} \right)$$

where x_1 is the income of the first blue-collar worker, x_2 is the second worker's income, and so on. It is likely that N equals several million. As you can appreciate, these formulas are seldom used in practical applications because populations are so large. It is unlikely that we would be able to record all the incomes in the population of Australian blue-collar workers. However, probability distributions represent populations.

Rather than record each of the many observations in a population, we list the values and their associated probabilities, as we did in deriving the probability distribution of the household size in Example 7.1, and the number of successes in three calls by the mutual fund salesperson in Example 7.2. These can be used to calculate the mean and variance of the population.

7.2a Population mean and variance

The **population mean** is the weighted average of all of its possible values. The weights are the probabilities. This parameter is also called the **expected value** of X and is represented by $E(X)$.

population mean

The average of all its possible values, denoted by μ .

expected value

The sum of all possible values a random variable can take times the corresponding probabilities.

Population mean or expected value

Given a discrete random variable X with possible values x that occur with probability $p(x)$, the expected value of X is

$$\mu = E(X) = \sum_{\text{all } x} xp(x)$$

The expected value of X should be interpreted simply as a weighted average of the possible values of X , rather than as a value that X is expected to assume. In fact, $E(X)$ may not even be a possible value of X , as we will illustrate in Example 7.3.

The population variance is the weighted average of the squared deviations from the mean.

Population variance

Let X be a discrete random variable with possible values x that occur with probability $p(x)$, and let $E(X) = \mu$. The *variance* of X is defined as

$$\sigma^2 = V(X) = \sum_{\text{all } x} (x - \mu)^2 p(x)$$

Notice that a variance is always non-negative, since each item in the summation is non-negative. The notion of variance is, therefore, chiefly used to compare the variabilities of different distributions, which may (for example) represent the possible outcomes of alternative courses of action under consideration. One important application arises in finance, where variance is the most popular numerical measure of risk; the underlying assumption is that a large variance corresponds to a higher level of risk.

There is a shortcut calculation that simplifies the calculations for the population variance. This formula is not an approximation; it will yield the same value as the previous formula.

Shortcut calculation for population variance

$$\sigma^2 = V(X) = \sum_{\text{all } x} x^2 p(x) - \mu^2$$

The *standard deviation* is defined as in Chapter 5. As was the case in Chapter 5 with a set of measurement data, we may wish to express the variability of X in terms of a measure that has the same unit as X . Once again, this is accomplished by taking the positive square root of the variance.

Population standard deviation

The *standard deviation* of a random variable X , denoted σ , is the positive square root of the variance of X .

$$\sigma = \sqrt{V(X)} = \sqrt{\sigma^2}$$

EXAMPLE 7.3

LO3

Describing the population of the number of car sales

Now that the new models are available, a car dealership has lowered the prices on last year's models in order to clear its holdover inventory. With prices slashed, a salesman estimates the following probability distribution of X , the total number of cars that he will sell next week:

x	0	1	2	3	4
p(x)	0.15	0.15	0.35	0.25	0.10

Find the mean, the variance and the standard deviation for the population of the number of cars he will sell next week.

Solution

The mean of X (or the expected number of cars the salesman will sell next week) is:

$$\begin{aligned}\mu &= E[X] = \sum_{\text{all } x} xp(x) \\ &= 0p(0) + 1p(1) + 2p(2) + 3p(3) + 4p(4) \\ &= 0(0.15) + 1(0.15) + 2(0.35) + 3(0.25) + 4(0.10) \\ &= 2.0 \text{ cars}\end{aligned}$$

The variance of X is:

$$\begin{aligned}\sigma^2 &= V(X) = \sum_{\text{all } x} (x - \mu)^2 p(x) \\ &= (0 - 2.0)^2 p(0) + (1 - 2.0)^2 p(1) + (2 - 2.0)^2 p(2) + (3 - 2.0)^2 p(3) + (4 - 2.0)^2 p(4) \\ &= (0 - 2.0)^2 (0.15) + (1 - 2.0)^2 (0.15) + (2 - 2.0)^2 (0.35) + (3 - 2.0)^2 (0.25) + (4 - 2.0)^2 (0.10) \\ &= 1.40 \text{ (cars)}^2\end{aligned}$$

To demonstrate the shortcut method, we'll use it to re-compute the variance:

$$\begin{aligned}\sum_{\text{all } x} x^2 p(x) &= 0^2 p(0) + 1^2 p(1) + 2^2 p(2) + 3^2 p(3) + 4^2 p(4) \\ &= 0^2 (0.15) + 1^2 (0.15) + 2^2 (0.35) + 3^2 (0.25) + 4^2 (0.10) \\ &= 5.40\end{aligned}$$

Thus,

$$\sigma^2 = \sum_{\text{all } x} x^2 p(x) - \mu^2 = 5.40 - (2.0)^2 = 1.40 \text{ (cars)}^2$$

The standard deviation is:

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.40} = 1.18 \text{ cars}$$

These parameters tell us that the mean and standard deviation of the number of cars the salesman will sell next week are 2 cars and 1.18 cars respectively.

A convenient alternative for calculation purposes is to record the probability distribution of X (and subsequent calculations) in a table such as **Table 7.4**. Rather than having a column for $(x - \mu)^2$, we have chosen to use the shortcut formula for variance, which entails finding the expected value of X^2 .

**TABLE 7.4** Calculations for $E(X)$ and $E(X^2)$

x	p(x)	xp(x)	x²	x²p(x)
0	0.15	0	0	0
1	0.15	0.15	1	0.15
2	0.35	0.70	4	1.40
3	0.25	0.75	9	2.25
4	0.10	0.40	16	1.60
Total	$\sum_{\text{all } x} xp(x) = 2.0$		$\sum_{\text{all } x} x^2 p(x) = 5.40$	

Therefore, from **Table 7.4**:

$$\mu = E(X) = \sum_{\text{all } x} xp(x) = 2.0 \text{ cars}$$

$$\sigma^2 = V(X) = \sum_{\text{all } x} x^2 p(x) - \mu^2 = 5.40 - (2.0)^2 = 1.40 \text{ (cars)}^2$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.40} = 1.18 \text{ cars}$$

While the mean of 2 cars is a measure of the centre of the distribution, the standard deviation of 1.18 cars measures the dispersion of the values about the centre.

7.2b Laws of expected value and variance

As you will discover, we often create new random variables that are functions of existing random variables. The formulas given below allow us to quickly determine the expected value and variance of these new random variables.

Laws of expected value and variance

In the notation used here, if X is the random variable and c and d are two constants, then $cX + d$ is another random variable with a mean (expected value) and variance, given below.

Law of expected value

$$E(c) = c$$

$$E(X + c) = E(X) + c$$

$$E(cX) = cE(X)$$

$$E(cX + d) = cE(X) + d$$

Law of variance

$$V(c) = 0$$

$$V(X + c) = V(X)$$

$$V(cX) = c^2 V(X)$$

$$V(cX + d) = c^2 V(X)$$

EXAMPLE 7.4

LO3

Describing the population of monthly profits

The monthly sales at a computer store have a mean of \$25 000 and a standard deviation of \$4000. Profits are calculated by multiplying sales value by 30% and subtracting fixed costs of \$6000. Find the mean and standard deviation of monthly profits.

Solution

We can describe the relationship between profits and sales by the following equation:

$$\text{Profit} = 0.30(\text{Sales}) - 6000$$

Also, $E(\text{Sales}) = \$25\,000$ and $SD(\text{Sales}) = \sigma_{\text{Sales}} = \4000 . Rather than performing the mind-numbing calculations in **Table 7.4**, we could simply use the laws of expected value and variance listed above with $c = 0.30$ and $d = -6000$.

The expected or mean profit is:

$$E(\text{Profit}) = E[0.30(\text{Sales}) - 6000]$$

Applying the law of expected value yields:

$$E(\text{Profit}) = 0.30E(\text{Sales}) - 6000 = 0.30(25\,000) - 6000 = \$1500$$

Thus, the mean monthly profit is \$1500.

The variance of profit is:

$$V(\text{Profit}) = V[0.30(\text{Sales}) - 6000]$$

Applying the law of variance yields:

$$V(\text{Profit}) = 0.30^2V(\text{Sales}) = 0.09(4000)^2 = 1440\,000$$

Thus, the standard deviation of monthly profit is:

$$\sigma_{\text{Profit}} = \sqrt{1440\,000} = 1200$$

The expected monthly profit is therefore \$1500 with a standard deviation of \$1200.

EXERCISES**Learning the techniques**

- 7.16** Let X be a random variable with the following probability distribution:

x	5	10	15	20	25
$p(x)$	0.05	0.30	0.25	0.25	0.15

- a Find the expected value and variance of X .
- b Find the expected value and variance of $Y = 4X - 3$.
- c Find $V(Y)$.

- 7.17** Let X be a random variable with the following probability distribution:

x	-10	-5	0	5	10
$p(x)$	0.10	0.20	0.20	0.20	0.30

- a Find the mean, variance and standard deviation of X .
- b Find the mean, variance and standard deviation of $2X$.
- c Find the mean, variance and standard deviation of $2X + 5$.

- 7.18** The number of pizzas delivered to university students each month is a random variable (X) with the following probability distribution:

x	0	1	2	3
$p(x)$	0.1	0.3	0.4	0.2

- a Find the probability that a student has received delivery of two or more pizzas this month.
- b Determine the mean and variance of the number of pizzas delivered to students each month.

- c** If the pizza shop makes a profit of \$3 per pizza, determine the mean and variance of the profit per student.
- 7.19** A shopping mall estimates the probability distribution of the number of stores mall customers actually enter, as shown in the table.
- | x | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|--------|------|------|------|------|------|------|------|
| $p(x)$ | 0.04 | 0.19 | 0.22 | 0.28 | 0.12 | 0.09 | 0.06 |
- a** Find the mean and standard deviation of the number of stores entered.
b Suppose that, on average, customers spend 10 minutes in each store they enter. Find the mean and standard deviation of the total amount of time customers spend in stores.
- Applying the techniques**
- 7.20 Self-correcting exercise.** The owner of a small firm has just purchased a personal computer that she expects will serve her for the next two years. The owner has been told that she 'must' buy a surge suppressor to provide protection for her new hardware against possible surges or variations in the electrical current, which have the capacity to damage the computer. The amount of damage to the computer depends on the strength of the surge. It has been estimated that there is a 1% chance of incurring \$400 damage, a 2% chance of incurring \$200 damage, and 10% chance of \$100 damage. An inexpensive suppressor, which would provide protection for only one surge can be purchased. How much should the owner be willing to pay if she makes decisions on the basis of expected value? Determine the standard deviation.
- 7.21** Suppose that you have the choice of receiving \$500 in cash or a gold coin that has a face value of \$100. The actual value of the gold coin depends on its gold content. You are told that the coin has a 40% chance of being worth \$400, a 30% chance of being worth \$900, and a 30% chance of being worth its face value. If you base your decision on expected value, which should you choose?
- 7.22** To examine the effectiveness of its four annual advertising promotions, a mail-order company has sent a questionnaire to each of its customers, asking how many of the previous year's promotions prompted orders that otherwise would not have been made. The following table summarises the data received, where the random variable X is the number of promotions indicated in the customers' responses.
- | x | 0 | 1 | 2 | 3 | 4 |
|--------|------|------|------|------|------|
| $p(x)$ | 0.10 | 0.25 | 0.40 | 0.20 | 0.05 |
- a** If we assume that the responses received were accurate evaluations of individual effectiveness, and that customer behaviour next year will be the same as last year, what is the expected number of promotions that each customer will take advantage of next year by ordering goods that otherwise would not be purchased?
b What is the standard deviation of X ?
- 7.23** Refer to Exercise 7.22. A previous analysis of historical data found that the mean value of orders for promotional goods is \$20, with the company earning a gross profit of 20% on each order.
- a** Calculate the expected value of the profit contribution next year.
b The fixed cost of conducting the four promotions next year is estimated to be \$15 000, with a variable cost of \$3.00 per customer for mailing and handling costs. Assuming that the survey results can be used as an accurate predictor of behaviour for existing and potential customers, how large a customer base must the company have to cover the cost of the promotions?
- 7.24** Suppose that you and a friend have contributed equally to a portfolio of \$10 000, which was invested in a risky venture. The income (X) that will be earned on this portfolio over the next year has the following probability distribution:
- | x | \$500 | \$1000 | \$2000 |
|--------|-------|--------|--------|
| $p(x)$ | 0.5 | 0.3 | 0.2 |
- a** Determine the expected value and the variances of the income earned on this portfolio.
b Determine the expected value and the variance of your share (half) of the income. Answer the question first by calculating the expected value and the variance directly from the probability distribution of the income you will receive. Then check your answer using the laws of expected value and variance.
- 7.25** The probability that a university graduate will be offered no jobs within a month of graduation is estimated to be 5%. The probability of receiving one, two and three job offers has similarly been estimated to be 43%, 31%, and 21% respectively.

- a Determine the probability that a graduate is offered fewer than two jobs.
- b Determine the probability that a graduate is offered more than one job.
- c Determine the expected number of job offers a graduate would receive.

7.26 It costs \$1 to buy a lottery ticket that has five prizes. The prizes and the probability that a player wins the prize are listed here. Calculate the expected value of the payoff.

Prize	Probability
\$1 million	1/10000000
\$200 000	1/1 000 000
\$50 000	1/500 000
\$10 000	1/50 000
\$1 000	1/10 000

- 7.27** After an analysis of incoming faxes, the manager of an accounting firm determined the probability distribution of the number of pages per fax as follows:

x	1	2	3	4	5	6	7
p(x)	0.05	0.12	0.20	0.30	0.15	0.10	0.08

Calculate the mean and variance of the number of pages per fax.

- 7.28** Refer to Exercise 7.27. Further analysis by the manager revealed that the cost of processing each page of a fax is \$0.25. Determine the mean and variance of the cost per fax.

7.3 Binomial distribution

Now that we have introduced probability distributions in general, we need to introduce several specific probability distributions. In this section we present the binomial distribution.

7.3a Binomial experiment

The binomial distribution is the result of a **binomial experiment**, which has the following properties:

Binomial experiment

- 1 The binomial experiment consists of a fixed number of repeated trials, denoted by n .
- 2 On each trial there are two possible outcomes, labelled as a **success** and a **failure**.
- 3 The probability of success is p . The probability of failure is $q = 1 - p$.
- 4 The trials are independent, which means that the outcome of one trial does not affect the outcome of any other trial.

If properties 2, 3, and 4 are satisfied, we say that each trial is a **Bernoulli trial**. Adding property 1 yields the binomial experiment. The random variable of a binomial experiment is defined as the total number of successes in the n trials. It is called the **binomial random variable**.

Here are several examples of binomial experiments.

- 1 Flip a coin 10 times. The two outcomes in each trial are heads and tails. The terms success and failure are arbitrary. We can label either outcome as a success. However, generally, we call success anything we're looking for. For example, if we were betting on heads, we would label heads a success. If the coin is fair, the probability of heads is 50%. Thus, $p = 0.5$. Finally, we can see that the trials are independent, because the outcome of one coin flip cannot possibly affect the outcomes of other flips. If X counts the total number of heads in 10 flips, then X is a binomial random variable and can take values 0, 1, 2, ..., 10.

binomial experiment

An experiment consisting of a number of repeated Bernoulli trials.

success

An arbitrary label given to one of the outcomes of a Bernoulli trial.

failure

The non-success outcome of a Bernoulli trial.

Bernoulli trial

A random experiment that has only two possible outcomes.

binomial random variable

The number of successes in the n trials of a binomial experiment.

- 2 Test 500 randomly selected computer chips produced at a manufacturing facility and determine whether they are defective. The number of trials is 500. There are two outcomes in each trial: the product is either defective or non-defective. Assuming the defective rate is 1% and labelling the occurrence of a defective chip to be a success, then $p = 0.01$ and $q = (1 - p) = 0.99$. If the computer chips are selected at random for testing, the trials are independent. If X counts the number of defective chips in the 500 chips tested, then X is a binomial random variable and can take values 0, 1, 2, ..., 500.
- 3 A political survey asks 1500 voters whether they would vote for the Australian Labor Party (ALP) if an election were held next week. The response would be either 'yes' or 'no'. Thus, we have two outcomes per trial. The trials are independent, because the choice of one voter does not affect the choice of other voters. We can label the answer 'yes' (vote for Labor) as a success and the answer 'no' as a failure.

If X counts the number of voters supporting the ALP, then X is a binomial random variable and can take values 0, 1, 2, ..., 1500. Also, for example, if every 6 out of 10 voters support ALP, then $p = 0.6$.

As you will discover, the third example is a very common application of statistical inference. The actual value of p , the probability of a success, is unknown, and the job of the statistics practitioner is to estimate its value. By understanding the probability distribution that uses p , we will be able to develop the statistical tools to estimate p .

Notice that in each example we made an assumption that allowed us to assign a value to p . Also note that we need to define what we mean by a success. In general, a success is defined arbitrarily and is not necessarily something we want to happen. The random variable of interest in these experiments is the total number of successes, and is called the binomial random variable.

7.3b Binomial probability distribution

binomial probability distribution

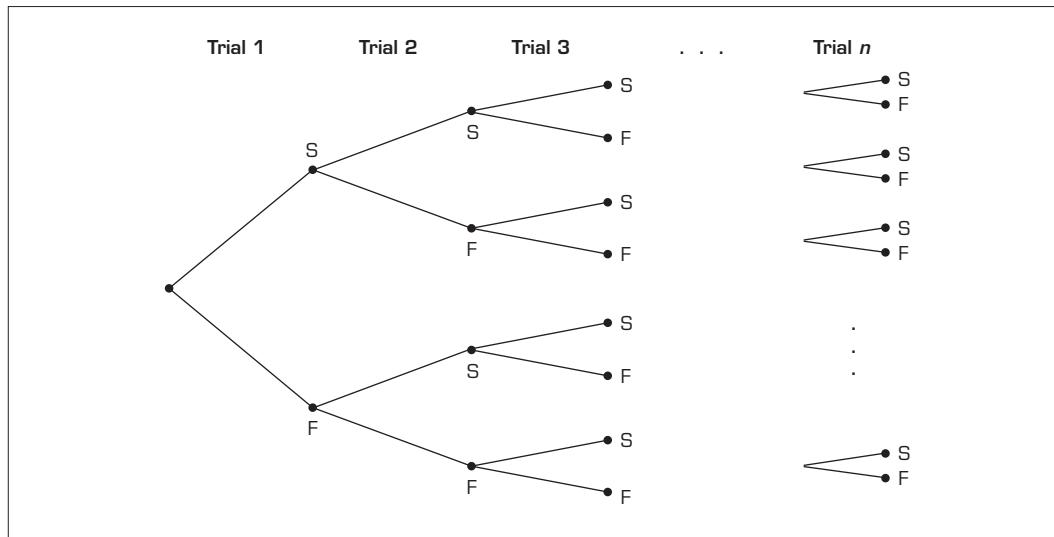
The probability distribution of the binomial random variable.

A binomial random variable is therefore a discrete random variable that can take on any one of the values 0, 1, 2, ..., n . The probability distribution of this random variable, called the **binomial probability distribution**, gives us the probability that a success will occur x times in the n trials, for $x = 0, 1, 2, \dots, n$. To proceed, we must be capable of calculating the probability associated with each value. Rather than working out binomial probabilities from scratch each time, we would do better to have a general formula for calculating the probabilities associated with any binomial experiment.

Using a probability tree, we draw a series of branches as depicted in **Figure 7.4**. The stages represent the outcomes for each of the n trials. At each stage there are two branches representing success and failure. To calculate the probability that there are x successes in n independent trials, we must multiply each success in the sequence by p . And, if there are x successes, there must be $n - x$ failures. We multiply each failure in the sequence by $1 - p$. Thus, the probability for each sequence of branches that represent x independent successes and $(n - x)$ independent failures has the probability:

$$p^x (1-p)^{n-x}$$

There are a number of branches that yield x successes and $n - x$ failures. For example, there are two ways to produce exactly one success and one failure in two trials: SF and FS. To count the number of branch sequences that produce x successes and $n - x$ failures, we use the well-known counting rule.

FIGURE 7.4 Probability tree for a binomial experiment

Counting rule

The number of different ways of choosing x objects from a total n objects is found using the combinatorial formula.

$$C_x^n = \frac{n!}{x!(n-x)!}$$

The notation $n!$ is read ‘ n factorial’ and defined as $n! = n(n - 1)(n - 2) \dots (2)(1)$. For example, $3! = (3)(2)(1) = 6$. Incidentally, although it may not appear to be logical, $0! = 1$.

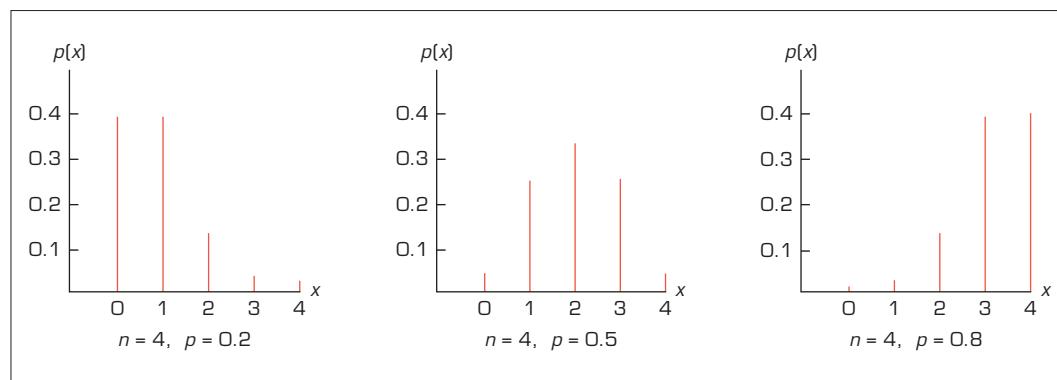
Pulling together the two components of the probability distribution yields the following general formulation of the binomial probability distribution.

Binomial probability distribution

If the random variable X is the total number of successes in the n independent trials of a binomial experiment that has probability p of a success on any given trial, the probability distribution of X is given by

$$p(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \text{ for } x = 0, 1, 2, \dots, n$$

Each pair of values (n, p) determines a distinct binomial distribution. Graphical representations of three binomial distributions are shown in **Figure 7.5**. Each of the $(n + 1)$ possible values of a binomial random variable X has a positive probability of occurring. The fact that some possible values of X do not have a vertical line above them in **Figure 7.5** simply means that the probability that those values will occur is too small to be displayed on the graph. A binomial distribution is symmetrical whenever $p = 0.5$, and it is asymmetrical otherwise.

FIGURE 7.5 Graphs of three binomial distributions**EXAMPLE 7.5**

LO4 LO5

Pat Statsdud and the statistics quiz

Pat Statsdud is a student taking a statistics course. Unfortunately, Pat is not a good student. Pat does not read the textbook before class, does not do homework and regularly misses classes. Pat intends to rely on luck to pass the next quiz. The quiz consists of 10 multiple-choice questions. Each question has five possible answers, only one of which is correct. Pat plans to guess the answer to each question.

- What is the probability that Pat gets no answers correct?
- What is the probability that Pat gets two answers correct?
- What is the probability that Pat gets all answers correct?

Solution

The experiment consists of 10 identical trials, each with two possible outcomes and where success is defined as a correct answer. Because Pat intends to guess, the probability of success is $1/5$ or 0.2. Finally, the trials are independent because the outcome of any of the questions does not affect the outcomes of any other questions. These four properties tell us that the experiment is binomial with $n = 10$ and $p = 0.2$. Let X denote the number of correct answers in the 10 questions.

- From

$$p(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, 2, \dots, 10$$

we get the probability of no successes where $n = 10$, $p = 0.2$, and $x = 0$. Hence,

$$p(0) = \frac{10!}{0!(10-0)!} (0.2)^0 (1-0.2)^{10-0}$$

The combinatorial part of the formula is $10!/(0! \times 10!)$, which is 1. This is the number of ways to get 0 correct and 10 incorrect. Obviously, there is only one way to produce $x = 0$ and because $(0.2)^0 = 1$,

$$p(0) = (0.2)^0 (0.8)^{10} = 0.1074$$

- The probability of two correct answers is computed similarly by substituting $n = 10$, $p = 0.2$ and $x = 2$:

$$\begin{aligned} p(x) &= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\ p(2) &= \frac{10!}{2!(10-2)!} (0.2)^2 (1-0.2)^{10-2} \\ &= \frac{(10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1)}{(2 \times 1)(8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1)} (0.04)(0.1678) \\ &= 45(0.04)(0.1678) \\ &= 0.3020 \end{aligned}$$



In this calculation we discovered that there are 45 ways to get exactly two correct and eight incorrect answers, and that each such outcome has probability 0.006712. Multiplying the two numbers produces a probability of 0.3020.

- c The probability of all correct answers (10) is computed similarly by substituting $n = 10$, $p = 0.2$ and $x = 10$:

$$\begin{aligned} p(x) &= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\ p(10) &= \frac{10!}{10!(10-10)!} (0.2)^{10} (1-0.2)^{10-10} \\ &= \frac{10}{10 \times 1} (0.2)^{10} \\ &= 0.0000 \end{aligned}$$

The probability of Pat getting all (10) correct answers is almost zero. Similarly we can calculate the probability distribution of X , which is presented in **Table 7.5**.

TABLE 7.5 Binomial distribution ($n = 10$, $p = 0.2$)

x	0	1	2	3	4	5	6	7	8	9	10
p(x)	0.1074	0.2684	0.3020	0.2013	0.0881	0.0264	0.0055	0.0008	0.0001	0.0000	0.0000

7.3c Cumulative probability

The formula of the binomial distribution allows us to determine the probability that X equals individual values. There are many circumstances where we wish to find the probability that a random variable is less than or equal to a value. That is, we want to determine $P(X \leq x)$, where x is that value. Such a probability is called a **cumulative probability**.

cumulative probability
The probability that a random variable X is less than or equal to x , $P(X \leq x)$.

EXAMPLE 7.6

LO5

Will Pat Statsdud fail the quiz?

Find the probability that Pat fails the quiz. A mark is considered a failure if it is less than 50%.

Solution

In this quiz, a mark of less than 5 is a failure. Because the marks must be integers, a mark of 4 or less is a failure. We wish to determine $P(X \leq 4)$. So:

$$P(X \leq 4) = p(0) + p(1) + p(2) + p(3) + p(4)$$

From **Table 7.5**, we know $p(0) = 0.1074$, $p(1) = 0.2684$, $p(2) = 0.3020$, $p(3) = 0.2013$ and $p(4) = 0.0881$. Thus:

$$P(X \leq 4) = 0.1074 + 0.2684 + 0.3020 + 0.2013 + 0.0881 = 0.9672$$

There is a 96.72% probability that Pat will fail the quiz by guessing the answer for each question.

7.3d Binomial table

There is another way to determine binomial probabilities. Table 1 in Appendix B provides cumulative binomial probabilities for selected values of n and p . We can use this table to answer the question in Example 7.6, where we need $P(X \leq 4)$. Refer to Table 1 in Appendix B, find $n = 10$, and in that table find $p = 0.20$. The values in that column are for $P(X \leq x)$ for $x = 0, 1, 2, \dots, 10$, which are shown in **Table 7.6**.

TABLE 7.6 Cumulative binomial probabilities with $n = 10$ and $p = 0.2$

X	0	1	2	3	4	5	6	7	8	9	10
P(X ≤ x)	0.1074	0.3758	0.6778	0.8791	0.9672	0.9936	0.9991	0.9999	1.0000	1.0000	1.0000

The first cumulative probability is $P(X \leq 0)$, which is $p(0) = 0.1074$. The probability we need for Example 7.6 is $P(X \leq 4) = 0.9672$, which is the same value we obtained manually using four decimal places.

We can use the table and the complement rule to determine probabilities of the type $P(X \geq x)$. For example, to find the probability that Pat will pass the quiz from Example 7.6, we note that:

$$P(X \leq 4) + P(X \geq 5) = 1$$

Thus:

$$P(X \geq 5) = 1 - P(X \leq 4) = 1 - 0.9672 = 0.0328$$

Using the binomial table to find the binomial probability $P(X \geq x)$

$$P(X \geq x) = 1 - P(X \leq x - 1)$$

The table is also useful in determining the probability of an individual value of X . For example, to find the probability that Pat will get exactly two right answers, we note that

$$P(X \leq 2) = p(0) + p(1) + p(2)$$

and

$$P(X \leq 1) = p(0) + p(1)$$

The difference between these two cumulative probabilities is $p(2) = P(X = 2)$. Thus:

$$\begin{aligned} p(2) &= P(X = 2) = P(X \leq 2) - P(X \leq 1) \\ &= 0.6778 - 0.3758 = 0.3020 \end{aligned}$$

Using the binomial table to find the binomial probability $P(X = x)$

$$P(X = x) = p(x) = P(X \leq x) - P(X \leq x - 1)$$

7.3e Using the computer to find binomial probabilities

There are a few software packages that calculate probabilities in addition to their more traditional duties involving statistical procedures. Excel is one of these, allowing us to easily produce the probability distributions for various random variables.

We can use Excel to calculate cumulative probabilities and the probability of individual values of binomial random variables.

COMMANDS

To calculate the probabilities associated with a binomial random variable X with the number of trials n and the probability of a success p , type the following into any active cell:

=BINOMDIST([x],[n],[p],[True] or [False])

Typing 'True' calculates a cumulative probability, $P(X \leq x)$, and typing 'False' calculates the probability of an individual value of X , $P(X = x)$.

For example, to calculate $P(X = 3)$ for a binomial random variable with $n = 10$ and $p = 0.2$, type into any active cell **=BINOMDIST(3,10,.2,False)**, which will calculate the probability value of 0.201327. To calculate $P(X \leq 3)$, type **=BINOMDIST(3,10,.2,True)**, which will calculate the value of 0.879126.

7.3f Mean and variance of binomial distribution

Statisticians have developed general formulas for the mean, the variance and the standard deviation of a binomial random variable.

Mean and variance of binomial random variables

If X is a binomial random variable, the mean and the variance of X are

$$E(X) = \mu = np$$

$$V(X) = \sigma^2 = npq$$

$$SD(X) = \sigma = \sqrt{npq}$$

where n is the number of trials, p is the probability of success in any trial, and $q = (1 - p)$ is the probability of failure in any trial.

EXAMPLE 7.7

Pat Statsdud has been cloned!

Suppose that a professor has a class full of students like Pat (a nightmare!). What is the mean number of correct answers? What is the standard deviation?

Solution

Let X denote the number of correct answers in the 10 questions, the experiment is binomial with $n = 10$ and $p = 0.2$. The mean number of correct answers for a class of Pat Statsduds is

$$\mu = E(X) = np = 10(0.2) = 2$$

The variance is

$$\sigma^2 = npq = 10(0.2)(1 - 0.2) = 10(0.2)(0.8) = 1.6$$

The standard deviation is

$$\sigma = \sqrt{1.6} = 1.26$$

EXAMPLE 7.8

LO5

Likely credit card payment for shoe purchase

A shoe store's records show that 30% of customers making a purchase use a credit card to pay. This morning, 20 customers purchased shoes from the store. Answer the following, making use of Table 1 in Appendix B.

- Find the probability that at least 12 of the customers used a credit card.
- What is the probability that at least three customers, but not more than six, used a credit card?
- What is the expected number of customers who used a credit card? What is the standard deviation?
- Find the probability that exactly 14 customers did not use a credit card.
- Find the probability that at least nine customers did not use a credit card.

Solution

If making payment with a credit card is designated as a success, we have a binomial experiment with $n = 20$ and $p = 0.3$. Let X denote the number of customers who used a credit card.

- We must first express the probability we seek in the form $P(X \leq k)$, as this is the form in which probabilities are tabulated in the binomial table (Table 1 in Appendix B).

$$\begin{aligned} P(X \geq 12) &= P(X = 12) + P(X = 13) + \cdots + P(X = 20) \\ &= P(X \leq 20) - P(X \leq 11) \end{aligned}$$

The probabilities in a binomial distribution ($n = 20$) must sum to 1, so $P(X \leq 20) = 1$. From Table 1 in Appendix B, $P(X \leq 11) = 0.995$. Therefore,

$$P(X \geq 12) = 1 - 0.995 = 0.005$$

The probability that at least 12 customers used a credit card is 0.005.

- Expressing the probability we seek in the form used for the probabilities tabulated in the binomial table, we have

$$\begin{aligned} P(3 \leq X \leq 6) &= P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6) \\ &= P(X \leq 6) - P(X \leq 2) \\ &= 0.608 - 0.035 \\ &= 0.573 \end{aligned}$$

The probability that between three and six customers used a credit card is 0.573.

- The expected number of customers who used a credit card is

$$E(X) = np = 20(0.3) = 6$$

The standard deviation is

$$\sigma = \sqrt{npq} = \sqrt{20(0.3)(0.7)} = \sqrt{4.2} = 2.05$$

- Let Y denote the number of customers who did not use a credit card. The probability that a customer did not use a credit card is $(1 - 0.3) = 0.7$. This part of the example can be solved in either of two ways:
 - You can interchange the designations of success and failure and work with $p = 0.7$.
 - You can express the required probability in terms of the number of customers who did not use a credit card, and proceed with $p = 0.3$.

Method (i) is probably easier to use with the tables in the text. In many cases, however, binomial tables with p values above 0.5 are not available, and method (ii) must be used.



Using method (i) begins with recognising that, since the original assignment of the designations *success* and *failure* was arbitrary, we may interchange them. If *not* using a credit card is designated as a success, then $p = 0.7$. From the binomial table, we find that:

$$\begin{aligned} P(Y=14) &= P(Y \leq 14) - P(Y \leq 13) \\ &= 0.584 - 0.392 \\ &= 0.192 \end{aligned}$$

In method (ii) we retain the original designation, according to which using a credit card is a success and $p = 0.3$. If 14 customers did not use a credit card, the number of customers who did use one is $(20 - 14) = 6$. Hence,

$$\begin{aligned} P(Y=14) &= P(X=6) \\ &= P(X \leq 6) - P(X \leq 5) \\ &= 0.608 - 0.416 \\ &= 0.192 \end{aligned}$$

Using either method, we find that the probability that exactly 14 customers did not use a credit card is 0.192.

- e** Again, let Y denote the number of customers who did not use a credit card. If *not* using a credit card is designated as a success, then $p = 0.7$. Expressing the required probability in terms of values tabulated in the binomial table, we have

$$\begin{aligned} P(Y \geq 9) &= 1 - P(Y \leq 8) \\ &= 1 - 0.005 \\ &= 0.995 \end{aligned}$$

The probability that at least nine customers did not use a credit card is 0.995.

EXERCISES

Learning the techniques

- 7.29** Let x be a binomial random variable. Use the formula to calculate the following probabilities:
- $P(X = 2)$, if $n = 8$ and $p = 0.1$
 - $P(X = 5)$, if $n = 9$ and $p = 0.5$
 - $P(X = 9)$, if $n = 10$ and $p = 0.95$
- 7.30** Use Table 1 in Appendix B to check your answers to Exercise 7.29.
- 7.31** Given a binomial random variable x with $n = 15$ and $p = 0.3$, find the following probabilities, using Table 1 in Appendix B.
- $P(X \leq 2)$
 - $P(X \geq 7)$
 - $P(X = 6)$
 - $P(4 \leq X \leq 8)$

e $P(4 < X < 8)$

f $P(X \geq 12)$

Applying the techniques

- 7.32 Self-correcting exercise.** A multiple-choice quiz has 15 questions. Each question has five possible answers, of which only one is correct.
- What is the probability that sheer guesswork will yield at least seven correct answers?
 - What is the expected number of correct answers by sheer guesswork?
- 7.33** According to Stats NZ, 62.8% of women available to work were employed in June quarter 2018. A random sample of 25 women available to work in New Zealand was drawn. What is the probability that 15 or more are employed?

- 7.34** A sign on the petrol pumps of a chain of petrol stations encourages customers to have their oil checked, claiming that one out of every four cars should have its oil topped up.
- Of the next 10 cars entering a petrol station, what is the probability that exactly three of them should have their oil topped up?
 - What is the probability that (i) at least half of the next 10 cars entering a petrol station should have their oil topped up and (ii) at least half of the next 20 cars should have their oil topped up?
- 7.35** A student majoring in accounting at Griffith University is trying to decide on the number of firms to which she should apply. Given her work experience, academic results and extracurricular activities, she has been told by a placement counsellor that she can expect to receive a job offer from 80% of the firms to which she applies. Wanting to save time, the student applied to five firms only. Assuming the counsellor's estimate is correct, find the probability that the student receives:
- no offers
 - at most two offers
 - between two and four offers (inclusive)
 - five offers.
- 7.36** An auditor is preparing for a physical count of inventory as a means of verifying its value. Items counted are reconciled with a list prepared by the storeroom supervisor. Normally 20% of the items counted cannot be reconciled without reviewing invoices. The auditor selects 10 items.
- a** Find the probability of each of the following:
- Up to four items cannot be reconciled.
 - At least six items cannot be reconciled.
 - Between four and six items (inclusive) cannot be reconciled.
- b** If it normally takes 20 minutes to review the invoice for an item that cannot be reconciled and one hour for the balance of the count, how long should the auditor expect the physical count to take?
- 7.37** The leading brand of dishwasher detergent has a 30% market share. A sample of 25 dishwasher detergent customers was taken. What is the probability that 10 or fewer customers chose the leading brand?
- 7.38** A certain type of tomato seed germinates 90% of the time. A gardener planted 25 seeds.
- What is the probability that exactly 20 seeds germinate?
 - What is the probability that 20 or more seeds germinate?
 - What is the probability that 24 or fewer seeds germinate?
 - What is the expected number of seeds that germinate?
- 7.39** According to the Australian Bureau of Statistics, in August 2018, 15% of employees reported being a member of a trade union in their main job. A random sample of 50 employees was drawn. What is the probability that 5 or more are members of a trade union in their main job?

7.4 Poisson distribution

Poisson probability distribution

The probability distribution of the Poisson random variable.

Poisson random variable

The number of successes that occur in a period of time or an interval of space in a Poisson experiment.

Another useful discrete probability distribution is the **Poisson probability distribution**, named after its French creator. Like the binomial random variable, the **Poisson random variable** is the number of occurrences of events, which we'll continue to call successes. The difference between the two random variables is that a binomial random variable is the number of successes in a set number of trials, whereas a Poisson random variable is the number of successes in an interval of time or specific region of space. Here are several examples of Poisson random variables:

- The number of cars arriving at a service station in one hour. (The interval of time is one hour.)
- The number of flaws in a bolt of cloth. (The specific region is a bolt of cloth.)
- The number of accidents in one day on a particular stretch of highway. (The interval is defined by both time, one day, and space, the particular stretch of highway.)

7.4a Poisson experiment

A **Poisson experiment** possesses the properties described in the box.

Poisson experiment

- 1** The number of successes that occur in any interval is independent of the number of successes that occur in any other interval.
- 2** The probability of a success in an interval is the same for all equal-size intervals.
- 3** The probability of a success in an interval is proportional to the size of the interval.
- 4** The probability of more than one success in an interval approaches 0 as the interval becomes smaller.

Poisson experiment
An experiment with rare outcomes.

As a general rule, a Poisson random variable is the number of occurrences of a *relatively rare* event that occurs *randomly* and *independently*. The number of hits on an active website is not a Poisson random variable, because the hits are not rare. The number of people arriving at a restaurant is not a Poisson random variable, because restaurant patrons usually arrive in groups, which violates the independence condition.

7.4b Poisson probability distribution

There are several ways to derive the probability distribution of a Poisson random variable. However, all are beyond the mathematical level of this book. We simply provide the formula and illustrate how it is used.

Probability distribution of Poisson random variables

If X is a Poisson random variable, the probability distribution of X is given by:

$$P(X = x) = p(x) = \frac{e^{-\mu} \mu^x}{x!} \text{ for } x = 0, 1, 2, \dots$$

where μ is the average number of successes occurring in the given time interval or region, and $e = 2.71828\dots$ is the base of the natural logarithms.

EXAMPLE 7.9

LO6

Probability distribution of the number of typographical errors in textbooks

A statistics instructor has observed that the number of typographical errors in new editions of textbooks varies considerably from book to book. After some analysis, he concludes that the number of errors is Poisson distributed with a mean of 1.5 per 100 pages. The instructor randomly selects 100 pages of a new book. What is the probability that there are no typographical errors?

Solution

Let X be the number of typographical errors in the 100 pages. The variable X is a Poisson random variable with mean 1.5. We want to determine the probability that X is equal to 0. Thus, we substitute into the formula for the Poisson distribution:

$$P(X = 0) = p(0) = \frac{e^{-1.5} 1.5^0}{0!} = \frac{(2.71828)^{-1.5} (1)}{1} = 0.2231$$

The probability that in the 100 pages selected there are no typographical errors is 0.2231.

Notice that in Example 7.9 we wanted to find the probability of 0 typos in 100 pages given a mean of 1.5 typos in 100 pages. The next example illustrates how we calculate the probability of events where the intervals or regions do not match.

EXAMPLE 7.10

LO6

Probability of the number of typographical errors in 400 pages

Refer to Example 7.9. Suppose that the instructor has just received a copy of a new statistics book. He notices that there are 400 pages.

- a What is the probability that there are no typos?
- b What is the probability that there are five or fewer typos?

Solution

The specific region that we're interested in is 400 pages. Let X be the number of typographical errors in the 400-page region. To calculate Poisson probabilities associated with this region, we must determine the mean number of typos per 400 pages. Because the mean is specified as 1.5 per 100 pages, we multiply this mean by four to convert to 400 pages. Thus, X is a Poisson random variable with mean $\mu = 6$ typos per 400 pages.

- a The probability of no typos in the 400-page book is:

$$P(X=0) = p(0) = \frac{e^{-6} 6^0}{0!} = \frac{(2.71828)^{-6} (1)}{1} = 0.002479$$

- b We want to determine the probability that the number of typos in the 400-page book (X) is five or less. That is, we want to calculate:

$$P(X \leq 5) = p(0) + p(1) + p(2) + p(3) + p(4) + p(5)$$

To produce this probability, we need to compute the six probabilities in the summation:

$$\begin{aligned} p(0) &= 0.002479 \\ p(1) &= \frac{e^{-\mu} \mu^x}{x!} = \frac{e^{-6} 6^1}{1!} = \frac{(2.71828)^{-6} (6)}{1} = 0.01487 \\ p(2) &= \frac{e^{-\mu} \mu^x}{x!} = \frac{e^{-6} 6^2}{2!} = \frac{(2.71828)^{-6} (36)}{2} = 0.04462 \\ p(3) &= \frac{e^{-\mu} \mu^x}{x!} = \frac{e^{-6} 6^3}{3!} = \frac{(2.71828)^{-6} (216)}{6} = 0.08924 \\ p(4) &= \frac{e^{-\mu} \mu^x}{x!} = \frac{e^{-6} 6^4}{4!} = \frac{(2.71828)^{-6} (1296)}{24} = 0.1339 \\ p(5) &= \frac{e^{-\mu} \mu^x}{x!} = \frac{e^{-6} 6^5}{5!} = \frac{(2.71828)^{-6} (7776)}{120} = 0.1606 \end{aligned}$$

Thus,

$$\begin{aligned} P(X \leq 5) &= 0.002479 + 0.01487 + 0.04462 + 0.08924 + 0.1339 + 0.1606 \\ &= 0.4457 \end{aligned}$$

The probability of observing five or fewer typos in this book is 0.4457.

7.4c Poisson table

As was the case with the binomial distribution, a table is available that makes it easier to compute Poisson probabilities of individual values of x as well as cumulative and related probabilities.

Like Table 1 in Appendix B for binomial probabilities, Table 2 in Appendix B can be used to determine the cumulative Poisson probabilities of the type $P(X \leq x)$ for selected values

of μ . This table makes it easy to find cumulative probabilities like that in Example 7.10, part (b), where we found $P(X \leq 5)$. To do so, find $\mu = 6$ in Table 2 in Appendix B. The values in that column are $P(X \leq x)$ for $x = 0, 1, 2, \dots$, which are shown in **Table 7.7**.

TABLE 7.7 Cumulative Poisson probabilities for $\mu = 6$

X	$P(X \leq x)$	X	$P(X \leq x)$
0	0.0025	10	0.9574
1	0.0174	11	0.9799
2	0.0620	12	0.9912
3	0.1512	13	0.9964
4	0.2851	14	0.9986
5	0.4457	15	0.9995
6	0.6063	16	0.9998
7	0.7440	17	0.9999
8	0.8472	18	1.0000
9	0.9161		

Theoretically, a Poisson random variable has no upper limit. The table provides cumulative probabilities until the sum is 1.000 (using four decimal places).

The first cumulative probability is $P(X \leq 0)$, which is $P(X = 0) = p(0) = 0.0025$. This is the same value as the one we obtained manually in Example 7.10, part (a). The probability we need for part (b), is $P(X \leq 5) = 0.4457$, which is the same value we obtained manually.

To find the probability that in Example 7.10 there are six or more typos, we note that $P(X \leq 5) + P(X \geq 6) = 1$. Thus, $P(X \geq 6) = 1 - P(X \leq 5)$. Using Table 2 in Appendix B for $\mu = 6$ or **Table 7.7**, we have

$$P(X \geq 6) = 1 - P(X \leq 5) = 1 - 0.4457 = 0.5543$$

Using the Poisson table to find the Poisson probability $P(X \geq x)$

$$P(X \geq x) = 1 - P(X \leq x - 1)$$

We can also use the table to determine the probability of one individual value of X . For example, to find the probability that the book contains exactly 10 typos, we note that

$$P(X \leq 10) = p(0) + p(1) + \dots + p(9) + p(10)$$

and

$$P(X \leq 9) = p(0) + p(1) + \dots + p(9)$$

The difference between these two cumulative probabilities is $p(10)$. Thus, from **Table 7.7**,

$$\begin{aligned} P(X = 10) &= p(10) = P(X \leq 10) - P(X \leq 9) \\ &= 0.9574 - 0.9161 = 0.0413. \end{aligned}$$

Using the Poisson table to find the Poisson probability $P(X = x)$

$$P(X = x) = p(x) = P(X \leq x) - P(X \leq [x - 1])$$

7.4d Using the computer to find Poisson probabilities

We can use Excel to calculate cumulative probabilities and the probability of individual values of Poisson random variables.

COMMANDS

To calculate the probabilities associated with a Poisson random variable X with mean μ , type the following into any active cell:

=POISSON([x],[Mean],[True] or [False])

Typing 'True' calculates a cumulative probability, $P(X \leq x)$, and typing 'False' calculates the probability of an individual value of X , $P(X = x)$.

For example, to calculate the $P(X \leq 3)$ of a Poisson random variable with mean 2.5, type into an active cell, **=POISSON(3,2.5,true)**. This will calculate the probability value of 0.757576. Likewise, to calculate the $P(X = 3)$, type in **=POISSON(3,2.5,false)**. This will calculate the value of 0.213763.

There is no limit to the number of values a Poisson random variable can assume. The Poisson random variable is a discrete random variable with infinitely many possible values (i.e. $x = 0, 1, 2, \dots$), unlike the binomial random variable, which has only a finite number of possible values (i.e. $x = 0, 1, 2, \dots n$).

Mean and variance of Poisson random variables

If X is a Poisson random variable, the mean and the variance of X are

$$E(X) = \mu$$

$$V(X) = \sigma^2 = \mu$$

$$SD(X) = \sigma = \sqrt{\mu}$$

where μ is the average number of successes that occur in a specified interval.

EXERCISES

Learning the techniques

- 7.40** Let X be a Poisson random variable with $\mu = 5$.

Use Table 2 in Appendix B to find the following probabilities.

- a $P(X \leq 5)$
- b $P(X = 5)$
- c $P(X \geq 7)$

- 7.41** Suppose X is a Poisson random variable whose distribution has a mean of 2.5. Use Table 2 in Appendix B to find the following probabilities.

- a $P(X \leq 3)$
- b $P(X = 6)$
- c $P(X \geq 2)$
- d $P(X > 2)$

- 7.42** Graph the probability distribution of a Poisson random variable with $\mu = 0.5$.

Applying the techniques

- 7.43 Self-correcting exercise.** The marketing manager of a mail-order company has noted that she usually receives 10 complaint calls from customers during a week (consisting of five working days), and that the calls occur at random. Find the probability of her receiving five such calls in a single day.

- 7.44** The number of calls received by a switchboard operator between 9 a.m. and 10 a.m. has a Poisson distribution with a mean of 12. Find the probability that the operator received at least five calls during the following periods:

- a between 9 a.m. and 10 a.m.
- b between 9 a.m. and 9.30 a.m.
- c between 9 a.m. and 9.15 a.m.

7.45 The numbers of accidents that occur on an assembly line have a Poisson distribution, with an average of three accidents per week.

- a Find the probability that a particular week will be accident free.
- b Find the probability that at least three accidents will occur in a week.
- c Find the probability that exactly five accidents will occur in a week.
- d If the accidents occurring in different weeks are independent of one another, find the expected number of accidents in a year.

7.46 During the summer months (December to February, inclusive), an average of five marriages per month take place in a small city. Assuming that these marriages occur randomly and independently of

one another, find the probability of the following occurring:

- a Fewer than four marriages will occur in December.
- b At least 14 but not more than 18 marriages will occur during the entire three months of summer.
- c Exactly 10 marriages will occur during the two months of January and February.

7.47 The number of arrivals at a greengrocer's between 1 p.m. and 3 p.m. has a Poisson distribution with a mean of 14.

- a Find the probability that the number of arrivals between 1 p.m. and 3 p.m. is at least eight.
- b Find the probability that the number of arrivals between 1.30 p.m. and 2 p.m. is at least eight.
- c Find the probability that there is exactly one arrival between 2 p.m. and 3 p.m.

REAL-LIFE APPLICATIONS

Applications in operations management

Waiting lines

Everyone is familiar with waiting in lines. We wait in line at banks, grocers and fast-food restaurants. There are also waiting lines in firms where trucks wait to load and unload, and on assembly lines where stations wait for new parts. Management scientists have developed mathematical models that allow managers to determine the operating characteristics of waiting lines. Some of the operating characteristics are:

- the probability that there are no units in the system
- the average number of units in the waiting line
- the average time a unit spends in the waiting line
- the probability that an arriving unit must wait for service.

The Poisson probability distribution is used extensively in waiting line (also called queuing) models. Many models assume that the arrival of units

for service is Poisson distributed with a specific value of μ . Exercises 7.48 and 7.49 require the calculation of the probability of a number of arrivals.



Source: © Dreamstime.com/Verdelho

7.48 Cars arriving for petrol at a particular station follow a Poisson distribution with a mean of 5 per hour.

- a Determine the probability that over the next hour only one car will arrive.
- b Compute the probability that in the next 3 hours more than 20 cars will arrive.

7.49 The number of users of an ATM is Poisson distributed. The mean number of users per 5-minute interval is 1.5. Find the probability of:

- a no users in the next 5 minutes
- b 5 or fewer users in the next 15 minutes
- c 3 or more users in the next 10 minutes.

7.5 Bivariate distributions

bivariate distribution

Joint probability distribution of two variables.

Thus far, we have dealt with the distribution of a single variable. However, there are circumstances where we need to know about the relationship between two variables. Recall that we have addressed this problem statistically in Chapter 4 by drawing the scatter diagram, and in Chapter 5 by calculating the covariance and the coefficient of correlation. In this section we present the **bivariate distribution**, which provides probabilities of combinations of two variables. When we need to distinguish between the bivariate distributions and the distributions of one variable, we'll refer to the latter as univariate distributions.

The joint probability that two random variables X and Y will assume the values x and y is denoted as $p(x, y)$.

As is the case with univariate distributions, the joint probability must satisfy the following two requirements.

Requirements for a discrete bivariate distribution

1 $0 \leq p(x, y) \leq 1$ for all pairs of values (x, y)

2 $\sum_{\text{all } x} \sum_{\text{all } y} p(x, y) = 1$

EXAMPLE 7.11

LO7

Bivariate distribution of the number of house sales

Xavier and Yvette are real estate agents. Let X denote the number of houses that Xavier will sell in a month, and let Y denote the number of houses Yvette will sell in a month. An analysis of their past monthly performances has the following joint probabilities.

Bivariate probability distribution

		0	1	2
		0.12	0.42	0.06
y	0	0.21	0.06	0.03
	1	0.07	0.02	0.01

We interpret these joint probabilities in the same way we did in Chapter 6. For example, the probability that Xavier sells 0 houses and Yvette sells 1 house in the month is $p(0, 1) = 0.21$.

7.5a Marginal probabilities

As we did in Chapter 6, we can calculate the marginal probabilities by summing down columns or across rows.

The marginal probability distribution of X in Example 7.11 is

$$P(X = 0) = p(0,0) + p(0,1) + p(0,2) = 0.12 + 0.21 + 0.07 = 0.4$$

$$P(X = 1) = p(1,0) + p(1,1) + p(1,2) = 0.42 + 0.06 + 0.02 = 0.5$$

$$P(X = 2) = p(2,0) + p(2,1) + p(2,2) = 0.06 + 0.03 + 0.01 = 0.1$$

The marginal probability distribution of X is

x	0	1	2
$p(x)$	0.4	0.5	0.1

The marginal probability distribution of Y in Example 7.11 is

$$\begin{aligned} P(Y=0) &= p(0,0)+p(1,0)+p(2,0)=0.12+0.42+0.06=0.6 \\ P(Y=1) &= p(0,1)+p(1,1)+p(2,1)=0.21+0.06+0.03=0.3 \\ P(Y=2) &= p(0,2)+p(1,2)+p(2,2)=0.07+0.02+0.01=0.1 \end{aligned}$$

The marginal probability distribution of Y is

y	0	1	2
$p(y)$	0.6	0.3	0.1

Notice that both marginal probability distributions meet the requirements: the probabilities are between 0 and 1, and they sum to 1.

7.5b Describing the bivariate distribution

As we did with the univariate distribution, we often describe the bivariate distribution by computing the mean, the variance and the standard deviation of each variable. We do so by utilising the marginal probabilities.

The expected value, variance and standard deviation of X in Example 7.11 is

$$\begin{aligned} \mu_x &= E(X)=\sum xp(x)=0(0.4)+1(0.5)+2(0.1)=0.7 \\ \sigma_x^2 &= V(X)=\sum(x-\mu_x)^2 p(x) \\ &= (0-0.7)^2(0.4)+(1-0.7)^2(0.5)+(2-0.7)^2(0.1)=0.41 \\ \sigma_x &= \sqrt{V(X)}=\sqrt{0.41}=0.64 \end{aligned}$$

The expected value, variance and standard deviation of Y in Example 7.11 is

$$\begin{aligned} \mu_y &= E(Y)=\sum yp(y)=0(0.6)+1(0.3)+2(0.1)=0.5 \\ \sigma_y^2 &= V(Y)=\sum(y-\mu_y)^2 p(y) \\ &= (0-0.5)^2(0.6)+(1-0.5)^2(0.3)+(2-0.5)^2(0.1)=0.45 \\ \sigma_y &= \sqrt{V(Y)}=\sqrt{0.45}=0.67 \end{aligned}$$

There are two more parameters we can and need to compute. Both deal with the relationship between the two variables. They are the covariance and the coefficient of correlation. Recall that both were introduced in Chapter 5, in which the formulas were based on the assumption that we knew each of the N observations of the population. In this chapter we compute parameters such as the covariance and the coefficient of correlation from the bivariate distribution.

Covariance

The covariance of two discrete variables is defined as:

$$\sigma_{xy} = COV(X,Y)=\sum_{\text{all } x} \sum_{\text{all } y} (x - \mu_x)(y - \mu_y)p(x,y)$$

Notice that we multiply the deviations from the mean for both X and Y and then multiply by the joint probability.

The calculations are simplified by the following shortcut method.

Shortcut calculation for covariance

$$\sigma_{XY} = COV(X, Y) = \sum_{\text{all } x} \sum_{\text{all } y} xy p(x, y) - \mu_x \mu_y$$

The coefficient of correlation is calculated in the same way as in Chapter 5.

Coefficient of correlation

$$\rho = \frac{\sigma_{XY}}{\sigma_x \sigma_y}$$

Therefore, the covariance between two variables X and Y can be written in terms of the coefficient of correlation and the variances as

$$COV(X, Y) = \sigma_{XY} = \rho \sigma_x \sigma_y$$

EXAMPLE 7.12

LO7

Describing the bivariate distribution of the number of house sales

Calculate the covariance and the coefficient of correlation between the numbers of houses sold by the two agents in Example 7.11.

Solution

We start by computing the covariance.

$$\begin{aligned}\sigma_{XY} &= COV(X, Y) = \sum_{\text{all } x} \sum_{\text{all } y} (x - \mu_x)(y - \mu_y)p(x, y) \\ &= (0 - 0.7)(0 - 0.5)(0.12) + (1 - 0.7)(0 - 0.5)(0.42) + (2 - 0.7)(0 - 0.5)(0.06) \\ &\quad + (0 - 0.7)(1 - 0.5)(0.21) + (1 - 0.7)(1 - 0.5)(0.06) + (2 - 0.7)(1 - 0.5)(0.03) \\ &\quad + (0 - 0.7)(2 - 0.5)(0.07) + (1 - 0.7)(2 - 0.5)(0.02) + (2 - 0.7)(2 - 0.5)(0.01) \\ &= -0.15\end{aligned}$$

As we did with the shortcut method for the variance, we'll recalculate the covariance using its shortcut method.

$$\begin{aligned}\sum_{\text{all } x} \sum_{\text{all } y} xy p(x, y) &= (0)(0)(0.12) + (1)(0)(0.42) + (2)(0)(0.06) \\ &\quad + (0)(1)(0.21) + (1)(1)(0.06) + (2)(1)(0.03) \\ &\quad + (0)(2)(0.07) + (1)(2)(0.02) + (2)(2)(0.01) \\ &= 0.2\end{aligned}$$

Using the expected values computed previously, we find



$$\sigma_{XY} = \sum_{\text{all } x} \sum_{\text{all } y} xy p(x,y) - \mu_x \mu_y = 0.2 - (0.7)(0.5) = -0.15$$

We also computed the standard deviations above. Thus, the coefficient of correlation is

$$\rho = \frac{\sigma_{XY}}{\sigma_x \sigma_y} = \frac{-0.15}{(0.64)(0.67)} = -0.35$$

There is a weak negative relationship between the two variables, the number of houses Xavier will sell in a month (X) and the number of houses Yvette will sell in a month (Y).

7.5c Sum of two variables

The bivariate distribution allows us to develop the probability distribution of any combination of the two variables. Of particular interest to us is the sum of two variables. The analysis of this type of distribution leads to an important statistical application in finance, which we present in the next section.

To demonstrate how to develop the probability distribution of the sum of two variables from their bivariate distribution, return to Example 7.11. The sum of the two variables X and Y is the total number of houses sold per month. The possible values of $X + Y$ are 0, 1, 2, 3 and 4. The probability that $X + Y = 2$, for example, is obtained by summing the joint probabilities of all pairs of values of X and Y that sum to 2:

$$P(X + Y = 2) = p(0,2) + p(1,1) + p(2,0) = 0.07 + 0.06 + 0.06 = 0.19$$

We calculate the probabilities of the other values of $X + Y$ similarly, producing the following table.

Probability distribution of $X + Y$ in Example 7.11

$x + y$	0	1	2	3	4
$p(x + y)$	0.12	0.63	0.19	0.05	0.01

We can compute the expected value, the variance and the standard deviation of $X + Y$ in the usual way:

$$\mu_{X+Y} = E(X + Y) = 0(0.12) + 1(0.63) + 2(0.19) + 3(0.05) + 4(0.01) = 1.2$$

$$\begin{aligned} \sigma_{X+Y}^2 &= V(X + Y) = (0 - 1.2)^2 (0.12) + (1 - 1.2)^2 (0.63) + (2 - 1.2)^2 (0.19) \\ &\quad + (3 - 1.2)^2 (0.05) + (4 - 1.2)^2 (0.01) \\ &= 0.56 \end{aligned}$$

$$\sigma_{X+Y} = \sqrt{V(X + Y)} = \sqrt{0.56} = 0.75$$

It can be noted that, in Example 7.11, we calculated the individual expected values of X and Y as $E(X) = 0.7$ and $E(Y) = 0.5$. Now we have calculated $E(X + Y) = 1.2$, which is the sum of $E(X)$ and $E(Y)$. Similarly, we can also note that in Example 7.11 we calculated the individual variances of X and Y as $V(X) = 0.41$ and $V(Y) = 0.45$. In Example 7.12, we calculated $COV(X,Y) = -0.15$. Now we have calculated $V(X + Y) = 0.56$, which is the sum of $V(X)$ and $V(Y)$ plus twice $COV(X,Y)$.

We can derive a number of laws that enable us to compute the expected value and variance of the sum of two variables.

Laws of expected value and variance of the sum of two variables**Law of expected value of two random variables**

$$E(X + Y) = E(X) + E(Y)$$

Law of variance of two random variables

$$V(X + Y) = V(X) + V(Y) + 2COV(X, Y)$$

If X and Y are independent, $COV(X, Y) = 0$ and thus $V(X + Y) = V(X) + V(Y)$.

EXAMPLE 7.13

LO7

Describing the population of the total number of house sales

Use the rules of expected value and variance of the sum of two variables to calculate the mean and variance of the total number of houses sold per month in Example 7.11.

Solution

Using the law of expected value, we compute the expected value of $X + Y$:

$$E(X + Y) = E(X) + E(Y) = 0.7 + 0.5 = 1.2$$

which is the same value we produced directly from the probability distribution of $X + Y$.

We apply the law of variance to determine the variance of $X + Y$:

$$V(X + Y) = V(X) + V(Y) + 2COV(X, Y) = 0.41 + 0.45 + 2(-0.15) = 0.56$$

This is the same value we obtained from the probability distribution of $X + Y$.

We will encounter several applications where we will need the laws of expected value and variance for the sum of two variables. Additionally, we will demonstrate an important application in operations management where we need the formulas for the expected value and variance of the sum of more than two variables. See Exercises 7.35–7.38 (on page 284).

EXERCISES

- 7.50** The table lists the bivariate distribution of X and Y .

$y \backslash x$	1	2
1	0.5	0.1
2	0.1	0.3

- a Find the marginal probability distribution of X .
- b Find the marginal probability distribution of Y .
- c Compute the mean and variance of X .
- d Compute the mean and variance of Y .
- e Compute the covariance and the coefficient of correlation of X and Y .

- 7.51** Refer to Exercise 7.50. Use the laws of expected value and variance of the sum of two variables to compute the mean and variance of $X + Y$.

- 7.52** Refer to Exercise 7.50. Consider the random variables $X + Y$.

- a Determine the distribution of $X + Y$.
- b Determine the mean and variance of $X + Y$.
- c Does your answer to part (b) equal the answer to Exercise 7.51?

- 7.53** The joint probability distribution of X and Y is shown in the table below.

$y \backslash x$	1	2	3
1	0.42	0.12	0.06
2	0.28	0.08	0.04

- a Determine the marginal distributions of X and Y .
- b Compute the covariance and coefficient of correlation between X and Y .
- c Develop the probability distribution of $X + Y$.

- 7.54** The following distributions of X and of Y have been developed. If X and Y are independent, determine the joint probability distribution of X and Y .

x	0	1	2
$p(x)$	0.6	0.3	0.1
y	1	2	
$p(y)$	0.7	0.3	

- 7.55** After analysing several months of sales data, the owner of an appliance store produced the following joint probability distribution of the number of refrigerators (x) and stoves (y) sold daily.

Stoves (y)	Refrigerators (x)		
	0	1	2
0	0.08	0.14	0.12
1	0.09	0.17	0.13
2	0.05	0.18	0.04

- a** Find the marginal probability distribution of the number of refrigerators sold daily.
- b** Find the marginal probability distribution of the number of stoves sold daily.
- c** Compute the mean and variance of the number of refrigerators sold daily.
- d** Compute the mean and variance of the number of stoves sold daily.
- e** Compute the covariance and the coefficient of correlation.

- 7.56** Refer to Exercise 7.55. Find the following conditional probabilities.

- a** $P(1 \text{ refrigerator}|0 \text{ stoves})$
- b** $P(0 \text{ stoves}|1 \text{ refrigerator})$
- c** $P(2 \text{ refrigerators}|2 \text{ stoves})$

REAL-LIFE APPLICATION

PERT/CPM

PERT (Project Evaluation and Review Technique) and CPM (Critical Path Method) are related management science techniques that help operations managers control the activities and the amount of time it takes to complete a project. Both techniques are based on the order in which the activities must be performed. For example, in building a house, the excavation of the foundation must precede the pouring of the foundation, which in turn precedes the framing.

A *path* is defined as a sequence of related activities that leads from the starting point to the completion of a project. In most projects there are several paths with differing amounts of time needed for their completion. The longest path is called the *critical path* because any delay in the activities along this path will result in a delay in the completion of the project.

In some versions of PERT/CPM, the activity completion times are fixed and the chief task of the operations manager is to determine the critical path. In other versions, each activity's completion time is considered to be a random variable, where the mean and variance can be estimated. By extending the laws of expected value and variance for the sum of two variables to more than two variables, we produce the



Source: Shutterstock.com/Pressmaster

following, where X_1, X_2, \dots, X_k are the times for the completion of activities 1, 2, ..., k respectively. These times are independent random variables.

Laws of expected value and variance for the sum of k independent variables where $k \geq 2$

- 1** $E(X_1 + X_2 + \dots + X_k) = E(X_1) + E(X_2) + \dots + E(X_k)$
- 2** $V(X_1 + X_2 + \dots + X_k) = V(X_1) + V(X_2) + \dots + V(X_k)$

Using these laws we can then produce the expected value and variance for the complete project. Exercises 7.57–7.60 address this problem.

- 7.57** There are four activities along the critical path for a project. The expected values and variances of the completion times of the activities are listed here. Determine the expected value and variance of the completion time of the project.

Activity	Expected completion time (days)	Variance
1	18	8
2	12	5
3	27	6
4	8	2

- 7.58** The operations manager of a large plant wishes to overhaul a machine. After conducting a PERT/CPM analysis, he has developed the following critical path:

- 1 Disassemble machine
- 2 Determine parts that need replacing
- 3 Find needed parts in inventory
- 4 Reassemble machine
- 5 Test machine

He has estimated the mean (in minutes) and variances of the completion times as follows:

Activity	Mean	Variance
1	35	8
2	20	5
3	20	4
4	50	12
5	20	2

Determine the mean and standard deviation of the completion time of the project.

- 7.59** In preparing to launch a new product, a marketing manager has determined the critical path for her department. The activities and the mean and variance of the completion time (in days) for each

activity along the critical path are shown in the accompanying table. Determine the mean and variance of the completion time of the project.

Activity	Completion time	
	Mean	Variance
Develop survey questionnaire	8	2
Pre-test the questionnaire	14	5
Revise the questionnaire	5	1
Hire survey company	3	1
Conduct survey	30	8
Analyse data	30	10
Prepare report	10	3

- 7.60** A professor of business statistics is about to begin work on a new research project. Because time is quite limited, a PERT/CPM critical path has been developed, which consists of the following activities:
- 1 Conduct a search for relevant research articles
 - 2 Write a proposal for a research grant
 - 3 Perform the analysis
 - 4 Write the article and send to journal
 - 5 Wait for reviews
 - 6 Revise on the basis of the reviews and resubmit

The mean (in days) and variance (in days²) of the completion times are as follows:

Activity	Mean	Variance
1	10	9
2	3	0
3	30	100
4	5	1
5	100	400
6	20	64

Compute the mean and standard deviation of the completion time of the entire project.

7.6 Applications in finance: Portfolio diversification and asset allocation

In this section we introduce an important application in finance that is based on the previous section.

In Chapter 5, we described how the variance or the standard deviation can be used to measure the risk associated with an investment. Most investors tend to be risk averse, which means that they prefer to have lower risk associated with their investments. One of the ways in which financial analysts lower the risk that is associated with the share market is through

diversification. This strategy was first mathematically developed by Harry Markowitz in 1952. His model paved the way for the development of modern portfolio theory, which is the concept underlying mutual funds.

7.6a Portfolios with two shares

To illustrate the basics of portfolio diversification, consider an investor who forms a portfolio consisting of only two shares by investing \$4000 in one share and \$6000 in a second share. Suppose that the results after one year are as listed below. (For the definition of return on investment see ‘Real-life applications: Return on investment’ on page 97).

One-year results

Share	Initial investment	Value of investment after one year	Rate of return on investment
1	\$4000	\$5000	$R_1 = 0.25$ (25%)
2	\$6000	\$5400	$R_2 = -0.10$ (-10%)
Total	\$10000	\$10400	$R_p = 0.04$ (4%)

Another way of calculating the portfolio return R_p is to calculate the weighted average of the individual share returns R_1 and R_2 , where the weights w_1 and w_2 are the proportions of the initial \$10000 invested in shares 1 and 2 respectively. In this illustration $w_1 = 0.4$ and $w_2 = 0.6$. (Note that w_1 and w_2 must always sum to 1 because the two shares constitute the entire portfolio.) The weighted average of the two returns is

$$\begin{aligned} R_p &= w_1 R_1 + w_2 R_2 \\ &= (0.4)(0.25) + (0.6)(-0.10) = 0.04 \end{aligned}$$

This is how portfolio returns are calculated. However, when the initial investments are made, the investor does not know what the returns will be. In fact, the returns are random variables. We are interested in determining the expected value and variance of the portfolio. The formulas below were derived from the laws of expected value and variance introduced in the two previous sections.

Mean and variance of a portfolio of two shares

$$\begin{aligned} E(R_p) &= w_1 E(R_1) + w_2 E(R_2) \\ V(R_p) &= w_1^2 V(R_1) + w_2^2 V(R_2) + 2w_1 w_2 \text{COV}(R_1, R_2) \\ &= w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + 2w_1 w_2 \rho \sigma_1 \sigma_2 \end{aligned}$$

where w_1 and w_2 are the proportions or weights of shares 1 and 2, $E(R_1)$ and $E(R_2)$ are their expected values, σ_1 and σ_2 are their standard deviations, and ρ is the coefficient of correlation (recall that $\text{COV}(R_1, R_2) = \rho \sigma_1 \sigma_2$).

EXAMPLE 7.14

LO3

Describing the population of the returns on a portfolio

An investor has decided to form a portfolio by putting 25% of his money into McDonald's shares and 75% into Cisco Systems shares. The investor assumes that the expected returns will be 8% and 15% respectively, and that the standard deviations will be 12% and 22% respectively.

- a** Find the expected return on the portfolio.
- b** Calculate the standard deviation of the returns on the portfolio assuming that:
 - i** returns on the two shares are perfectly positively correlated
 - ii** the coefficient of correlation is 0.5
 - iii** returns on the two shares are uncorrelated.

Solution

- a** The expected values of the two shares are

$$E(R_1) = 0.08 \text{ and } E(R_2) = 0.15$$

The weights are $w_1 = 0.25$ and $w_2 = 0.75$. Thus,

$$E(R_p) = w_1 E(R_1) + w_2 E(R_2) = 0.25(0.08) + 0.75(0.15) = 0.1325$$

- b** The standard deviations are

$$\sigma_1 = 0.12 \text{ and } \sigma_2 = 0.22$$

Thus,

$$\begin{aligned} V(R_p) &= w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + 2w_1 w_2 \rho \sigma_1 \sigma_2 \\ &= (0.25^2)(0.12^2) + (0.75^2)(0.22^2) + 2(0.25)(0.75)\rho(0.12)(0.22) \\ &= 0.0281 + 0.0099\rho \end{aligned}$$

- i** When returns on the two shares are perfectly positively correlated, i.e. $\rho = 1$:

$$V(R_p) = 0.0281 + 0.0099(1) = 0.0380$$

$$\text{Standard deviation} = SD(R_p) = \sqrt{V(R_p)} = \sqrt{0.0380} = 0.1949$$

- ii** When $\rho = 0.5$:

$$V(R_p) = 0.0281 + 0.0099(0.5) = 0.0331$$

$$\text{Standard deviation} = SD(R_p) = \sqrt{V(R_p)} = \sqrt{0.0331} = 0.1819$$

- iii** When returns on the two shares are uncorrelated, i.e. $\rho = 0$:

$$V(R_p) = 0.0281 + 0.0099(0) = 0.0281$$

$$\text{Standard deviation} = SD(R_p) = \sqrt{V(R_p)} = \sqrt{0.0281} = 0.1676$$

Notice that the variance of the portfolio returns decreases as the coefficient of correlation decreases.

7.6b Portfolio diversification in practice

The formulas introduced in this section require that we know the expected values, variances and covariance (or coefficient of correlation) of the investments we are interested in. The question arises: 'How do we determine these parameters?' (Incidentally, this question is rarely addressed in finance textbooks!) The most common procedure is to estimate the parameters from historical data, using sample statistics.

7.6c Portfolios with more than two shares

We can extend the formulas that describe the mean and variance of the returns of a portfolio of two shares to a portfolio of any number of shares.

Mean and variance of a portfolio of k shares

$$E(R_p) = \sum_{i=1}^k w_i E(R_i)$$

$$V(R_p) = \sum_{i=1}^k w_i^2 \sigma_i^2 + 2 \sum_{i=1}^k \sum_{j=i+1}^k w_i w_j COV(R_i, R_j)$$

where R_i is the return of the i th share, w_i is the proportion of the portfolio invested in share i , and k is the number of shares in the portfolio.

When k is greater than 2, the calculations can be tedious and time consuming. For example, when $k = 3$, we need to know the values of three weights, three expected values, three variances and three covariances. When $k = 4$, there are four expected values, four variances and six covariances. (The number of covariances required in general is $k(k - 1)/2$.) To assist you, we have created an Excel workbook, **Portfolio allocation.xlsx** that is provided on the companion website (accessible through <http://login.cengagebrain.com>), to perform the calculations when $k = 2, 3$ or 4 . To demonstrate, we will return to the problem described in this chapter's introduction.

SPOTLIGHT ON STATISTICS

Where to invest the super?: Solution

Because of the large number of calculations, we will solve this problem using only Excel. From the file we calculate the means of the returns for each share.



Source: iStock.com/Faberrin

Excel means

	A	B	C	D	E
1		Residential	Banking	Commercial	Resources
2	Mean	0.00031	0.00234	0.00791	0.00800

Next we calculate the sample variance–covariance matrix. (The commands are the same as those described in Chapter 5 (see page 185) – simply include all the columns of the returns of the investments you wish to include in the portfolio and then convert the population variances and covariances to sample values by multiplying the values by $73/72$ ($n/n - 1$).

Excel variance–covariance matrix (sample)

	A	B	C	D	E
1		Residential	Banking	Commercial	Resources
2	Residential	0.00354			
3	Banking	0.00109	0.001117		
4	Commercial	0.00155	0.00202	0.00549	
5	Resources	0.00095	0.00419	0.00255	0.01042

Notice that the variances of the returns are listed on the diagonal. Thus, for example, the variance of the 72 monthly returns of Residential is 0.00354. The covariances appear below the diagonal. The covariance between the returns of Residential and Commercials is 0.00155.

The means and the variance–covariance matrix are copied to the spreadsheet using the commands described below. The weights are typed, producing the output below.

Excel worksheet: Portfolio diversification – Portfolio 1

	A	B	C	D	E	F
1	Portfolio of 4 shares					
2			Residential	Banking	Commercial	Resources
3	Variance–covariance	Residential	0.00354			
4		Banking	0.00109	0.01117		
5		Commercial	0.00155	0.00202	0.00549	
6		Resources	0.00095	0.004131	0.00255	0.01042
7						
8	Expected returns		0.00031	0.00234	0.00791	0.00800
9	Weights		0.25	0.25	0.25	0.25
10						
11						
12	Portfolio return					
13	Expected value	0.0046				
14	Variance	0.0035				
15	Standard deviation	0.0588				

The expected return on the portfolio is 0.0046 and the variance is 0.0035.

COMMANDS

- 1 Open the file containing the returns (**XMO7-00**).
- 2 Calculate the means of the columns containing the returns of the shares in the portfolio.
- 3 Using the commands described in Chapter 5 (page 185), calculate the (population) variance–covariance matrix. Convert the population values to sample values by multiplying each value by $(n/n - 1)$.
- 4 Open the **Portfolio Diversification** workbook. Use the tab to select the **4 Shares** worksheet. DO NOT CHANGE ANY CELLS THAT APPEAR IN BOLD PRINT. DO NOT SAVE ANY WORKSHEETS.
- 5 Copy the means into cells C7 to E7. (Use **Copy**, **Paste Special** with **Values** and **Number** formats.)
- 6 Copy the variance–covariance matrix (including row and column labels) into columns C, D and E.
- 7 Type the weights into cells C9 to E9.

The mean, variance and standard deviation of the portfolio will be printed. Use similar commands for 2-share and 3-share portfolios.

The results for portfolio 2 are

	A	B
1	Portfolio return	
2	Expected value	0.0049
3	Variance	0.0032
4	Standard deviation	0.0561

The results for portfolio 3 are

	A	B
1	Portfolio return	
2	Expected value	0.0044
3	Variance	0.0048
4	Standard deviation	0.0695

Portfolio 2 is better than portfolios 1 and 3 because its expected value is larger and its variance is smaller. To minimise risk and maximise return, the investor should choose portfolio 2.

In this example we showed how to calculate the expected return, variance and standard deviation from a sample of returns on the investments for any combination of weights. (We illustrated the process with three sets of weights.) It is possible to determine the ‘optimal’ weights that minimise risk for a given expected value or maximise expected return for a given standard deviation. This is an extremely important function of financial analysts and investment advisers. Solutions can be determined using a management science technique called *linear programming*, a subject taught by most schools of business and faculties of management.

EXERCISES

Exercises 7.61–7.71 can be solved using Excel’s **Portfolio allocation.xlsx** workbook, which can be obtained from the companion website.

- 7.61** A portfolio is composed of two shares. The following parameters associated with the returns of the two shares are given.

Share	1	2
Proportion of portfolio	0.30	0.70
Mean	0.12	0.25
Standard deviation	0.02	0.15

For each of the following values of the coefficient of correlation, determine the mean and the standard deviation of the return on the portfolio.

- a $\rho = 0.5$
- b $\rho = 0.2$
- c $\rho = 0.0$

- 7.62** Describe what happens to the expected value and standard deviation of the portfolio returns when the coefficient of correlation decreases.

- 7.63** An investor is given the following information about the returns on two shares:

Share	1	2
Mean	0.09	0.13
Standard deviation	0.15	0.21

- a If she is most interested in maximising her returns, which share should she choose?

- b If she is most interested in minimising her risk, which share should she choose?

- 7.64** Refer to Exercise 7.63. Calculate the expected value and variance of the portfolio composed of 60% share 1 and 40% share 2. The coefficient of correlation is 0.4.

- 7.65** Refer to Exercise 7.63. Calculate the expected value and variance of the portfolio composed of 30% share 1 and 70% share 2. The coefficient of correlation is 0.4.

- 7.66 XR07-64** A financial analyst recorded the quarterly returns on investment for three shares.

- a Calculate the mean and variance of each of the three shares.
- b If you wish to construct a portfolio that maximises the expected return, what should you do?
- c If you wish to construct a portfolio that minimises the risk, what should you do?

- 7.67** Refer to Exercise 7.66.

- a Find the expected value and variance of the following portfolio:
Share 1: 30%
Share 2: 40%
Share 3: 30%
- b How do the expected value and variance of the portfolio compare with those of Exercise 7.66 part (a)?

7.68 Refer to Exercise 7.66.

- a Find the expected value and variance of the following portfolio:
Share 1: 10%
Share 2: 10%
Share 3: 80%
- b How do the expected value and variance of the portfolio compare with those of Exercise 7.66 part (a) and Exercise 7.67?

7.69 XR07-69 The quarterly rates of return for four shares are recorded.

- a Calculate the mean and variance of each of the four shares.
- b If you wish to construct a portfolio that maximises the expected return, what should you do?
- c If you wish to construct a portfolio that minimises the risk, what should you do?

7.70 Refer to Exercise 7.69.

- a Find the expected value and variance of the following portfolio:
Share 1: 25%
Share 2: 25%
Share 3: 25%
Share 4: 25%
- b How do the expected value and variance of the portfolio compare with those of Exercise 7.69 part (a)?

7.71 Refer to Exercise 7.69.

- a Find the expected value and variance of the following portfolio:
Share 1: 20%
Share 2: 20%
Share 3: 10%
Share 4: 50%
- b How do the expected value and variance of the portfolio compare with those of Exercise 7.67 part (a) and Exercise 7.68?

Study Tools

CHAPTER SUMMARY

The concept of a random variable permits us to summarise the results of an experiment in terms of numerically valued events. Specifically, a random variable assigns a numerical value to each simple event of an experiment. There are two types of random variables. A discrete random variable is one whose values are countable. A continuous random variable can assume an uncountable number of values. In this chapter we discussed discrete random variables and their probability distributions.

We defined the *expected value*, *variance* and *standard deviation* of a population represented by a discrete probability distribution. We also presented two most important discrete distributions: the *binomial* and the *Poisson*. Finally, in this chapter we also introduced *bivariate discrete distributions* for which an important application in finance was discussed.

The data files for Examples, Exercises and Cases are provided on the companion website (accessible through <https://login.cengagebrain.com/>).

SYMBOL

Symbol	Pronounced	Represents
$\sum_{\text{all } x} x$	Sum of x for all values of x	Summation
C_x^n	n -choose- x	Number of combinations
$n!$	n -factorial	$n(n - 1)(n - 2) \dots (3)(2)(1)$
e	exponential	2.718...

SUMMARY OF FORMULAS

Expected value (mean)	$E(X) = \mu = \sum_{\text{all } x} xp(x)$
Variance	$V(X) = \sigma^2 = \sum_{\text{all } x} (x - \mu)^2 p(x) = \sum_{\text{all } x} x^2 p(x) - \mu^2$
Standard deviation	$SD(X) = \sigma = \sqrt{V(X)}$
Covariance	$\begin{aligned} COV(X, Y) &= \sigma_{XY} = \sum_{\text{all } x} (x - \mu_X)(y - \mu_Y)p(x, y) \\ &= \sum_{\text{all } x} \sum_{\text{all } y} xy p(x, y) - \mu_X \mu_Y \end{aligned}$
Coefficient of correlation	$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$
Laws of expected value and variance	1 $E(cX + d) = cE(X) + d$ 2 $V(cX + d) = c^2 V(X)$
Binomial probability	$p(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \text{ for } x = 0, 1, 2, \dots, n$ $E(X) = \mu = np$ $V(X) = \sigma^2 = npq, \quad q = 1 - p$ $SD(X) = \sigma = \sqrt{npq}$

Poisson probability	$P(X = x) = \frac{e^{-\mu} \mu^x}{x!}, x = 0, 1, 2, \dots$ $E(X) = \mu$ $V(X) = \sigma^2 = \mu$ $SD(X) = \sigma = \sqrt{\mu}$
Laws of expected value and variance of the sum of two variables	1 $E(X + Y) = E(X) + E(Y)$ 2 $V(X + Y) = V(X) + V(Y) + 2COV(X, Y) = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}$
Laws of expected value and variance for the sum of k variables, where $k \geq 2$ if the variables are independent	1 $E(X_1 + X_2 + \dots + X_k) = E(X_1) + E(X_2) + \dots + E(X_k)$ 2 $V(X_1 + X_2 + \dots + X_k) = V(X_1) + V(X_2) + \dots + V(X_k)$
Mean and variance of a portfolio of two shares: $R_p = w_1 R_1 + w_2 R_2$	1 $E(R_p) = w_1 E(R_1) + w_2 E(R_2)$ 2 $V(R_p) = w_1^2 V(R_1) + w_2^2 V(R_2) + 2w_1 w_2 COV(R_1, R_2)$ $= w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + 2w_1 w_2 \rho \sigma_1 \sigma_2$
Mean and variance of a portfolio of k shares: $R_p = \sum_{i=1}^k w_i R_i$	1 $E(R_p) = \sum_{i=1}^k w_i E(R_i)$ 2 $V(R_p) = \sum_{i=1}^k w_i^2 \sigma_i^2 + 2 \sum_{i=1}^k \sum_{j=i+1}^k w_i w_j COV(R_i, R_j)$

SUPPLEMENTARY EXERCISES

- 7.72** Let X represent the household size of Aboriginal and Torres Strait Islander households in Northern Territory (NT). The frequency distribution of X based on the 2016 census is given below:

x	1	2	3	4	5	6 or more
No. of families	1545	2506	1987	1913	1427	3371

Source: Australian Bureau of Statistics, *Census of Population and Housing 2016*, ABS, Canberra

- a Construct the probability distribution of X and present it in graphical form.
- b Comment on the symmetry or skewness of the distribution.
- c What is the most likely number of persons in an NT Aboriginal and Torres Strait Islander household?
- d If an Aboriginal and Torres Strait Islander household in NT is selected at random, what is the probability that the household has more than three people?

- 7.73** The number of magazine subscriptions per household (X) is represented by the following probability distribution.

x	0	1	2	3	4
$p(x)$	0.48	0.35	0.08	0.05	0.04

- a Calculate the mean number of magazine subscriptions per household.
- b Find the standard deviation.

- 7.74** A maintenance worker in a large paper-manufacturing plant knows that, on average, the main pulper (which beats solid materials to a pulp) breaks down six times per 30-day month. Find the probability that, on a given day, the pulper will have to be repaired:
- a exactly once
 - b at least once
 - c at least once, but not more than twice.

- 7.75** A study of drivers reveals that, when lost, 45% will stop and ask for directions, 30% will consult a map and 25% will continue driving until the location has been determined. Suppose that a sample of 200

drivers was asked to report what they do when lost. Find the following probabilities.

- a At least 100 will stop and ask directions.
- b At most 55 will continue driving.
- c Between 50 and 75 (inclusive) will consult a map.

7.76 The scheduling manager for an electricity supply company knows that there is an average of 12 emergency calls regarding power failures per month. Assume that a month consists of 30 days.

- a Find the probability that the company will receive at least 12 emergency calls during a specified month.
- b Suppose that the company can handle a maximum of three emergency calls per day. What is the probability that there will be more emergency calls than the company can handle on a given day?

7.77 Lotteries are an important income source for various governments around the world. However, the availability of lotteries and other forms of gambling have created a social problem – gambling addicts. A critic of government-controlled gambling contends that 30% of people who regularly buy lottery tickets are gambling addicts. If we randomly select 10 people from those who report that they regularly buy lottery tickets, what is the probability that more than five of them are addicts?

7.78 A pharmaceutical researcher working on a cure for baldness noticed that middle-aged men who are balding at the crown of their head have a 45% probability of suffering a heart attack over the next decade. In a sample of 100 middle-aged balding men, what are the following probabilities?

- a More than 50 will suffer a heart attack in the next decade.
- b Fewer than 44 will suffer a heart attack in the next decade.
- c Exactly 45 will suffer a heart attack in the next decade.

7.79 An airline boasts that 77.4% of its flights were on time. If we select five flights at random, what is the probability that all five are on time?

7.80 Many mobile phone service providers offer family plans wherein parents who subscribe can get discounts for other family members. Suppose that the number of mobile phones per family is Poisson distributed with a mean of 1.5. If one family is randomly selected, calculate the following probabilities.

- a Family has only 1 mobile phone.
- b Family has 3 or more mobile phones.
- c Family has 4 or fewer mobile phones.

7.81 A newly arrived business migrant intends to place one-quarter of his funds in a real-estate venture and the remaining three-quarters in a portfolio of shares. The real-estate venture has an expected return of 28% with a standard deviation of 20%, and the share portfolio has an expected return of 12% with a standard deviation of 6%. Assume that the returns on these two investments are independent.

- a What are the expected value and the standard deviation of the return on the total funds invested? (*Hint:* Let X be the return on the real-estate venture, let Y be the return on the share portfolio, and express the return on the total funds invested as a function of X and Y .)
- b Using the variance of the possible returns on an investment as a measure of its relative riskiness, rank the real-estate venture, the share portfolio, and the combination of the two, in order of increasing riskiness.

7.82 After watching several seasons of soccer a statistician produced the following bivariate probability distribution of scores.

Visiting team	Home team			
	0	1	2	3
0	0.14	0.11	0.09	0.10
1	0.12	0.10	0.05	0.02
2	0.09	0.07	0.04	0.01
3	0.03	0.02	0.01	0.00

- a What is the probability that the home team wins?
- b What is the probability of a tie?
- c What is the probability that the visiting team wins?

Case Studies

CASE 7.1 Has there been a shift in the location of overseas-born population within Australia over the 50 years from 1996 to 2016?

C07-02 In 1966, Australia's overseas-born population was only 18% of the total Australian population. By 2016, 50 years later, this proportion had increased to 26%. The table below gives the breakdown of the states and territories where overseas-born Australian population lived in 1966 and 2016. Analyse the data to identify any changes between 1966 and 2016 in the distribution of the overseas-born population living in the various Australian states and territories.

Overseas-born Australian population of each state and territory, 1966 and 2016

State or Territory	Number of persons	
	1966	2016
New South Wales	733804	2072454
Victoria	680549	1680256
Queensland	201831	1015875
South Australia	245935	384097
Western Australia	198764	797695
Tasmania	35840	61240
Northern Territory	8416	45403
Australian Capital Territory	25416	105161
Australia	2130555	6162181

Source: Census of Population and Housing, 1966 and 2016, Australian Bureau of Statistics, 2016, Canberra, Australia.

CASE 7.2 How about a carbon tax on motor vehicle ownership?

An environmental specialist proposed that one way of reducing carbon emissions into our environment is by introducing a new carbon tax on motor vehicle ownership. He proposes the following taxation system based on the number of motor vehicles (excluding electric vehicles) per dwelling.

Number of motor vehicles per dwelling	Proposed carbon tax per year (\$)
0	0
1	200
2	400
3 or more	1000

One of the government officials in the revenue department is tempted by this proposal, as it would generate significant revenue for the government. He asks his research officer to collect data on motor vehicle ownership per dwelling to work out the expected tax revenue from this proposal and is willing to consider it seriously if the total tax revenue exceeds \$100 million.

According to the Australian Bureau of Statistics' *2016 Census of Population and Housing*, the following are the data on motor vehicles ownership by dwelling. The total number of dwellings in Australia is 9924 992. Will the official consider the environmental specialist's proposal?

Number of motor vehicles	Number of dwellings
0	643304
1	2945022
2	3025454
3 or more	1505374
Total	8 119 154

Source: Australian Bureau of Statistics,
2016 Census of Population and Housing, 2016, ABS, Canberra.

CASE 7.3 How about a carbon tax on motor vehicle ownership? – New Zealand

Consider Case 7.2. The Australian government official in the revenue department would also like to know how much revenue a smaller country such as New Zealand would gain if the same taxation system was imposed for New Zealand households. He asks his research officer to collect data on motor vehicle ownership per dwelling to work out the expected tax revenue from this proposal for New Zealand. According to the New Zealand 2006, 2013 and 2018 Census data, the following are the data on motor vehicles ownership (excluding electric vehicles) by household. Obtain the expected tax revenue from the environmental specialist's proposal for New Zealand for the three years.

Number of motor vehicles	Number of households		
	2006	2013	2018
No motor vehicle	112 758	116 379	100 302
One motor vehicle	527 844	552 813	514 992
Two motor vehicles	531 624	565 095	598 062
Three or more motor vehicles	222 201	237 471	303 942
Total households stated	1394 430	1471 758	1517 298

Source: Motor vehicles per household
in New Zealand, 2006, 2013 and 2018 Censuses, <http://nzdotstat.stats.govt.nz/wbos/>, NZ Stats, New Zealand.

CASE 7.4 Internet usage by children

C07-04 Internet usage has increased astronomically in the past decade or so. Increasingly children are becoming very active users for various purposes, including education, gaming and communicating with family and friends. The following table presents the data for the number of hours of internet usage by various age groups of Australian children. Analyse the patterns of internet usage of the various age groups of children.

Number of hours of internet use per week	Number of children ('000) by age group			
	5–8 years	9–11 years	12–14 years	All
2 hours or less	456.9	269.6	125.1	851.6
3–4 hours	135.2	136.3	97.3	368.8
5–9 hours	111.4	168.1	192.5	472.0
10–19 hours	39.3	99.3	230.5	369.1
20 hours or more	8.2	27.9	97.6	133.7
Total	751.0	701.2	743.0	2195.2

Source: Australian Bureau of Statistics, *Children's Internet Usage, Australia*, Table 19, October 2012, cat. no. 4901.0, ABS, Canberra

CASE 7.5 COVID-19 deaths in Australia by age and gender III

C07-05 Consider again Case 3.1, in which data on the number of COVID-19 related deaths as at 30 June 2020, by age group, state and gender are recorded and presented in the following table.

- a Develop a probability distribution of the COVID-19 deaths across the states by age group in Australia.
- b Also develop a probability distribution of the COVID-19 deaths by age group and gender.

Number of COVID-19 related deaths by states, gender and age group, Australia, 30 June 2020

Age group	State/Territory								Gender		
	NSW	VIC	QLD	WA	SA	TAS	ACT	NT	Australia	Male	Female
40-49 years				1					1	1	
50-59 years	1	1							2	1	1
60-69 years	4	3	1		1		1		10	5	5
70-79 years	11	6	3	6	3	5			34	22	12
80-89 years	16	9	2	2		5	2		36	19	17
90-99 years	16	1				3			20	8	12
Total	48	20	6	9	4	13	3	0	103	56	47

Source: covid19data.com.au. CC BY 4.0 <https://creativecommons.org/licenses/by/4.0/>

Continuous probability distributions

Learning objectives

This chapter introduces continuous probability distributions such as the normal distribution.

At the completion of this chapter, you should be able to:

- L01** discuss the basic differences between discrete and continuous random variables
- L02** calculate probabilities using a uniform distribution
- L03** understand how to use the standard normal probability table and calculate probabilities from a normal distribution
- L04** calculate probabilities using an exponential distribution
- L05** approximate binomial probabilities using a normal distribution.

CHAPTER OUTLINE

Introduction

8.1 Probability density functions

8.2 Uniform distribution

8.3 Normal distribution

8.4 Exponential distribution

SPOTLIGHT ON STATISTICS

Would the pizza business survive next year?

While every business aims to make a profit, the first milestone for any business is to achieve the breakeven situation.

A well-educated businessman who wants to buy an existing pizza business in a Melbourne outer suburb wants to model the total pizza sales by looking at the past sales records of the business to make sure that his new pizza business will survive. He analyses the pizza business' past total sales, he finds that total sales are normally distributed with a mean of \$3m and standard deviation of \$0.5m. In order to cover the costs, total sales for the year should exceed the breakeven level of \$2m.

The businessman will purchase the business only if he can be confident that it is highly likely that the business would survive next year. Will the business survive next year?

He would also like to know the sales level, which has only a 9% likelihood of being exceeded next year.

After introducing the normal distribution, we will provide an answer to this question (see pages 331–2).



Source: iStock.com/RaStudio

Introduction

This chapter completes our presentation of probability by introducing continuous random variables and their distributions. In Chapter 7 we introduced discrete probability distributions that are used to calculate the probability associated with discrete random variables. In Section 7.6 we introduced the binomial distribution, which allows us to determine the probability that the random variable equals a particular value (the number of successes). In this way we connected the population represented by the probability distribution with a sample of nominal data. In this chapter we introduce continuous probability distributions, which are used to calculate the probability associated with numerical (quantitative) variables. By doing so, we develop the link between a population and a sample of numerical data.

Section 8.1 introduces probability density functions and Section 8.2 uses the uniform density function to demonstrate how probability is calculated. In Section 8.3 we focus on the normal distribution, one of the most important distributions because of its role in the development of statistical inference. Section 8.4 introduces the exponential distribution, a distribution that has proven to be useful in various management science applications. Finally, we present a summary of the chapter.

Three more continuous distributions, which are used extensively in statistical inference, will be introduced in later chapters: *t distribution* (Chapter 10), *chi-squared distribution* (Appendix 14.A) and *F distribution* (Appendix 16.A).

8.1 Probability density functions

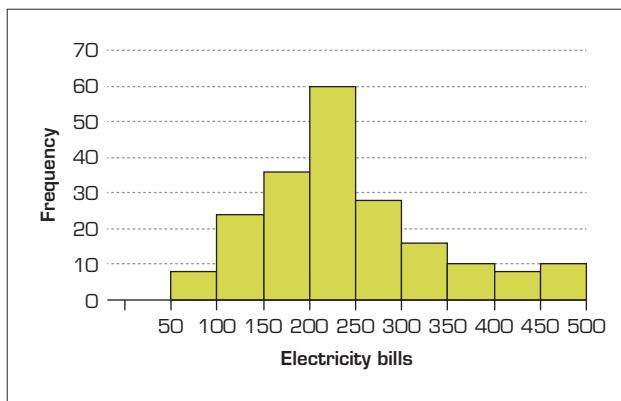
continuous random variable

A random variable that can assume an uncountable number of values in an interval.

Up to this point, we have focused our attention on discrete distributions – distributions of random variables (X) that have either a finite number of possible values (e.g. $x = 0, 1, 2, \dots, n$) or a countably infinite number of values ($x = 0, 1, 2, \dots$). In contrast, as discussed in Chapter 7, a **continuous random variable** has an uncountably infinite number of possible values and can assume any value in the interval between two points a and b ($a < x < b$). Whereas discrete random variables typically involve counting, continuous random variables typically involve measurement attributes such as length, weight, time and temperature.

A continuous random variable is one that can assume an uncountable number of values. Because this type of random variable is so different from a discrete random variable, we need to treat it completely differently. First, we cannot list the possible values because there is an infinite number of them. Second, because there is an infinite number of values, the probability of each individual value is virtually 0. Consequently, we can determine the probability of a range of values only. To illustrate how this is done, consider the histogram we created for the monthly electricity bills (Example 4.1), which is depicted in **Figure 8.1**.

FIGURE 8.1 Histogram for Example 4.1



Relative frequencies for Example 4.1

Class limits	Relative frequency
50 up to 100	8/200 = 0.04
100 up to 150	24/200 = 0.12
150 up to 200	36/200 = 0.18
200 up to 250	60/200 = 0.30
250 up to 300	28/200 = 0.14
300 up to 350	16/200 = 0.08
350 up to 400	10/200 = 0.05
400 up to 450	8/200 = 0.04
450 up to 500	10/200 = 0.05
	Total = 1.00

We found, for example, that the relative frequency of the interval 50 to 100 was 8/200. Using the relative frequency approach, we estimate that the probability that a randomly selected electricity bill will fall between \$50 and \$100 is $8/200 = 0.04$. We can similarly estimate the probabilities of the other intervals in the histogram.

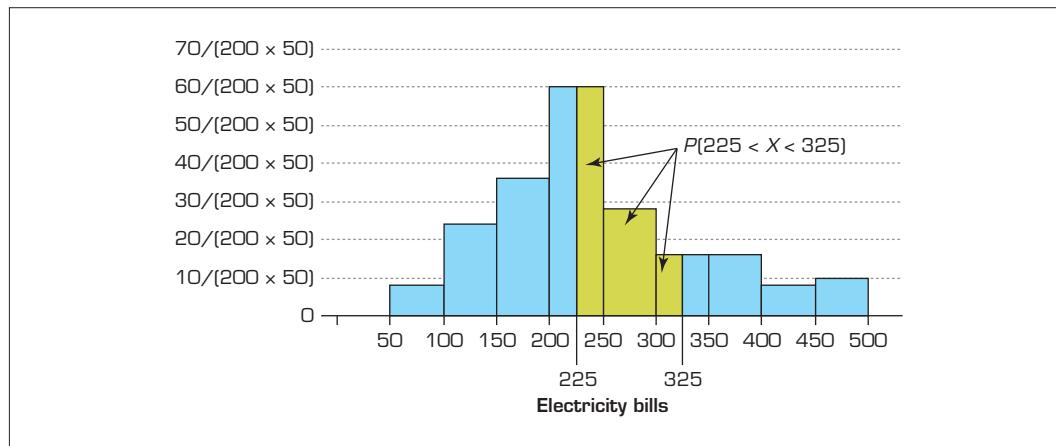
Notice that the sum of the probabilities equals 1. To proceed, we set the values along the vertical axis so that the area in all the rectangles together adds to 1. We accomplish this by dividing each relative frequency by the width of the interval, which is 50. The result is a rectangle over each interval whose area equals the probability that the random variable will fall into that interval.

To determine probabilities of ranges other than those created when we drew the histogram, we apply the same approach. For example, the probability that an electricity bill will fall between \$225 and \$325 is equal to the area between 225 and 325, as shown in **Figure 8.2**.

The areas in the shaded rectangles are calculated and added together as follows:

Interval	Height of rectangle	Base multiplied by height
$225 < X \leq 250$	$60/(200 \times 50) = 0.0060$	$(250 - 225) \times 0.0060 = 0.150$
$250 < X \leq 300$	$28/(200 \times 50) = 0.0028$	$(300 - 250) \times 0.0028 = 0.140$
$300 < X \leq 325$	$16/(200 \times 50) = 0.0016$	$(325 - 300) \times 0.0016 = 0.040$
		Total = 0.33

FIGURE 8.2 Histogram for Example 4.1: Relative frequencies divided by interval width

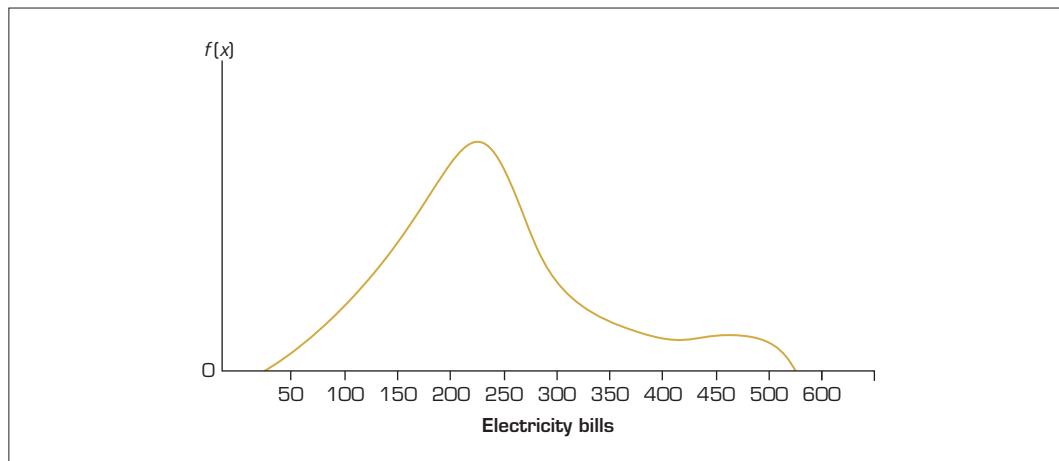


We estimate that the probability that a randomly selected electricity bill falls between \$225 and \$325 is 0.33.

If the histogram is drawn with a large number of small intervals, we can smooth the edges of the rectangles to produce a smooth curve, as shown in **Figure 8.3**. In many cases it is possible to determine a function that approximates the curve. The function is called a **probability density function**.

probability density function (pdf)

A function $f(x)$ such that (1) $f(x)$ is non-negative, (2) the total area under $f(x)$ is 1, (3) the area under $f(x)$ between the lines $x = a$ and $x = b$ gives the probability that the value of X is between a and b , where X is a continuous random variable.

FIGURE 8.3 Density function for Example 4.1

The requirements of a probability density function are stated in the following box.

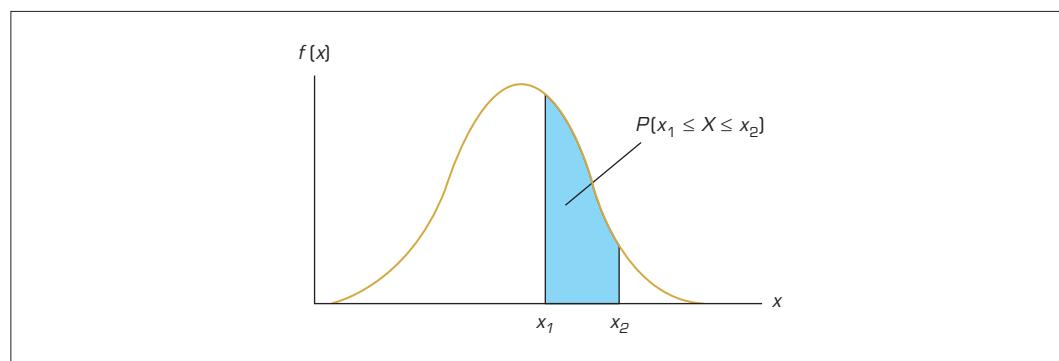
Requirements for a probability density function

The following requirements apply to a probability density function $f(x)$ whose range is $a \leq X \leq b$.

- 1** $f(x) \geq 0$ for all x between a and b .
- 2** The total area under the curve between a and b is 1.0.

It is important to note that $f(x)$ is not a probability. That is, $f(x) \neq P(X = x)$. As previously mentioned, when variable X is continuous, the probability that X will take any specific value is zero: $P(X = x) = 0$. Given a probability density function $f(x)$, the area under the graph of $f(x)$ between the two values x_1 and x_2 is the probability that X will take a value between x_1 and x_2 . This area is shaded in **Figure 8.4**.

A continuous random variable X has an expected value and a variance, just as a discrete random variable does.

FIGURE 8.4 Probability density function $f(x)$. Shaded area is $P(x_1 \leq X \leq x_2)$ 

8.2 Uniform distribution

To illustrate how we find the area under the curve that describes a probability density function, consider the **uniform probability distribution**, also called the *rectangular probability distribution*.

Uniform probability density function

The uniform distribution is described by the function

$$f(x) = \frac{1}{b-a}, \quad \text{where } a \leq x \leq b$$

The function is graphed in **Figure 8.5**. You can see why the distribution is called rectangular.

To calculate the probability of any interval, simply find the area under the curve. For example, to find the probability that X falls between x_1 and x_2 , determine the area of the rectangle whose base is $x_2 - x_1$ and whose height is $1/(b-a)$. **Figure 8.6** depicts the area we wish to find. Thus:

$$P(x_1 < X < x_2) = \text{base} \times \text{height} = (x_2 - x_1) \frac{1}{(b-a)}$$

FIGURE 8.5 Uniform distribution

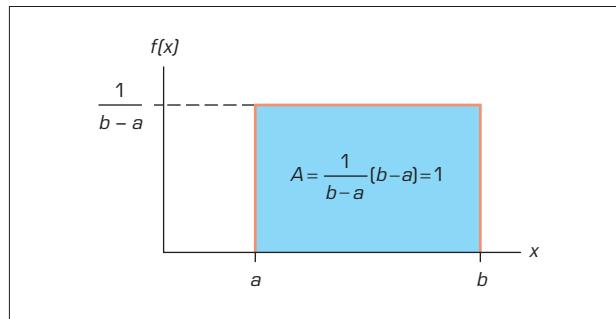
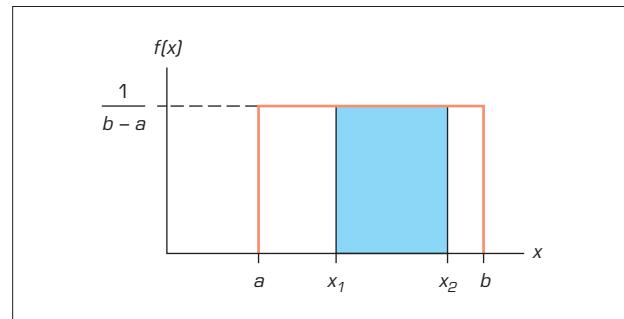


FIGURE 8.6 $P(x_1 < X < x_2)$



EXAMPLE 8.1

LO2

Uniformly distributed petrol sales

The volume of petrol sold daily at a service station is uniformly distributed with a minimum of 2000 litres and a maximum of 5000 litres.

- a Find the probability that daily sales will fall between 2500 and 3000 litres.
- b What is the probability that the service station will sell at least 4000 litres?
- c What is the probability that the station will sell exactly 2500 litres?
- d What is the value of daily sales below which 25% of the daily sales (lowest quarter) will lie?
- e What is the value of daily sales above which 10% of the daily sales will lie?

Solution

If X denotes the amount of petrol sold daily at a service station, then X can take any value in the interval $2000 \leq X \leq 5000$. The probability density function is:

$$f(x) = \frac{1}{5000 - 2000} = \frac{1}{3000}, \quad 2000 \leq x \leq 5000$$



- a The probability that X falls between 2500 and 3000 is the area under the curve between 2500 and 3000, as depicted in **Figure 8.7(a)**. The area of a rectangle is the base times the height. Thus,

$$\begin{aligned} P(2500 \leq X \leq 3000) &= \text{base} \times \text{height} \\ &= (3000 - 2500) \left(\frac{1}{3000} \right) = 0.1667 \end{aligned}$$

- b The probability that X is at least 4000 litres means the probability that X falls between 4000 and 5000, as depicted in **Figure 8.7(b)**. Thus,

$$\begin{aligned} P(X \geq 4000) &= P(4000 \leq X \leq 5000) \\ &= (5000 - 4000) \left(\frac{1}{3000} \right) = 0.3333 \end{aligned}$$

- c The probability that X is equal to 2500 litres is:

$$P(X = 2500) = 0$$

Because there is an uncountable infinite number of values of X , the probability of each individual value is zero. Moreover, as you can see from **Figure 8.7(c)**, the area of a line is 0.

- d The probability that the value of daily sales is less than a sales value c is 0.25 (i.e. in the lowest quarter), as depicted in **Figure 8.7(d)**. That is, we need to find c such that:

$$P(2000 \leq X \leq c) = 0.25$$

This means

$$\left(\frac{1}{(5000 - 2000)} \right) (c - 2000) = 0.25$$

$$c = 2750 \text{ litres}$$

That is, the volume of daily sales that lie below 2750 litres falls in the lowest quarter of the sales.

- e The probability that the value of daily sales is above a sales value d is 0.10, as depicted in **Figure 8.7(e)**. That is, we need to find d such that:

$$P(d \leq X \leq 5000) = 0.10$$

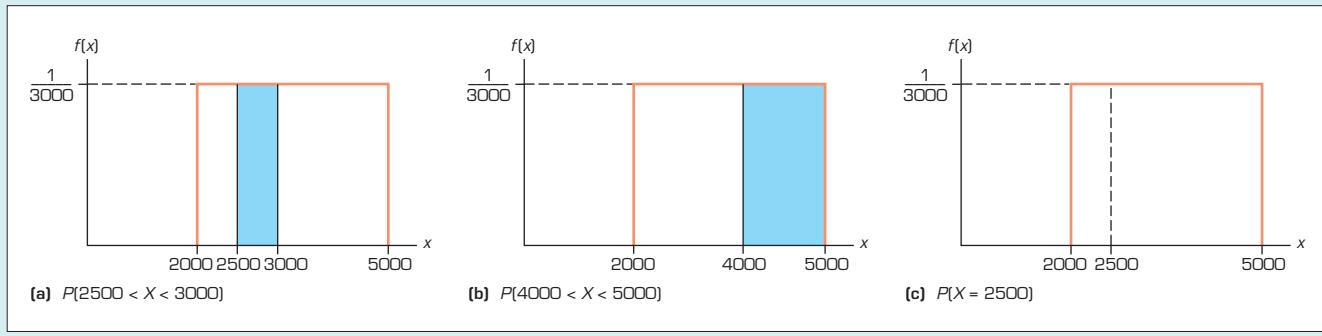
This means

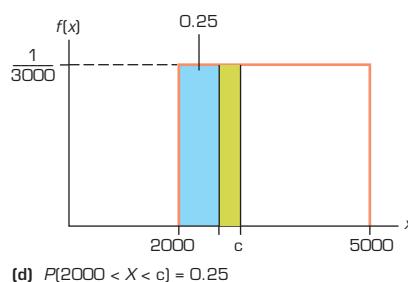
$$\left(\frac{1}{(5000 - 2000)} \right) (5000 - d) = 0.10$$

$$d = 4700 \text{ litres}$$

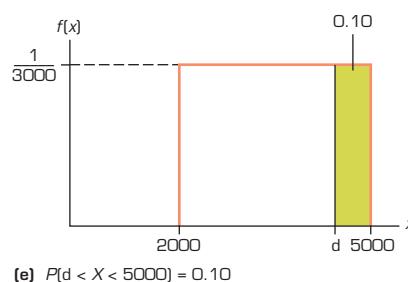
That is, volume of daily sales higher than 4700 litres will fall in the top 10% of the daily sales.

FIGURE 8.7 Probabilities from a uniform distribution



**FIGURE 8.7** (Continued)

$$(d) P(2000 < X < c) = 0.25$$



$$(e) P(d < X < 5000) = 0.10$$

EXERCISES

Learning the techniques

- 8.1** A random variable X is uniformly distributed between 5 and 25.
- Draw the density function.
 - Find $P(X > 25)$.
 - Find $P(10 < X < 15)$.
 - Find $P(5.0 < X < 5.1)$.
- 8.2** Consider an investment whose return (X) is uniformly distributed between \$20 and \$60. Draw the density function and calculate the following probabilities.
- $P(20 < X < 40)$
 - $P(X < 25)$
 - $P(35 < X < 65)$
- 8.3** Consider a random variable X having the uniform density function $f(x)$, with $a = 20$ and $b = 30$.
- Define and graph the density function $f(x)$.
 - Verify that $f(x)$ is a probability density function.
 - Find $P(22 \leq X \leq 30)$.
 - Find $P(X = 25)$.

Applying the techniques

- 8.4 Self-correcting exercise.** The volume of petrol sold daily at an independent service station is uniformly distributed with a minimum of 20 000 litres and maximum of 50 000 litres.
- Find the probability that daily sales will fall between 25 000 and 30 000 litres.
 - What is the probability that the service station will sell at least 40 000 litres?
 - What is the probability that the service station will sell exactly 25 000 litres?

- 8.5** The amount of time it takes for a student to complete a statistics quiz is uniformly distributed between 30 and 60 minutes. One student is selected at random. Find the probability of the following events.
- The student requires more than 55 minutes to complete the quiz.
 - The student completes the quiz in a time between 30 and 40 minutes.
 - The student completes the quiz in exactly 37.23 minutes.
- 8.6** Refer to Exercise 8.5. The lecturer wants to reward (with bonus marks) students who are in the lowest quarter of completion times. What completion time should be used for the cut-off for awarding bonus marks?
- 8.7** Refer to Exercise 8.5. The lecturer would like to track (and possibly help) students who are in the top 10% of completion times. What completion time should be used?
- 8.8** The weekly output of a steel mill is a uniformly distributed random variable that lies between 110 and 175 tonnes.
- Compute the probability that the steel mill will produce more than 150 tonnes next week.
 - Determine the probability that the steel mill will produce between 120 and 160 tonnes next week.
- 8.9** Refer to Exercise 8.8. The operations manager labels any week that is in the bottom 20% of production a ‘bad week’. How many tonnes should be used to define a bad week?

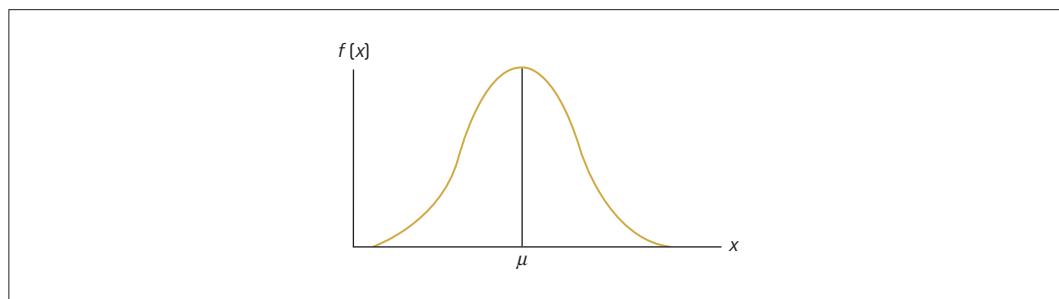
8.3 Normal distribution

normal distribution

The most important continuous distribution. The curve is bell-shaped, and describes many phenomena that occur both in nature and in business.

The normal distribution is the most important of all probability distributions because of its crucial role in statistical inference. The graph of the **normal distribution** is the familiar symmetrical, bell-shaped curve shown in **Figure 8.8**. One reason for the importance of the normal distribution is that it usefully models or describes the distributions of numerous random variables that arise in practice, such as the heights or weights of a group of people, the total annual sales of a firm, the results of a class of students, and the measurement errors that arise in the performance of an experiment. In examples such as these, the observed measurements tend to cluster in a symmetrical fashion about the central value, giving rise to a bell-shaped distribution curve.

FIGURE 8.8 Symmetrical, bell-shaped normal distribution



A second reason for the importance of the normal distribution is that this distribution provides a useful approximation to many other distributions, including discrete ones such as the binomial distribution. Finally, as we shall see in Chapter 10, the normal distribution is the cornerstone distribution of statistical inference, representing the distribution of the possible estimates of a population parameter that may arise from different samples. This last point, in fact, is primarily responsible for the importance of the normal distribution.

Normal distribution

A random variable X with mean μ and variance σ^2 is normally distributed if its probability density function is given by

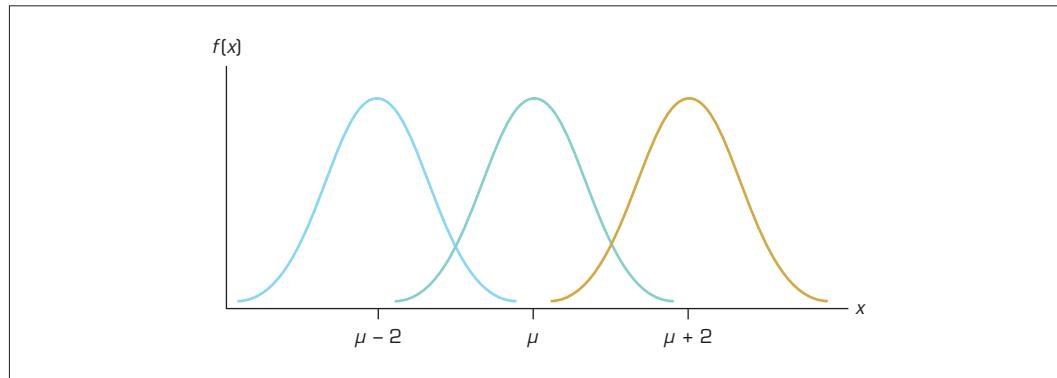
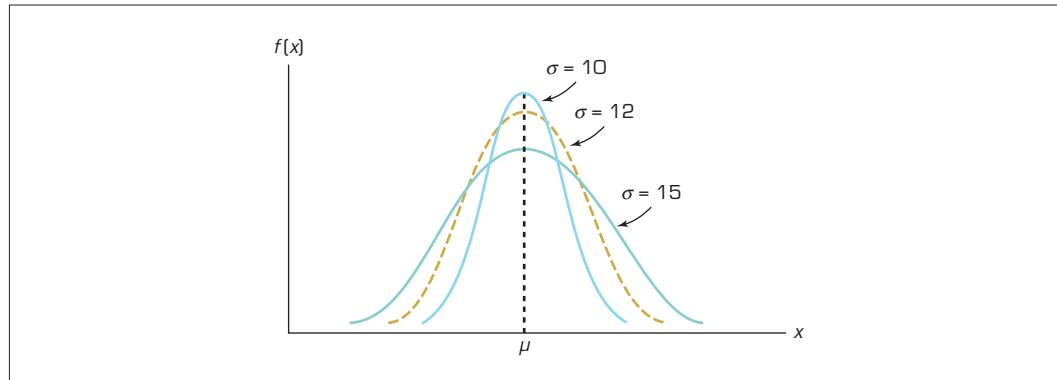
$$f(x) = \left(\frac{1}{\sigma\sqrt{2\pi}} \right) e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \quad -\infty < x < \infty$$

where $\pi = 3.14159\dots$ and $e = 2.71828\dots$

Figure 8.8 depicts a normal distribution. Notice that the curve is symmetric about its mean and the random variable ranges between $-\infty$ and ∞ .

The normal distribution is described by two parameters, the mean μ and the standard deviation σ . In **Figure 8.9**, we demonstrate the effect of changing the value of μ . Obviously, increasing μ shifts the curve to the right and decreasing μ shifts it to the left.

Figure 8.10 describes the effect of σ . Larger values of σ widen the curve and smaller values narrow it.

FIGURE 8.9 Normal distributions with the same variance but different means**FIGURE 8.10** Normal distributions with the same mean but different standard deviations

8.3a Calculating normal probabilities

To calculate the probability that a **normal random variable** falls into any interval, we need to compute the area in the interval under the curve. Unfortunately, the function is not as simple as the uniform distribution, precluding the use of simple mathematics or even integral calculus. Instead we will resort to using a probability table much as we did in determining binomial and Poisson probabilities in Chapter 7. Recall that to determine binomial probabilities from Table 1 of Appendix B we needed a table for each value of n and a separate column for selected values of p . Similarly, to find Poisson probabilities, we needed a separate column for each value of μ that we chose to include in Table 2 of Appendix B. It would appear, then, that we will need a separate table for normal probabilities for a selected set of values of μ and σ . Fortunately, this won't be necessary.

Instead we reduce the number of tables needed to one by standardising the normal random variable. We standardise a random variable by subtracting its mean and dividing by its standard deviation. When the variable is normal, the transformed variable is called a **standard normal random variable** and denoted by Z . That is,

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

It can be shown that the mean and variance of this standard normal random variable are

$$\mu_Z = 0 \text{ and } \sigma_Z^2 = 1$$

The distribution of the standard normal random variable Z is called the **standard normal distribution** or **z -distribution**. If we know the mean and standard deviation of a normally distributed random variable, we can always transform the probability statement about X into a probability statement about Z . Consequently, we need only one table: Table 3 in Appendix B, the standard normal probability table, which is reproduced here as **Table 8.1**.

normal random variable

A random variable that is normally distributed.

standard normal random variable

Labelled Z , a normal random variable with a mean of 0 and a standard deviation of 1.

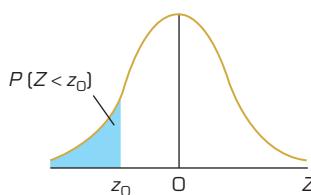
standard normal distribution (z -distribution)

Normal distribution with a mean of 0 and a standard deviation of 1.

TABLE 8.1 Reproduction of Table 3 in Appendix B: Standard normal curve areas, $P[Z < z_0]$

<i>z</i>	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

The table provides the probability that a standard normal random variable Z takes a value less than z_0 , $P(Z < z_0)$, the area of which is shaded in the graph below.



Because the normal curve is symmetric about its mean and the total area under the curve is 1, we can state the following:

$$P(Z > 0) = P(Z < 0) = 0.5 \text{ and } P(Z > z_0) = 1 - P(Z \leq z_0)$$

The z -values in the table range from -3.09 to 3.09 . As the table lists no values beyond 3.09 , we approximate any area beyond 3.10 as 0. That is,

$$P(Z > 3.10) = P(Z < -3.10) \approx 0$$

To use the standard normal table (see Table 8.1), we simply find the value of z and read the probability. The numbers in the left column describe the values of Z to one decimal place and the column headings specify the second decimal place. Thus, to use the standard normal table, we must always round z to two decimal places.

For example, the probability $P(Z < 1.00)$ is found in Table 8.1 by finding 1.0 in the left margin and under the heading 0.00 finding 0.8413, see Figure 8.11. That is, $P(Z < 1.00) = 0.8413$. Further, the probability $P(Z < 1.04)$ is found in the same row but under the heading 0.04. It is 0.8508. That is, $P(Z < 1.04) = 0.8508$.

FIGURE 8.11 Calculating $P(Z < 1.00)$

z	.00	.01	.02	.03	.04
0.8	.7881	.7910	.7939	.7967	.7995
0.9	.8159	.8186	.8212	.8238	.8264
1.0	<u>.8413</u>	.8438	.8461	.8485	.8508
1.1	.8643	.8665	.8686	.8708	.8729
1.2	.8849	.8869	.8888	.8907	.8925

Suppose we want to find $P(-0.5 < Z < 1.0)$. This probability is actually the difference between two probabilities as shown in Figure 8.12.

$$P(-0.5 < Z < 1.0) = P(Z < 1.0) - P(Z < -0.5)$$

Both probabilities can be easily determined from the table:

$$P(Z < 1.0) = 0.8413 \text{ and } P(Z < -0.5) = 0.3085$$

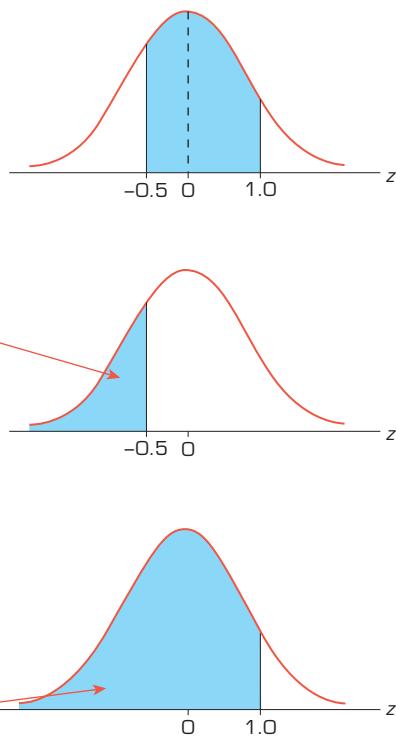
Thus,

$$P(-0.5 < Z < 1.0) = 0.8413 - 0.3085 = 0.5328$$

Figure 8.12 shows how this calculation is performed.

FIGURE 8.12 Calculating $P(-0.5 < Z < 1.0)$

z	0.00	0.01	0.02
-0.8	.2119	.2090	.2061
-0.7	.2420	.2389	.2358
-0.6	.2743	.2709	.2676
-0.5	.3085	.3050	.3015
-0.4	.3446	.3409	.3372
-0.3	.3821	.3783	.3745
-0.2	.4207	.4168	.4129
-0.1	.4602	.4562	.4522
-0.0	.5000	.4960	.4920
0.0	.5000	.5040	.5080
0.1	.5398	.5438	.5478
0.2	.5793	.5832	.5871
0.3	.6179	.6217	.6255
0.4	.6554	.6591	.6628
0.5	.6915	.6950	.6985
0.6	.7257	.7291	.7324
0.7	.7580	.7611	.7642
0.8	.7881	.7910	.7939
0.9	.8159	.8186	.8212
1.0	.8413	.8438	.8461



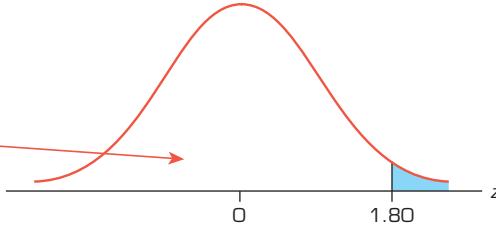
We can also determine the probability that the standard normal random variable is greater than some value of z . For example, we find the probability that Z is greater than 1.80 by determining the probability that Z is less than 1.80 and subtracting that value from 1. By applying the complement rule, we get

$$P(Z > 1.80) = 1 - P(Z < 1.80) = 1 - 0.9641 = 0.0359$$

See Figure 8.13.

FIGURE 8.13 Calculating $P(Z > 1.80)$

z	.00	.01	.02
1.6	.9452	.9463	.9474
1.7	.9554	.9564	.9573
1.8	.9641	.9649	.9656
1.9	.9713	.9719	.9726
2.0	.9772	.9778	.9783



EXAMPLE 8.2

LO3

Calculating the standard normal probabilities

Determine the following probabilities:

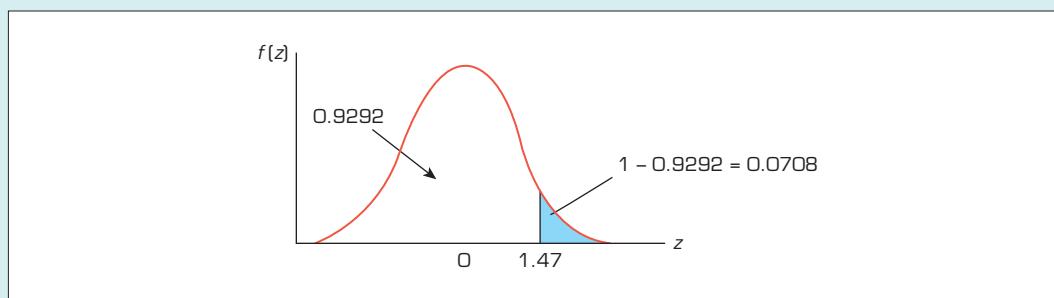
- $P(Z \geq 1.47)$
- $P(-2.25 \leq Z \leq 1.85)$
- $P(0.65 \leq Z \leq 1.36)$

Solution

- a It is always advisable to begin by sketching a diagram and indicating the area of interest under the normal curve, as shown in **Figure 8.14**. Since the entire area under the normal curve equals 1, the required probability is

$$P(Z \geq 1.47) = 1 - P(Z < 1.47)$$

FIGURE 8.14 Shaded area is $P(Z \geq 1.47)$ in Example 8.2(a)



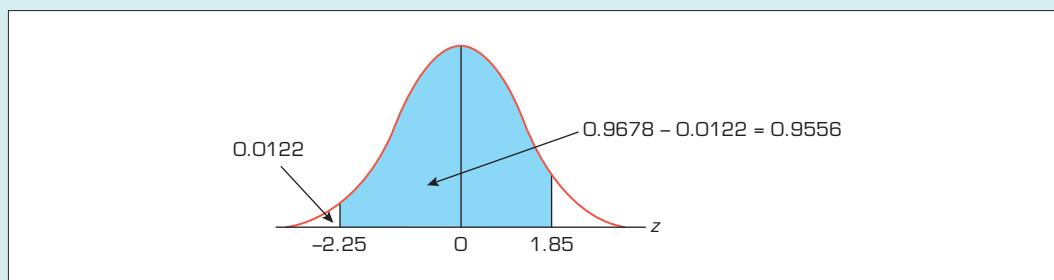
The probability $P(Z < 1.47)$ can be found in **Table 8.1** (or in Table 3, Appendix B). Locating $z = 1.47$ in **Table 8.1**, we find that this area is 0.9292. Therefore, the required probability is

$$\begin{aligned} P(Z \geq 1.47) &= 1 - P(Z < 1.47) \\ &= 1 - 0.9292 \\ &= 0.0708 \end{aligned}$$

- b Whenever the area of interest straddles the mean, as in **Figure 8.15**, we must express it as the difference of the portions to the left and to the right of the mean. The required probability $P(-2.25 \leq Z \leq 1.85)$ is therefore

$$P(-2.25 \leq Z \leq 1.85) = P(Z < 1.85) - P(Z < -2.25)$$

FIGURE 8.15 Shaded area is $P(-2.25 \leq Z \leq 1.85)$ in Example 8.2(b)





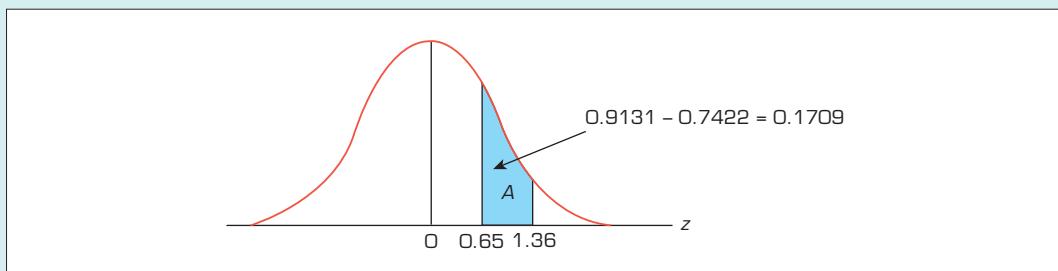
From **Table 8.1**, we find that $P(Z < 1.85) = 0.9678$ and $P(Z < -2.25) = 0.0122$. Therefore, the required probability is

$$\begin{aligned} P(-2.25 \leq Z \leq 1.85) &= P(Z < 1.85) - P(Z < -2.25) \\ &= 0.9678 - 0.0122 \\ &= 0.9556 \end{aligned}$$

- c $P(0.65 \leq Z \leq 1.36)$ corresponds to the shaded area A in **Figure 8.16**. We can express A as the difference between two areas. That is,

$$\begin{aligned} A = P(0.65 \leq Z \leq 1.36) &= P(Z < 1.36) - P(Z < 0.65) \\ &= 0.9131 - 0.7422 \\ &= 0.1709 \end{aligned}$$

FIGURE 8.16 Shaded area is $P(0.65 \leq Z \leq 1.36)$ in Example 8.2(c)



In general, a probability statement about a normal random variable X with mean μ and variance σ^2 is transformed into a statement about the standardised normal random variable Z . To illustrate how we proceed, consider the following example.

EXAMPLE 8.3

LO3

Normally distributed assembly time

Suppose that the amount of time to assemble a computer is normally distributed with a mean of 60 minutes and a standard deviation of 8 minutes. We would like to know the probability that a computer is assembled in a time between 60 and 70 minutes.

Solution

We want to find the probability

$$P(60 < X < 70)$$

Figure 8.17(a) describes a normal curve with mean $\mu = 60$ and standard deviation $\sigma = 8$, and the area we want to find. The first step is to standardise X by converting X to $Z = (X - \mu)/\sigma$. However, if we perform any operations on X , we must perform the same operations on 60 and 70. Thus:

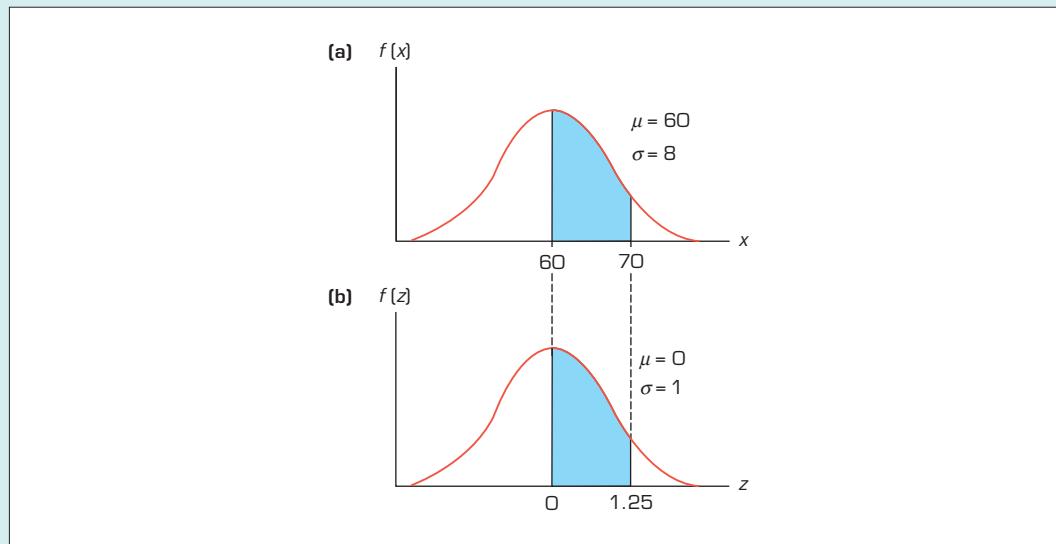
$$P(60 < X < 70) = P\left(\frac{60 - 60}{8} < \frac{X - \mu}{\sigma} < \frac{70 - 60}{8}\right) = P(0 < Z < 1.25)$$





Figure 8.17(b) describes the transformation that has taken place. Notice that the variable X was transformed into Z , 60 was transformed into 0, and 70 was transformed into 1.25. However, the area has not changed. That is, the probability that we wish to compute, $P(60 < X < 70)$, is identical to $P(0 < Z < 1.25)$.

FIGURE 8.17 Shaded area is $P(60 \leq X \leq 70) = P(0 < Z < 1.25)$



Note that,

$$P(0 < Z < 1.25) = P(Z < 1.25) - P(Z < 0)$$

We can now read the probability from the standard normal table in Table 3 of Appendix B or the table reproduced here in **Table 8.1**. Therefore:

$$\begin{aligned} P(60 < X < 70) &= P(0 < Z < 1.25) \\ &= P(Z < 1.25) - P(Z < 0) \\ &= 0.8944 - 0.5000 \\ &= 0.3944 \end{aligned}$$

EXAMPLE 8.4

L03

Likelihood of returns on investment

A venture capital company feels that the rate of return (X) on a proposed investment is approximately normally distributed, with a mean of 30% and a standard deviation of 10%.

- a Find the probability that the return will exceed 55%.
- b Find the probability that the return will be less than 22%.
- c Find the probability of losing money (i.e. the return is negative).

Solution

The rate of return, X , is normally distributed with $\mu = 30$ and $\sigma = 10$.

- a We need the $P(X > 55)$. The value of Z corresponding to $X = 55$ is

$$Z = \frac{X - \mu}{\sigma} = \frac{55 - 30}{10} = 2.5$$

Therefore:

$$P(X > 55) = P(Z > 2.5)$$

Figure 8.18(a) shows the required area $P(Z > 2.5)$, together with corresponding values of X .

Therefore:

$$\begin{aligned} P(X > 55) &= P(Z > 2.5) \\ &= 1 - P(Z < 2.5) \\ &= 1 - 0.9938 \quad (\text{from Table 8.1}) \\ &= 0.0062 \end{aligned}$$

The probability that the return will exceed 55% is 0.0062.

- b By the same logic as was used in part (a),

$$P(X < 22) = P\left(Z < \frac{22 - 30}{10}\right) = P(Z < -0.8) = 0.2119 \quad (\text{from Table 8.1})$$

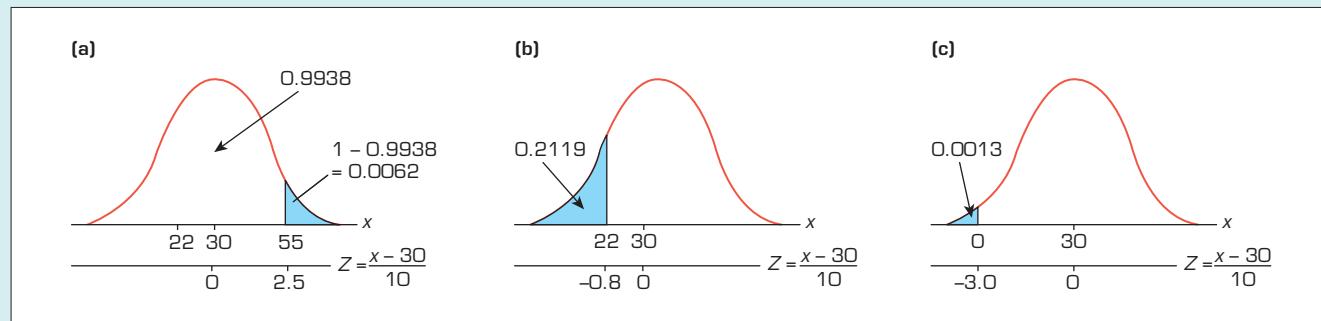
Figure 8.18(b) shows the required area $P(X < 22) = P(Z < -0.8)$. The probability that the return will be less than 22% is 0.2119.

- c The investment loses money when the return is negative. The required probability of negative return is $P(X < 0)$. Therefore,

$$P(X < 0) = P\left(Z < \frac{0 - 30}{10}\right) = P(Z < -3.0) = 0.0013 \quad (\text{from Table 8.1})$$

The probability of losing money is 0.0013. **Figure 8.18(c)** shows the required area.

FIGURE 8.18 Corresponding values of X and Z for Example 8.4



REAL-LIFE APPLICATIONS

Measuring risk

In previous chapters, we discussed several probability and statistical applications in finance in which we wanted to measure and perhaps reduce the risk associated with investments. In Example 4.2, we drew histograms to gauge the spread of the histogram of the returns on two investments. Then in Section 5.2, Example 5.11, we computed the standard deviation and variance as numerical measures of risk.

In Section 7.5, we developed an important application in finance in which we emphasised reducing the variance of the returns on a portfolio. However, we have not demonstrated why risk is measured by the variance and standard deviation. The following example corrects this deficiency.

EXAMPLE 8.5

LO3

Probability of a negative return on investment

Consider an investment whose return (X) is normally distributed, with a mean of 10% and a standard deviation of 5%.

- Find the probability of losing money (that is, the return is negative).
- Find the probability of losing money when the standard deviation is equal to 10%.

Solution

The rate of return, X , is normally distributed with $\mu = 10$ and $\sigma = 5$.

- The investment loses money when the return is negative. Thus we wish to determine $P(X < 0)$.

The first step is to standardise both X and 0 in the probability statement:

$$P(X < 0) = P\left(\frac{X - \mu}{\sigma} < \frac{0 - 10}{5}\right) = P(Z < -2.00) = 0.0228 \quad (\text{from Table 8.1})$$

Therefore, the probability of losing money is 0.0228.

- If we increase the standard deviation to 10%, the probability of suffering a loss becomes,

$$P(X < 0) = P\left(\frac{X - \mu}{\sigma} < \frac{0 - 10}{10}\right) = P(Z < -1.00) = 0.1587$$

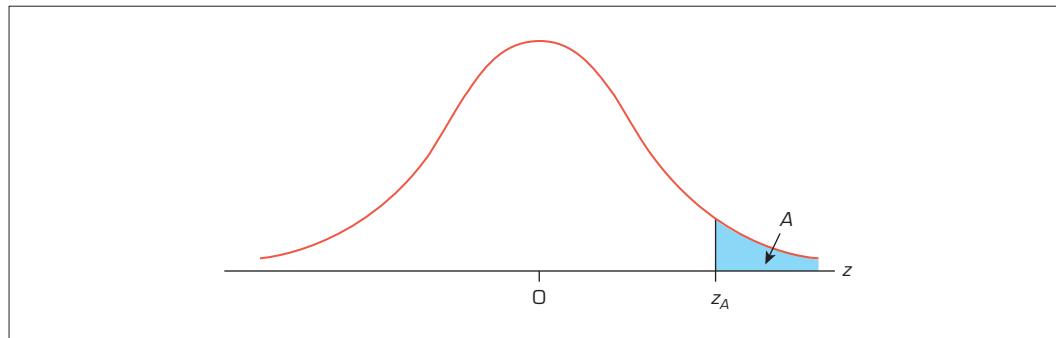
As you can see, increasing the standard deviation increases the probability of losing money. Note that increasing the standard deviation will also increase the probability that the return will exceed some relatively large amount. However, because investors tend to be risk averse, we emphasise the increased probability of negative returns when discussing the effect of increasing the standard deviation.

8.3b Finding values of Z

There is a family of problems that require us to determine the value of Z given a probability. We use the notation z_A to represent the value of Z such that the area to its right under the standard normal curve is A (see **Figure 8.19**). That is, z_A is a value of a standard normal random variable such that

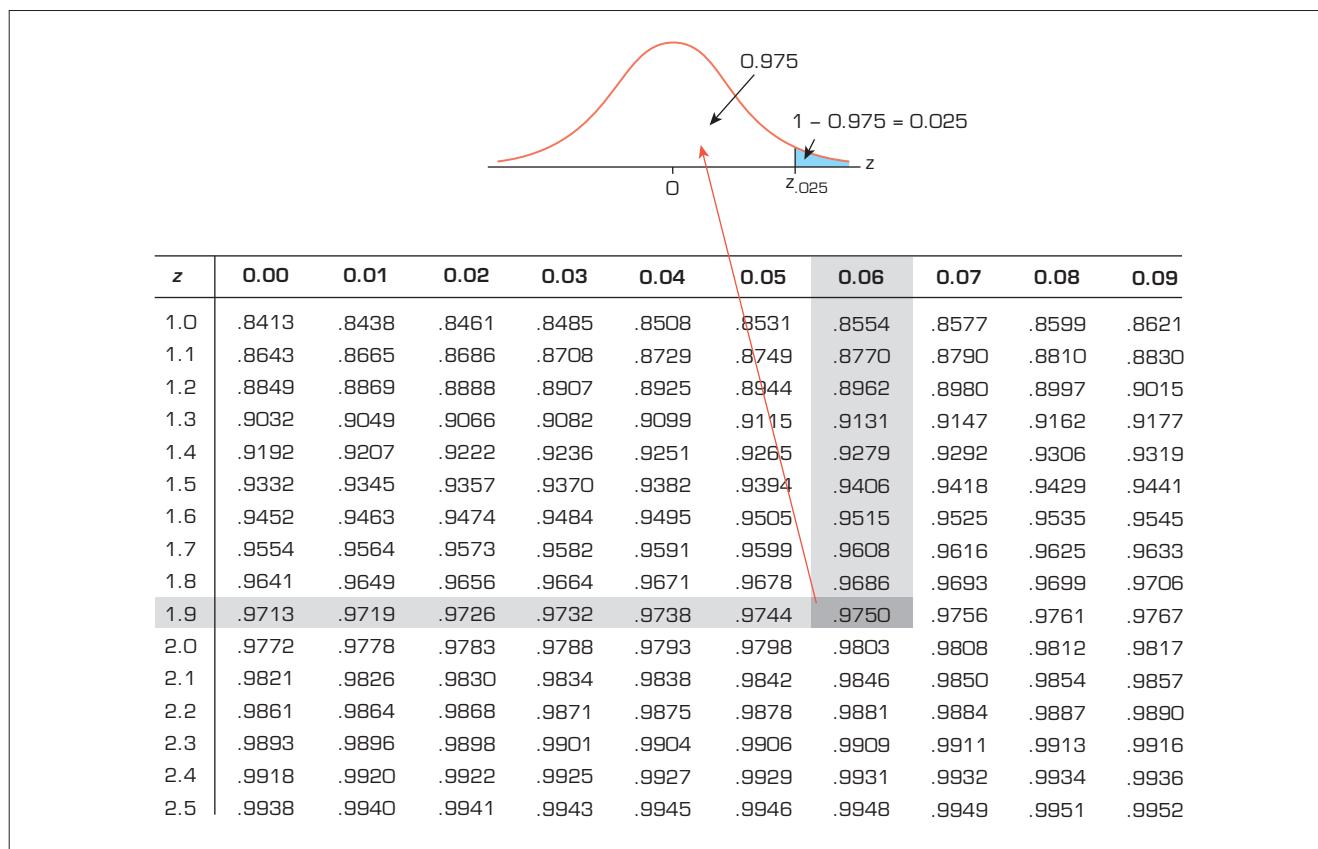
$$P(Z > z_A) = A$$

FIGURE 8.19 $P(Z > z_A) = A$



To find z_A for any value of A requires us to use the standard normal table backwards. As you saw in Example 8.2, to find a probability about Z we must find the value of z in the table and determine the probability associated with it. To use the table backwards, we need to specify a probability and then determine the z -value associated with it. We'll demonstrate by finding $z_{0.025}$. **Figure 8.20** depicts the standard normal curve and $z_{0.025}$.

FIGURE 8.20 Finding $z_{0.025}$



Because of the format of the standard normal table, we begin by determining the area less than $z_{0.025}$, which is $1 - 0.025 = 0.9750$. (Notice that we expressed this probability with four decimal places to make it easier for you to see what you need to do.) We now search through the probability part of the table looking for 0.9750. When we locate it, we see that the z -value associated with it is 1.96.

Thus, $z_{0.025} = 1.96$, which means that $P(Z > 1.96) = 0.025$ and $P(Z < 1.96) = 0.975$.

EXAMPLE 8.6

LO3

Finding $z_{0.05}$

Find the value of a standard normal random variable z_0 such that the probability that the random variable is greater than z_0 is 5%.

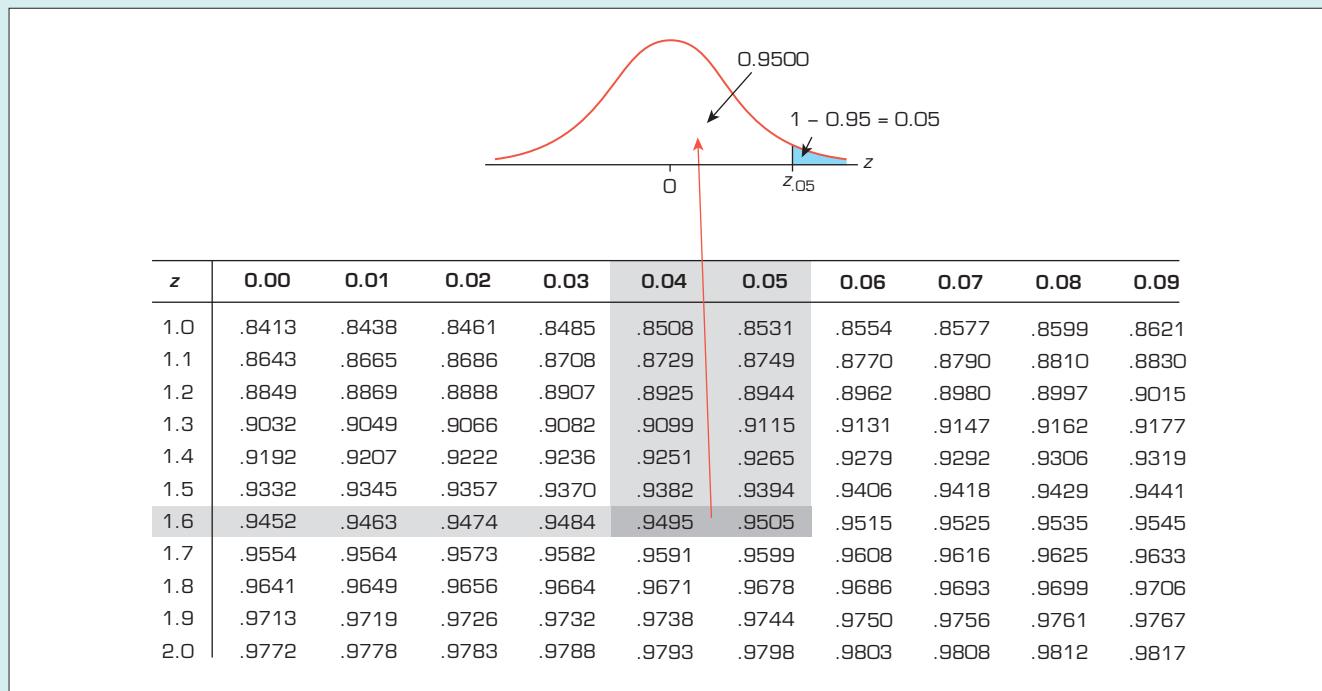
Solution

We wish to determine $z_{0.05}$, such that $P(Z > z_{0.05}) = 0.05$. **Figure 8.21** depicts the normal curve and $z_{0.05}$. If 0.05 is the area in the tail, then the probability less than $z_{0.05}$ must be $1 - 0.05 = 0.95$. To find $z_{0.05}$, we search the table looking for the probability 0.9500. We don't find the exact value of this probability, but we find two values



that are equally close: 0.9495 and 0.9505. The z-values associated with these probabilities are 1.64 and 1.65 respectively. The average is taken as $z_{0.05}$; thus, $z_{0.05} = 1.645$.

FIGURE 8.21 Finding $z_{0.05}$



EXAMPLE 8.7

LO3

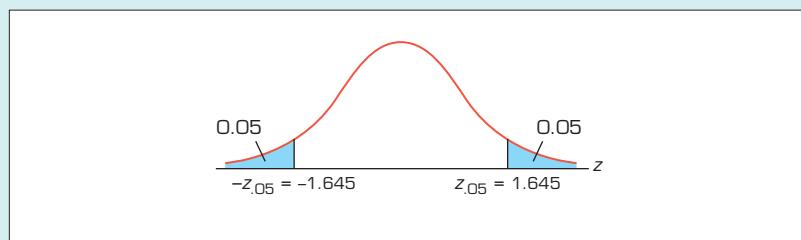
Finding $-z_{0.05}$

Find the value of a standard normal random variable z_0 such that the probability that the random variable is less than z_0 is 5%.

Solution

We need to find z_0 such that $P(Z < z_0) = 0.05$. As can be seen from **Figure 8.22**, we need to find $z_0 = -z_{0.05}$. In Example 8.6, we found $z_{0.05} = 1.645$. As the standard normal curve is symmetric about 0, $z_0 = -z_{0.05} = -1.645$.

FIGURE 8.22 $-z_{0.05}$



Alternatively, we can also look for the z-value that corresponds to 0.05 from **Table 8.1**. We cannot find 0.0500, but two values 0.0505 and 0.0495 are close. The z-values associated with these probabilities are -1.64 and -1.65 . The average is $-z_{0.05} = -1.645$.

EXAMPLE 8.8

LO3

Finding z_0

If Z is a standard normal variable, determine the value z_0 for which $P(Z \leq z_0) = 0.6331$.

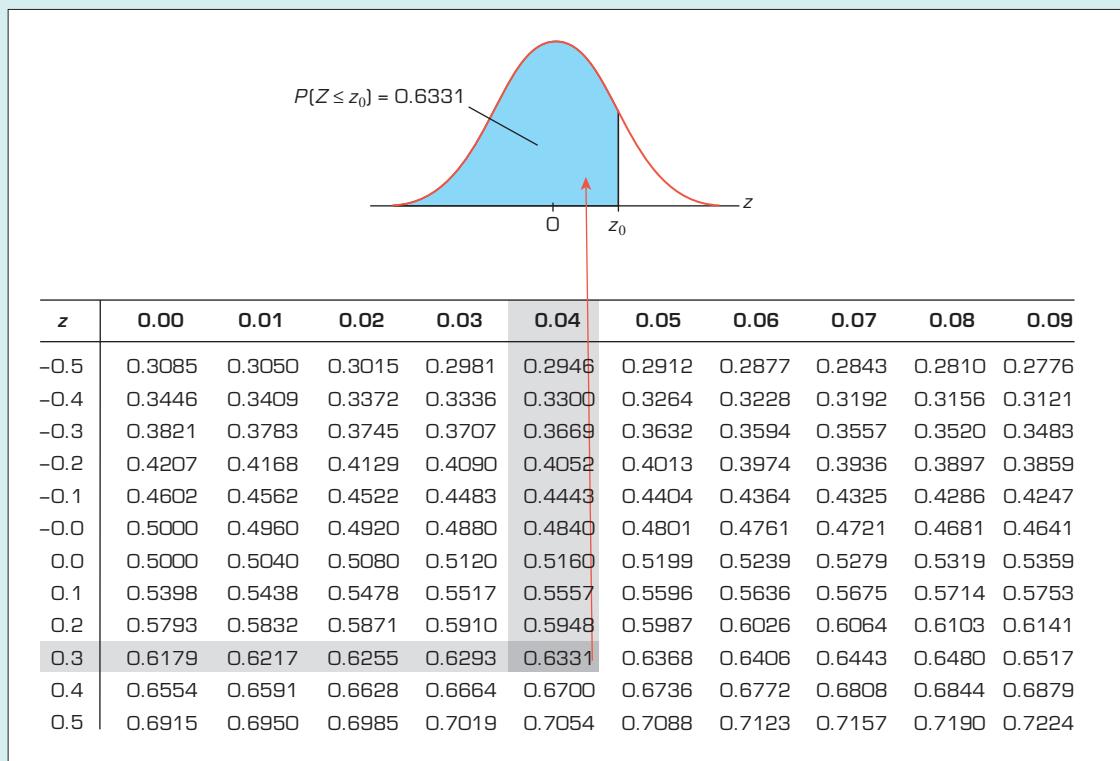
Solution

Since the area to the left of $z = 0$ is 0.5 and the given area is greater than 0.5, z_0 must be a positive number, as indicated in **Figure 8.23**. That is,

$$P(Z \leq z_0) = 0.6331$$

Figure 8.23 depicts the area and the corresponding z_0 value. Locating the area 0.6331 in the body of **Table 8.1**, we find that $P(Z \leq 0.34) = 0.6331$. Therefore, $z_0 = 0.34$.

FIGURE 8.23 Finding z_0 such that $P(Z \leq z_0) = 0.6331$



8.3c Using the computer to find normal probabilities

In the following Commands box, we describe how to use Excel to calculate normal probabilities. The probability that is produced is of the form $P(X < x)$, which means that the output is the probability that a normal random variable with a given mean and standard deviation falls between $-\infty$ and x . That is, the computer will calculate $P(-\infty < X < x)$ for any value of x . Note that the normal table (**Table 8.1** or Table 3 in Appendix B) lists the probabilities of the form $P(-\infty < Z < z)$.

We can employ Excel to calculate probabilities and values of x and z . To compute cumulative normal probabilities $P(X < x)$, proceed as follows.

EXCEL COMMANDS TO CALCULATE NORMAL PROBABILITIES

For a standard normal variable Z (mean 0 and standard deviation 1)

To calculate the probabilities $P(Z < z)$, type the following into any active cell:

=NORMSDIST([z])

For example, to calculate $P(Z < 1.96)$, type the following into any active cell: **=NORMSDIST(1.96)** This would give a result of 0.975. That is, $P(Z < 1.96) = 0.975$.

For a normal variable X with mean μ and standard deviation σ

To calculate the probabilities, $P(X < x)$, associated with a normal random variable X , type the following into any active cell:

=NORMDIST([x],[μ],[σ],True)

Typing 'True' yields a cumulative probability, $P(X \leq x)$, and typing 'False' will produce the value of the normal density function, a number with little meaning.

For example, in Example 8.4(b), to calculate the probability $P(X < 22)$, where X is normally distributed with mean 30 and standard deviation 10, we enter: **=NORMDIST(22,30,10,True)** This would give a result of 0.2119. That is, $P(X < 22) = 0.2119$.

Also, in Example 8.3, to calculate the probability $P(60 < X < 70)$, where X is normally distributed with mean 60 and standard deviation 8, we first write $P(60 < X < 70) = P(X < 70) - P(X < 60)$ and then to instruct Excel to calculate this probability, we enter **=NORMDIST(70,60,8,True)-NORMDIST(60,60,8,True)** Excel will produce the probability value of 0.3944. That is, $P(60 < X < 70) = 0.3944$.

To determine a value of Z or X given a cumulative probability, follow these commands.

EXCEL COMMANDS TO CALCULATE THE VALUE OF Z OR X

For a standard normal variable Z (mean 0 and standard deviation 1)

To calculate the value of z_0 , where, $P(Z < z_0) = A$, type the following into any active cell:

=NORMSINV([A])

For example, in Example 8.8, $P(Z < z_0) = 0.6331$. To calculate z_0 , enter the following into an active cell:

=NORMSINV(0.6331). Excel will produce the value of z_0 as 0.34.

To calculate the value of z_0 , where $P(Z > z_0) = A$, type the following into any active cell:

=NORMSINV([1-A])

For example, if $P(Z > z_0) = 0.025$, to calculate z_0 , as $A = 0.025$ gives $1 - A = 1 - 0.025 = 0.975$, enter the following into any active cell: **=NORMSINV(0.975)**. Excel will produce the value of $z_0 = 1.96$.

For a normal variable X with mean μ and standard deviation σ

To calculate a value x_0 , given the probability $P(X < x_0) = A$, type the following into any active cell:

=NORMINV([A],[μ],[σ])

For example, if $P(X < x_0) = 0.2119$, where X is normally distributed with mean 30 and standard deviation 10, to calculate x_0 , type the following into any active cell: **=NORMINV(0.2119,30,10)**. Excel would produce the value $x_0 = 22$.

To calculate a value x_0 given the probability $P(X > x_0) = A$, type the following into any active cell:

=NORMINV([1-A],[μ],[σ])

For example, if $P(X > x_0) = 0.0062$, where X is normally distributed with mean 60 and standard deviation 8, to calculate x_0 , as $A = 0.0062$ gives $1 - A = 1 - 0.0062 = 0.9938$, type the following into any active cell: **=NORMINV(0.9938,60,8)**. Excel would produce the value $x_0 = 80.00$.

8.3d z_A and percentiles

In Chapter 5 we introduced percentiles, which are measures of relative standing. The values of z_A are the $100(1 - A)$ th percentiles of a standard normal random variable. For example, $z_{0.05} = 1.645$, which means that 1.645 is the 95th percentile; 95% of all values of Z are below it and 5% are above it. We interpret other values of z_A similarly.

REAL-LIFE APPLICATIONS

Inventory management

Every organisation maintains some inventory, which is defined as a stock of items. For example, grocery stores hold inventories of almost all the products they sell. When the total number of products drops to a specified level, the manager arranges for the delivery of more products. An automobile repair shop keeps an inventory of a large number of replacement parts. A school keeps stock of items that it uses regularly, including chalk, pens, envelopes, file folders and paper clips. There are costs associated with inventories. These include the cost of capital, losses (theft and obsolescence) and warehouse space, as well as maintenance and record keeping. Management scientists have developed many models to help determine the optimum inventory level that balances the cost of inventory with the cost of shortages and the cost of making many small orders. Several of these models are deterministic – that is, they assume that the demand for the product is constant. However, in most realistic situations the demand is a random variable. One commonly applied probabilistic model assumes that the demand during *lead time* is a normally distributed random variable. *Lead time* is defined as the amount of time between when the order is placed and when it is delivered.

The quantity ordered is usually calculated by attempting to minimise the total costs, including the cost of ordering and the cost of maintaining inventory. (This topic is discussed in most management science



Source: iStock.com/stevecoleimages

courses.) Another critical decision involves the *reorder point*, which is the level of inventory at which an order is issued to its supplier. If the reorder point is too low, the company will run out of product, suffering the loss of sales and, potentially, customers, who will go to a competitor. If the reorder point is too high, the company will be carrying too much inventory, which costs money to buy and store. In some companies inventory has a tendency to walk out the back door or become obsolete. As a result, managers create a *safety stock*, which is the extra amount of inventory to reduce the times when the company has a shortage. They do so by setting a service level, which is the probability that the company will not experience a shortage. The method used to determine the reorder point is demonstrated with Example 8.9.

EXAMPLE 8.9

L03

Determining the reorder point

During spring, the demand for electric fans at a large home-improvement store is quite strong. The company tracks inventory using a computer system so that it knows how many fans are in the inventory at any time. The policy is to order a new shipment of 250 fans when the inventory level falls to the reorder point, which is 100. However, this policy has resulted in frequent shortages, resulting in lost sales because both lead time and demand are highly variable. The manager would like to reduce the incidence of shortages so that only 5% of orders will arrive after the inventory drops to 0 (resulting in a shortage). This policy is expressed as a 95% service level. From previous periods, the company has determined that demand during lead time is normally distributed with a mean of 200 and a standard deviation of 50. Determine the reorder point.



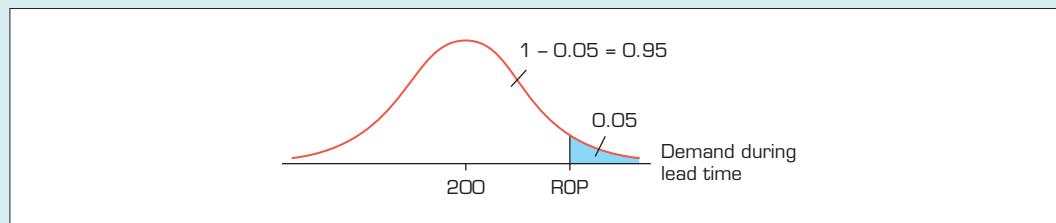
Solution

The reorder point is set so that the probability that demand during lead time exceeds this quantity is 5%.

Figure 8.24 depicts demand during lead time and the reorder point.

As we did in the solution to Example 8.8 we find the standard normal value such that the area to its left is $1 - 0.05 = 0.95$. The standardised value of the reorder point is $z_{0.05} = 1.645$. To find the reorder point (ROP), we must unstandardise $z_{0.05}$.

FIGURE 8.24 Distribution of demand during lead time



$$z_{0.05} = \frac{ROP - \mu}{\sigma}$$

$$1.645 = \frac{ROP - 200}{50}$$

$$\begin{aligned} ROP &= 50(1.645) + 200 \\ &= 282.25 \end{aligned}$$

The policy is to order a new batch of fans when there are 283 fans left in the inventory (we round 282.25 up to 283).

We will now provide the solution to the opening example in this chapter.

SPOTLIGHT ON STATISTICS

Would the pizza business survive next year? Solution

Let X be the total sales (in millions of dollars) for the pizza business next year. Based on the current sales records, total sales X is normally distributed with mean $\mu = \$3m$ and standard deviation $\sigma = \$0.5m$.

For the pizza business to survive next year, total sales should exceed the breakeven level of \$2m. That is, the probability of X exceeding 2 should be positive, $P(X > 2) > 0$. If this probability is high, that would mean the chance of survival of the business will be good.

We now calculate $P(X > 2)$.

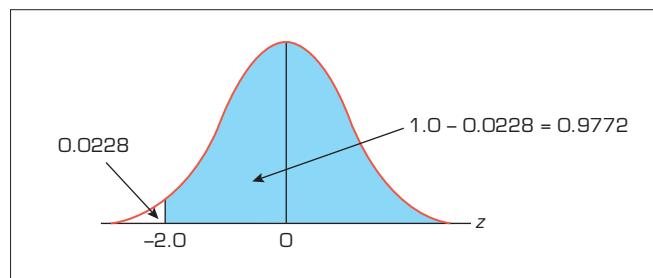
$$\begin{aligned} P(X > 2) &= P\left(Z > \frac{2.0 - 3.0}{0.5}\right) \\ &= P(Z > -2.0) \\ &= 1 - P(Z < -2.0) \\ &= 1 - 0.0228 \quad (\text{from Table 8.1}) \\ &= 0.9772 \quad (\text{see Figure 8.25}) \end{aligned}$$



Source: iStock.com/RaStudio

As the $P(X > 2)$ is very high, the businessman can purchase the pizza business with a high level of confidence that it will survive next year.

FIGURE 8.25 Finding $P(Z > -2.0)$



The businessman would also like to know the sales level that has only a 9% likelihood of being exceeded next year.

We've labelled the sales level that has only a 9% chance of being exceeded next year as $x_{0.09}$ such that:

$$P(X > x_{0.09}) = 0.09 \quad \text{or} \quad P(X < x_{0.09}) = 0.91$$

Let $z_{0.09}$ be the standardised value of $x_{0.09}$. First we find z such that

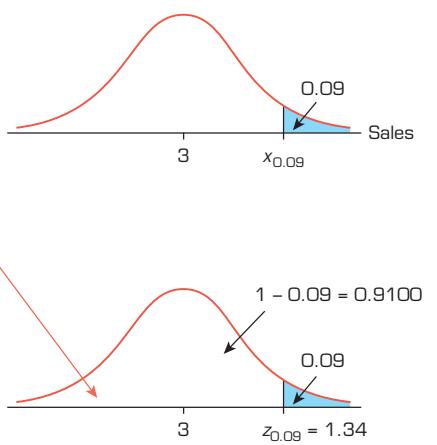
$$P(Z < z_{0.09}) = 0.91$$

In **Figure 8.26**, below the normal curve we depict the standard normal curve and $z_{0.09}$. We can determine the value of $z_{0.09}$ as we did in Example 8.6. In the standard normal table we find $z_{0.09} = 1.34$. Thus, the standardised value of $x_{0.09}$ is $z_{0.09} = 1.34$. To find $x_{0.09}$, we must *unstandardise* $z_{0.09}$. We do so by solving for $x_{0.09}$ in the equation:

$$z_{0.09} = \frac{x_{0.09} - \mu}{\sigma}$$

FIGURE 8.26 Finding $z_{0.09}$ such that $P(Z > z_{0.09}) = 0.09$

<i>z</i>	0.00	0.01	0.02	0.03	0.04
1.0	.8413	.8438	.8461	.8485	.8508
1.1	.8643	.8665	.8686	.8708	.8729
1.2	.8849	.8869	.8888	.8907	.8925
1.3	.9032	.9049	.9066	.9082	.9099
1.4	.9192	.9207	.9222	.9236	.9251
1.5	.9332	.9345	.9357	.9370	.9382
1.6	.9452	.9463	.9474	.9484	.9495
1.7	.9554	.9564	.9573	.9582	.9591
1.8	.9641	.9649	.9656	.9664	.9671
1.9	.9713	.9719	.9726	.9732	.9738
2.0	.9772	.9778	.9783	.9788	.9793



Substituting $z_{0.09} = 1.34$, $\mu = 3$, and $\sigma = 0.5$, we find:

$$1.34 = \frac{x_{0.09} - 3}{0.5}$$

Solving we get:

$$x_{0.09} = 1.34(0.5) + 3 = 3.67$$

We find that the sales level that has only a 9% chance of being exceeded next year is \$3.67m.

EXERCISES

Learning the techniques

8.10 Use Table 3 in Appendix B or NORMSDIST in Excel to find the area under the standard normal curve between the following values:

- a $z = 0$ and $z = 2.3$
- b $z = 0$ and $z = 1.68$
- c $z = 0.24$ and $z = 0.33$
- d $z = -2.75$ and $z = 0$
- e $z = -2.81$ and $z = -1.35$
- f $z = -1.73$ and $z = 0.49$

8.11 Use Table 3 in Appendix B or NORMSDIST in Excel to find the following probabilities:

- a $P(Z \leq -1.96)$
- b $P(Z \leq 2.43)$
- c $P(Z \geq 1.7)$
- d $P(Z \geq -0.95)$
- e $P(-2.97 \leq Z \leq -1.38)$
- f $P(-1.14 \leq Z \leq 1.55)$

8.12 Use Table 3 in Appendix B or NORMSINV in Excel to find the value z_0 for which:

- a $P(Z \leq z_0) = 0.95$
- b $P(Z \leq z_0) = 0.2$
- c $P(Z \geq z_0) = 0.25$
- d $P(Z \geq z_0) = 0.9$
- e $P(0 \leq Z \leq z_0) = 0.41$
- f $P(-z_0 \leq Z \leq z_0) = 0.88$

8.13 Determine $z_{\alpha/2}$, where $z_{\alpha/2}$ is the value of Z such that $P(Z > z_{\alpha/2}) = \alpha/2$ and locate its value on a graph of the standard normal distribution, for each of the following values of α :

- a 0.01
- b 0.02
- c 0.10

8.14 Let X be a normal random variable with a mean of 50 and a standard deviation of 8. Find the following probabilities manually using Table 3 in Appendix B or NORMDIST in Excel:

- a $P(X < 40)$
- b $P(X = 40)$
- c $P(X \geq 52)$
- d $P(X > 40)$
- e $P(35 < X \leq 64)$
- f $P(32 \leq X \leq 37)$

8.15 If X is a normal random variable with a mean of 50 and a standard deviation of 8, how many standard deviations away from the mean is each of the following values of X ?

- a $x = 52$
- b $x = 40$
- c $x = 35$
- d $x = 64$
- e $x = 32$
- f $x = 37$

Applying the techniques

8.16 **Self-correcting exercise.** The time required to assemble an electronic component is normally distributed, with a mean of 12 minutes and a standard deviation of 1.5 minutes. Find the probability that a particular assembly takes:

- a less than 14 minutes
- b less than 10 minutes
- c more than 14 minutes
- d more than 8 minutes
- e between 10 and 14 minutes.

8.17 The lifetime of a certain brand of tyres is approximately normally distributed with a mean of 65000 km and a standard deviation of 2500 km. The tyres carry a warranty for 60000 km.

- a What proportion of the tyres will fail before the warranty expires?
- b What proportion of the tyres will fail after the warranty expires but before they have lasted for 61000 km?

8.18 A marketing manager of a leading firm believes that total sales for the firm next year can be modelled by using a normal distribution with a mean of \$2.5 million and a standard deviation of \$300 000.

- a What is the probability that the firm's sales will exceed \$3 million?
- b What is the probability that the firm's sales will fall within \$150 000 of the expected level of sales?
- c In order to cover fixed costs, the firm's sales must exceed the breakeven level of \$1.8 million. What is the probability that sales will exceed the breakeven level?
- d Determine the sales level that has only a 9% chance of being exceeded next year.

- 8.19** Empirical studies have provided support for the belief that the annual rate of return of an ordinary share is approximately normally distributed. Suppose that you have invested in the shares of a company for which the annual return has an expected value of 16% and a standard deviation of 10%.
- Find the probability that your one-year return will exceed 30%.
 - Find the probability that your one-year return will be negative.
 - Suppose that this company embarks on a new, high-risk, but potentially highly profitable venture. As a result, the return on the share now has an expected value of 25% and a standard deviation of 20%. Answer parts (a) and (b) in light of the revised estimates regarding the share's return.
 - As an investor, would you approve of the company's decision to embark on the new venture?
- 8.20** The maintenance department of a city's electric power company finds that it is cost-efficient to replace all street-light bulbs at once, rather than to replace the bulbs individually as they burn out. Assume that the lifetime of a bulb is normally distributed, with a mean of 3000 hours and a standard deviation of 200 hours.
- If the department wants no more than 1% of the bulbs to burn out before they are replaced, after how many hours should all of the bulbs be replaced?
 - If two bulbs are selected at random from among those that have been replaced, what is the probability that at least one of them has burned out?
- 8.21** Travelbuys is an internet-based travel agency on whose website customers can see videos of the cities they plan to visit. The number of hits daily is a normally distributed random variable with a mean of 10000 and a standard deviation of 2400.
- What is the probability of getting more than 12 000 hits?
 - What is the probability of getting fewer than 9000 hits?
- 8.22** The heights of two-year-old children are normally distributed with a mean of 80 cm and a standard deviation of 3.6cm. Paediatricians regularly measure the heights of toddlers to determine whether there is a problem. There may be a problem when a child is in the top or bottom 5% of heights. Determine the heights of two-year-old children that could be a problem.
- 8.23** Refer to Exercise 8.22. Find the probability of these events.
- A two-year-old child is taller than 90 cm.
 - A two-year-old child is shorter than 85 cm.
 - A two-year-old child is between 75 and 85 cm tall.
- 8.24** University students average 7.2 hours of sleep per night, with a standard deviation of 40 minutes. If the amount of sleep is normally distributed, what proportion of university students sleep for more than 8 hours?
- 8.25** Refer to Exercise 8.24. Find the amount of sleep that is exceeded by only 25% of students.
- 8.26** Battery manufacturers compete on the basis of the amount of time their products last in cameras and toys. A manufacturer of alkaline batteries has observed that its batteries last for an average of 26 hours when used in a toy racing car. The amount of time is normally distributed with a standard deviation of 2.5 hours.
- What is the probability that the battery lasts between 24 and 28 hours?
 - What is the probability that the battery lasts longer than 28 hours?
 - What is the probability that the battery lasts less than 24 hours?
- 8.27** The daily withdrawals from an ATM located at a service station are normally distributed with a mean of \$50 000 and a standard deviation of \$8000. The operator of the ATM puts \$64 000 in cash at the beginning of the day. What is the probability that the ATM will run out of money?
- 8.28** The number of pages printed before replacing the cartridge in a laser printer is normally distributed with a mean of 11 500 pages and a standard deviation of 800 pages. A new cartridge has just been installed.
- What is the probability that the printer produces more than 12 000 pages before this cartridge must be replaced?
 - What is the probability that the printer produces fewer than 10 000 pages?
- 8.29** Refer to Exercise 8.28. The manufacturer wants to provide guidelines to potential customers advising them of the minimum number of pages they can expect from each cartridge. How many pages should

it advertise if the company wants to be correct 99% of the time?

- 8.30** The amount of time devoted to studying statistics each week by students who achieve a grade of A in the course is a normally distributed random variable with a mean of 7.5 hours and a standard deviation of 2.1 hours.
- What proportion of A-grade students study for more than 10 hours per week?
 - Find the probability that an A-grade student spends between 7 and 9 hours studying.
 - What proportion of A-grade students spend less than 3 hours studying?
 - What is the amount of time below which only 5% of all A-grade students spend studying?
- 8.31** It is said that sufferers of a cold virus experience symptoms for 7 days. However, the amount of time is actually a normally distributed random variable with a mean of 7.5 days and a standard deviation of 1.2 days.
- What proportion of cold sufferers experience symptoms for fewer than 4 days?
 - What proportion of cold sufferers experience symptoms for between 7 and 10 days?
- 8.32** How much money does a typical family of four spend at a McDonald's restaurant per visit? The amount is a normally distributed random variable with a mean of \$16.40 and a standard deviation of \$2.75.
- Find the probability that a family of four spends less than \$10.
 - What is the amount spent at McDonald's by less than 10% of families?
- 8.33** The final marks in a statistics course are normally distributed with a mean of 70 and a standard deviation of 10. The lecturer must convert all marks to letter grades such that 10% of the students are to receive an A, 30% a B, 40% a C, 15% a D and 5% an F. Determine the cut-offs for each letter grade.
- 8.34** Mensa is an organisation whose members possess IQs that are in the top 2% of the population. It is known that IQs are normally distributed with a mean of 100 and a standard deviation of 16. Find the minimum IQ needed to be a Mensa member.
- 8.35** A retailer of computing products sells a variety of computer-related products. One of the most popular products is an HP laser printer. The average weekly

demand is for 200 printers. Lead time for a new order to arrive from the manufacturer is one week. If the demand for printers was constant, the retailer would reorder when there were exactly 200 printers in inventory. However, the demand is a random variable. An analysis of previous weeks reveals that the weekly demand standard deviation is 30. The retailer knows that if a customer wants to buy an HP laser printer but she has none available, she will lose that sale as well as possibly additional sales. She wants the probability of running short in any week to be no more than 6%. How many HP laser printers should she have in stock when she reorders from the manufacturer?

- 8.36** The demand for a daily newspaper at a news stand at a busy intersection is known to be normally distributed with a mean of 150 and a standard deviation of 25. How many newspapers should the news stand operator order daily to ensure that he runs short on no more than 20% of days?
- 8.37** Every day a bakery prepares its famous rye bread. The statistically savvy baker determined that daily demand is normally distributed with a mean of 850 and a standard deviation of 90. How many loaves should the bakery bake if the probability of running short on any day is to be no more than 30%?
- 8.38** Refer to Exercise 8.37. Any rye loaves that are unsold at the end of the day are marked down and sold for half price. How many loaves should the baker prepare daily so that the proportion of days on which there are unsold loaves is no more than 60%?
- 8.39** The annual rate of return on a mutual fund is normally distributed with a mean of 14% and a standard deviation of 18%.
- What is the probability that the fund returns more than 25% next year?
 - What is the probability that the fund loses money next year?
- 8.40** In Exercise 7.61(b), we discovered that the expected return is 0.211 and the standard deviation is 0.1064. Working with the assumption that returns are normally distributed, determine the probability of the following events:
- The portfolio loses money.
 - The return on the portfolio is greater than 20%.

8.4 Exponential distribution

exponential distribution

A continuous distribution with probability density function $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$.

Another important continuous distribution is the **exponential distribution**.

8.4a Calculating exponential probabilities

Exponential probability density function

A random variable X is exponentially distributed if its probability density function is given by

$$f(x) = \lambda e^{-\lambda x}, x \geq 0$$

where $e = 2.71828\dots$ and λ is the parameter of the distribution.

exponential random variable

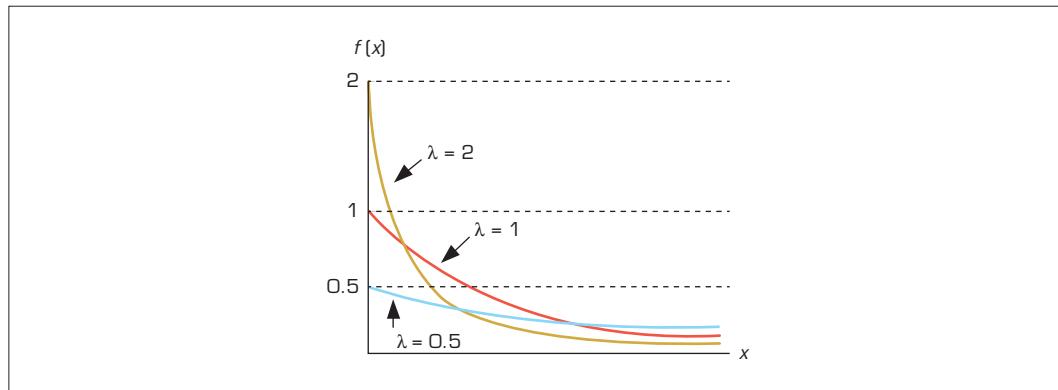
A random variable that is exponentially distributed.

Statisticians have shown that the mean and standard deviation of an **exponential random variable** are equal to each other and given by:

$$\mu = \sigma = \frac{1}{\lambda}$$

Recall that the normal distribution is a two-parameter distribution. The distribution is completely specified once the values of the two parameters are known. In contrast, the exponential distribution is a one-parameter distribution. The distribution is completely specified once the value of the parameter λ is known. **Figure 8.27** depicts three exponential distributions, corresponding to three different values of the parameter, λ . Notice that for any exponential density function $f(x)$, $f(0) = \lambda$ and $f(x)$ approaches 0 as x approaches infinity.

FIGURE 8.27 Graphs of three exponential distributions



The exponential density function is easier to work with than the normal; as a result, we can develop formulas for the calculation of the probability of any ranges of values. Using integral calculus, we can determine the following probability statements.

Probability associated with an exponential random variable

If X is an exponential random variable:

$$P(X > x) = e^{-\lambda x}$$

$$P(X < x) = 1 - e^{-\lambda x}$$

$$P(x_1 < X < x_2) = P(X < x_2) - P(X < x_1) = e^{-\lambda x_1} - e^{-\lambda x_2}$$

The value of $e^{-\lambda x}$ can be obtained with the aid of a calculator or computer.

EXAMPLE 8.10

LO4

Lifetime of alkaline batteries

The lifetime of an alkaline battery (measured in hours) is exponentially distributed with $\lambda = 0.05$.

- a What are the mean and the standard deviation of the battery's lifetime?
- b Find the probability that a battery will last between 10 and 15 hours.
- c What is the probability that a battery will last for more than 20 hours?

Solution

- a The mean and standard deviation are equal to $1/\lambda$. Thus,

$$\mu = \sigma = \frac{1}{\lambda} = \frac{1}{0.05} = 20 \text{ hours}$$

- b Let X denote the lifetime of a battery. The required probability is

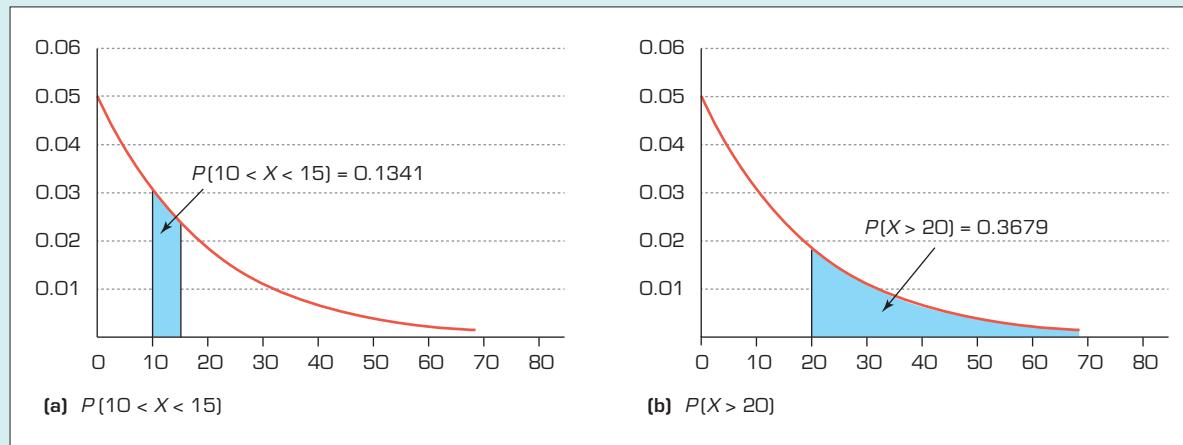
$$\begin{aligned} P(10 \leq X \leq 15) &= e^{-(0.05)(10)} - e^{-(0.05)(15)} \\ &= e^{-0.5} - e^{-0.75} \\ &= 0.6065 - 0.4724 \\ &= 0.1341 \end{aligned}$$

- c The required probability is

$$\begin{aligned} P(X > 20) &= e^{-(0.05)(20)} \\ &= e^{-1} \\ &= 0.3679 \end{aligned}$$

Figure 8.28 depicts these probabilities.

FIGURE 8.28 Probabilities for Example 8.10



EXAMPLE 8.11

LO4

Rate of arrival of cars at the Sydney Harbour tunnel

A toll collector for the Sydney Harbour tunnel has observed that cars arrive randomly and independently at an average rate of 360 cars per hour.

- Use the exponential distribution to find the probability that the next car will *not* arrive within half a minute.
- Use the Poisson distribution to find the probability required in part (a).

Solution

- Let X denote the time *in minutes* that will elapse before the next car arrives. It is important that X and λ be defined in terms of the same units. Thus, λ is the average number of cars arriving per minute: $\lambda = 360/60 = 6$. According to the formula for exponential probabilities, the probability that at least half a minute will elapse before the next car arrives is

$$\begin{aligned} P(X \geq 0.5) &= e^{-\lambda(0.5)} \\ &= e^{-3.0} \\ &= 0.0498 \end{aligned}$$

- Let Y be the number of cars that will arrive in the next half-minute. Then Y is a Poisson random variable, with $\mu = 0.5(\lambda) = 0.5(6) = 3$ cars per half-minute. We wish to find the probability that no cars will arrive within the next half-minute. Using the formula for a Poisson probability, we find

$$P(Y = 0) = \frac{(e^{-3})(3^0)}{0!} = 0.0498$$

Therefore, the probability obtained using the Poisson distribution is the same as that obtained using the exponential distribution.

8.4b Using the computer to find exponential probabilities

Below we provide Excel instructions to allow you to calculate exponential probabilities. The output is the probability that an exponential random variable with a given mean is less than x ; that is, the computer calculates $P(X < x)$.

COMMANDS

To calculate the cumulative probability $P(X \leq x)$ associated with an exponential random variable X with parameter λ , type the following into any active cell:

=EXPONDIST([X],[λ],True)

For Example 8.10c, we would find $P(X < 20)$ and subtract it from 1. To find $P(X < 20)$ with $\lambda = 0.05$, type **=EXPONDIST(20,0.05,True)**, which gives 0.6321 and hence $P(X > 20) = 1 - 0.6321 = 0.3679$, which is exactly the same as the probability value we produced manually.

REAL-LIFE APPLICATIONS**Waiting lines**

In Section 7.7 we described waiting-line models and described how the Poisson distribution is used to calculate the probabilities of the number of arrivals per time period. In order to calculate the operating characteristics of waiting lines, management scientists often assume that the times to complete a service

are exponentially distributed. In this application the parameter is the service rate, which is defined as the mean number of service completions per time period. For example, if service times are exponentially distributed with λ , this tells us that the service rate is 5 units per hour or 5 per 60 minutes. Recall that

the mean of an exponential distribution is $1/\lambda$. In this case, the service facility can complete a service in an average of 12 minutes. This was calculated as

$$\mu = \frac{1}{\lambda} = \frac{1}{5/h} = \frac{1}{5/60 \text{ min}} = \frac{60 \text{ min}}{5} = 12 \text{ min}$$

We can use this distribution to make a variety of probability statements.



Source: Dreamstime.com/Verdeleho

EXAMPLE 8.12

L04

Service rate at supermarket checkout counter

A checkout counter at a supermarket completes the process according to an exponential distribution with a service rate of six per hour. A customer arrives at the checkout counter. Find the probability of the following events.

- a The service is completed in less than 5 minutes.
- b The customer leaves the checkout counter more than 10 minutes after arriving.
- c The service is completed in a time between 5 and 8 minutes.

Solution

Let X be the time taken (in minutes) to complete the service. Note that service times in parts (a)–(c) are stated in minutes but the service rate is given per hour. One way to solve this problem is to convert the service rate so that the time period is 1 minute. (Alternatively, we can solve this by converting the probability statements so that the time periods are measured in fractions of an hour.) Let the service rate = $\lambda = 6$ per hour = $\frac{6}{60}$ per min = 0.1 per minute.

- a $P(X < 5) = 1 - e^{-\lambda x} = 1 - e^{-(0.1)(5)} = 1 - 0.6065 = 0.3935$
- b $P(X > 10) = e^{-\lambda x} = e^{-(0.1)(10)} = e^{-1} = 0.3679$
- c $P(5 < X < 8) = e^{-(0.1)(5)} - e^{-(0.1)(8)} = e^{-0.5} - e^{-0.8} = 0.6065 - 0.4493 = 0.1572$

EXERCISES

Learning the techniques

8.41 X is an exponential random variable with $\lambda = 1$.

Sketch the graph of the distribution of X by plotting and connecting the points representing $f(x)$ for $x = 0, 0.5, 1.0, 1.5$ and 2.0 .

8.42 X is an exponential random variable with $\lambda = 0.25$.

Sketch the graph of the distribution of X by plotting and connecting the points representing $f(x)$ for $x = 0, 2, 4, 6, 8, 10, 15, 20$.

8.43 Let X be an exponential random variable with $\lambda = 0.5$.

Find the following probabilities:

- a** $P(X > 1)$
- b** $P(X > 0.4)$
- c** $P(X < 0.5)$
- d** $P(X < 2)$

8.44 Let X be an exponential random variable with $\lambda = 3$.

Find the following probabilities:

- a** $P(X \geq 2)$
- b** $P(X \leq 4)$
- c** $P(1 \leq X \leq 3)$
- d** $P(X = 2)$

Applying the techniques

8.45 Self-correcting exercise. Suppose that customers arrived at a checkout counter at an average rate of two customers per minute, and that their arrivals follow the Poisson model.

a Sketch a graph of the (exponential) distribution of the time that will elapse before the next customer arrives by plotting and joining the points representing $f(t)$ for $t = 0, 0.5, 1.0, 1.5$ and 2.0 .

b Use the appropriate exponential distribution to find the probability that the next customer will arrive within (i) 1 minute, (ii) 2 minutes.

c Use the exponential distribution to find the probability that the next customer will not arrive within the next 1.5 minutes.

d Use the appropriate Poisson distribution to answer part (c).

8.46 A bank wishing to increase its customer base advertises that it has the fastest service and that virtually all of its customers are served in less than 10 minutes. A management scientist has studied the service times and concluded that service times are exponentially distributed with a mean of 5 minutes. Determine what the bank means when it claims 'virtually all' its customers are served in less than 10 minutes.

8.47 The time between breakdowns of ageing machines is known to be exponentially distributed with a mean of 25 hours. The machine has just been repaired. Determine the probability that the next breakdown occurs more than 50 hours from now.

8.48 A firm has monitored the duration of long-distance telephone calls placed by its employees, to help it decide which long-distance call package to purchase. The duration of calls was found to be exponentially distributed with a mean of 5 minutes.

- a** What proportion of calls last more than 2 minutes?
- b** What proportion of calls last more than 5 minutes?
- c** What proportion of calls are shorter than 10 minutes?

Study Tools

CHAPTER SUMMARY

This chapter dealt with *continuous random variables* and their distributions. Since a continuous random variable can assume an infinite number of values, the probability that the random variable equals any single value is 0. Consequently, we address the problem of computing the probability of a range of values. We showed that the probability of any interval is the area in the interval under the curve representing the *density function*.

We introduced the most important distribution in statistics, the normal distribution, and showed how to compute the probability that a *normal random variable* falls into any interval.

Additionally, we demonstrated how to use the normal table backwards to find values of a normal random variable given a probability. Next we introduced the *exponential distribution*, a distribution that is particularly useful in several management science applications, such as waiting lines and queuing.

The data files for Examples, Exercises and Cases are provided on the companion website (accessible through <https://login.cengagebrain.com/>).

SYMBOLS

Symbol	Pronounced	Represents
e		2.71828 ...
π	<i>pi</i>	3.14159 ...
z_A	z -sub- A or z - A	Value of Z such that area to its right is A

SUMMARY OF FORMULAS

Standardised normal random variable	$Z = \frac{X - \mu}{\sigma}$
Mean	$E(Z) = 0$
Variance	$V(Z) = 1$
Exponential	
Probability	$P(X \geq a) = e^{-\lambda a}$
Mean	$E(X) = \mu = 1/\lambda$
Variance	$V(X) = \sigma^2 = 1/\lambda^2$

SUPPLEMENTARY EXERCISES

- 8.49** Use Table 3 in Appendix B or NORMSDIST in Excel to find the following probabilities:

- a $P(Z < 0.52)$
- b $P(Z > 1.64)$
- c $P(1.23 < Z < 2.71)$
- d $P(-0.68 < Z < 2.42)$

- 8.50** Use Table 3 in Appendix B or NORMDIST in Excel to find the following probabilities, where X has a normal distribution with $\mu = 24$ and $\sigma = 4$:

- a $P(X \leq 26)$
- b $P(X > 30)$
- c $P(25 < X < 27)$
- d $P(18 \leq X \leq 23)$

- 8.51** Suppose that the actual amount of instant coffee that a filling machine puts into 250g jars varies from jar to jar, and that the actual fill may be considered a random variable having a normal distribution, with a standard deviation of 1g. If only two out of every 100 jars contain less than 250g of coffee, what must be the mean fill of these jars?

- 8.52** A soft-drink bottling plant in Perth uses a machine that fills bottles with drink mixture. The contents of the bottles filled are normally distributed, with a mean of 1 L and a standard deviation of 10mL.
- a Determine the volume exceeded by only 10% of the filled bottles.

- b** Determine the probability that the combined volume of two of these bottles is less than 1.8L. (*Hint:* If X_1 and X_2 are normally distributed variables, then $Y = X_1 + X_2$ is also normally distributed.)
- 8.53** Consumer advocates frequently complain about the large variation in the prices charged by different chemists for the same prescription. A survey of chemists by one such advocate revealed that the prices charged for 100 tablets of a drug were normally distributed, with about 90% of the prices ranging between \$8.25 and \$11.25. The mean price charged was \$9.75. What proportion of the pharmacies charged more than \$10.25 for the prescription?
- 8.54** Suppose that the heights of men are normally distributed, with a mean of 175cm and a standard deviation of 5cm. Find the minimum ceiling height of an aeroplane in which, at most, 2% of the men walking down the aisle will have to duck their heads.
- 8.55** Economists frequently make use of quintiles (i.e., the 20th, 40th, 60th and 80th percentiles) particularly when discussing incomes. Suppose that in a large city household incomes are normally distributed with a mean of \$50000 and a standard deviation of \$10000. An economist wishes to identify the quintiles. Unfortunately, he did not pass his statistics course. Help him by providing the quintiles.
- 8.56** A university has just approved a new executive MBA program. The director of the program believes that in order to maintain the prestigious image of the business school, the new program must be seen as having high standards. Accordingly, the Faculty Board decides that one of the entrance requirements will be that applicants must sit an admissions test very similar to the well-known GMAT (Graduate Management Admissions Test) in the US, and score in the top 1% of the scores. The director knows that the GMAT scores in US universities are normally distributed with a mean of 490 and a standard deviation of 61. Using this information, the director would like to know what is the minimum entry score for the executive MBA program.
- 8.57** The manager of a petrol station has observed that the times required by drivers to fill their car's tank and to pay are quite variable. In fact, the times are exponentially distributed with a mean of 7.5 minutes. What is the probability that a driver can complete the transaction in less than 5 minutes?
- 8.58** Because automatic teller machine (ATM) customers can perform a number of transactions, the times to complete them can be quite variable. A banking consultant has noted that the times are exponentially distributed with a mean of 125 seconds. What proportion of the ATM customers take more than 3 minutes to do their banking?

Case Studies

CASE 8.1 Average salary of popular business professions in Australia

C08-01 Based on a recent publication (<https://www.payscale.com/salary-calculator>), the average salaries of 10 popular business-related jobs in Australia are listed below. It is believed that salaries can be considered as following a normal distribution with a standard deviation of \$1500. A high school student would like to choose a degree that could lead to the profession which has a higher probability of gaining a reasonably good salary. What is the likelihood of him receiving an annual salary greater than \$60 000 for each of the 10 jobs listed in the table?

Average salary in Australia, 10 popular business jobs

Job title	Average salary (\$)
Accountant	57139
Business Data Analyst	69823
Finance Manager	93835
Financial Planner	76004
HR Manager	89328
Marketing Manager	78847
Personal Banker	53285
Retail Manager	51413
Supply Chain Manager	103724
Tax Accountant	56316

Source: <https://www.payscale.com/salary-calculator>

CASE 8.2 Fuel consumption of popular brands of motor vehicles

C08-02 With ever-increasing fuel prices, motorists are looking for cars with the most fuel efficiency; that is, low fuel consumption per kilometre. The fuel consumption of different brands of passenger cars are listed in the table. A car salesperson conveys to the purchaser that the standard deviation of fuel consumption is 3 litres per 100 km. What is the probability that a randomly selected car from each of the 6 brands will use less than 8 litres per 100 km? Which car has the highest probability of achieving this?

Vehicle name	Litres/100km
Ford Fiesta 1.0 Ecoboost start	8.0
Honda CR-V Turbo 4WD	7.0
Kia Sportage	7.8
Mazda CX-3	7.0
Nissan Navara Double Cab auto	8.2
Toyota Auris 1.2 Comfort	7.6

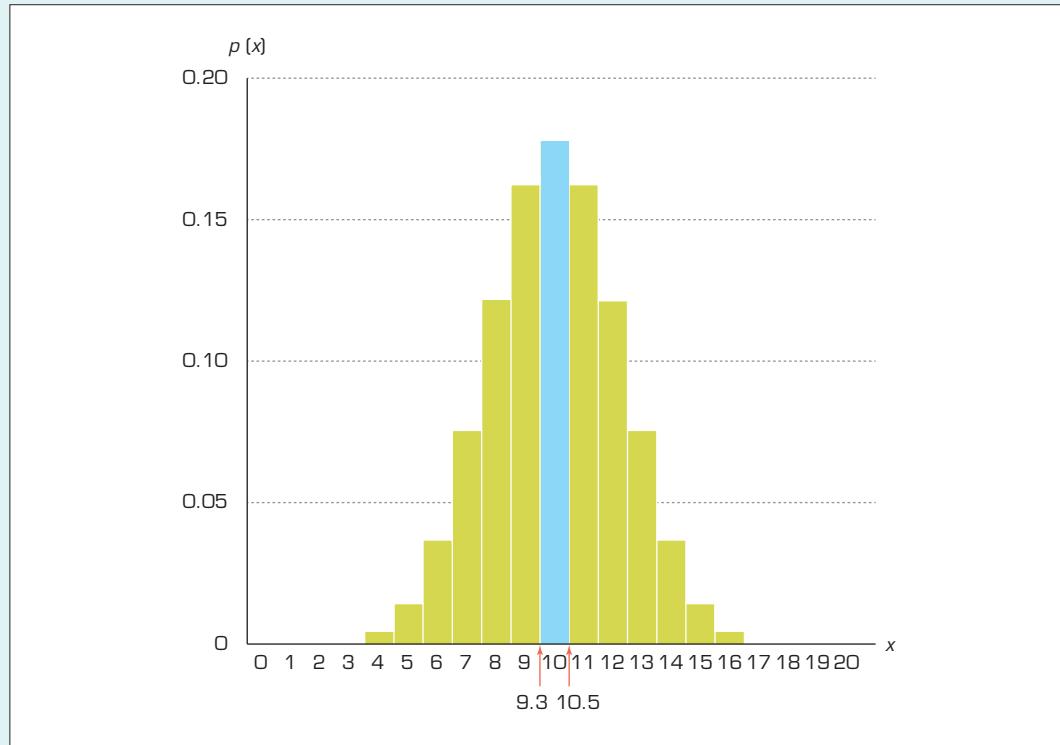
Appendix 8.A

Normal approximation to the binomial distribution

L05 Approximate binomial probabilities using a normal distribution

Recall that we introduced continuous probability distributions in Section 8.1. We developed the density function by converting a histogram so that the total area in the rectangles equalled 1. We can do the same for a binomial distribution. To illustrate, let X be a binomial random variable with $n = 20$ and $p = 0.5$. We can easily determine the probability of each value of X , where $x = 0, 1, 2, \dots, 19, 20$. A rectangle representing a value of x is drawn so that its area equals the probability. We accomplish this by letting the height of the rectangle equal the probability and the base of the rectangle equal 1. Thus the base of each rectangle for x is the interval from $x - 0.5$ to $x + 0.5$. **Figure A8.1** depicts this graph. As you can see, the rectangle representing $x = 10$ is the rectangle whose base is the interval 9.5 to 10.5 and whose height is $P(X = 10) = 0.176$.

FIGURE A8.1 Binomial distribution with $n = 20$ and $p = 0.5$



If we now smooth the ends of the rectangles, we produce a bell-shaped curve as seen in Figure A8.2. Thus, to use the normal approximation, all we need do is find the area under the normal curve between 9.5 and 10.5. To find normal probabilities requires us to first standardise X by subtracting the mean μ and dividing by the standard deviation σ . The values for μ and σ are derived from the binomial distribution being approximated. In Section 7.6 we pointed out that given a binomial distribution with n trials and probability p of a success on any trial, the mean and the standard deviation are

$$\mu = np$$

and

$$\sigma = \sqrt{np(1-p)}$$

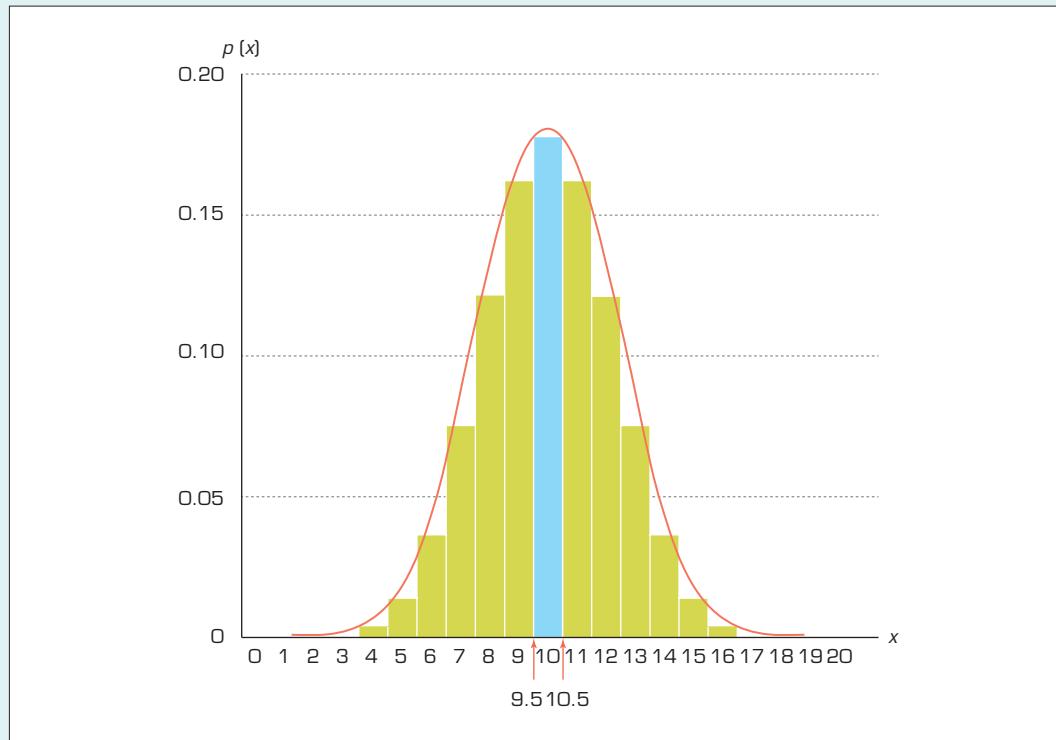
For $n = 20$ and $p = 0.5$, we have

$$\mu = np = 20(0.5) = 10$$

and

$$\sigma = \sqrt{np(1-p)} = \sqrt{20(0.5)(1-0.5)} = 2.24$$

FIGURE A8.2 Binomial distribution with $n = 20$ and $p = 0.5$ and normal approximation



To calculate the binomial probability $P(X=10)$ using the normal distribution as an approximation requires that we find the area under the normal curve between 9.5 and 10.5. That is

$$P(X=10) \approx P(9.5 < Y < 10.5)$$

where X is a binomial random variable with $n = 20$ and $p = 0.5$ and Y is a normal random variable approximating the binomial random variable X (with mean $\mu = 10$ and standard deviation $\sigma = 2.24$).

We standardise Y and use Table 3 in Appendix B to find

$$\begin{aligned} P(9.5 < Y < 10.5) &= P\left(\frac{9.5-10}{2.24} < \frac{Y-10}{2.24} < \frac{10.5-10}{2.24}\right) \\ &= P(-0.22 < Z < 0.22) \\ &= P(Z < 0.22) - P(Z < -0.22) \\ &= 0.5871 - 0.4129 \\ &= 0.1742 \end{aligned}$$

The actual probability that X equals 10 using the binomial table (Table 1 in Appendix B) is

$$P(X = 10) = P(X \leq 10) - P(X \leq 9) = 0.588 - 0.412 = 0.176$$

As you can see, the approximation is quite good. This approximation, in general, works best if $np \geq 5$ and $nq \geq 5$. In this example, note that we have $np = 20(0.5) = 10 > 5$ and $nq = 20(0.5) = 10 > 5$.

8.Aa Continuity correction factor

Notice that to draw a binomial distribution that is discrete, it was necessary to draw rectangles whose bases were constructed by adding and subtracting 0.5 to the values of X . The 0.5 is called the **continuity correction factor**.

continuity correction factor
A correction factor that allows for the approximation of a discrete random variable by a continuous random variable.

The approximation for any other value of X would proceed in the same manner. In general, the binomial probability $P(X = x)$ is approximated by the area under a normal curve between $x - 0.5$ and $x + 0.5$. To find the binomial probability $P(X \leq x)$, we calculate the area under the normal curve to the left of $x + 0.5$. For the same binomial random variable, the probability that its value is less than or equal to 8 is $P(X \leq 8) = 0.252$.

Since the binomial random variable X can take only values between 0 and 20, the normal approximation for $P(X \leq 8)$ is

$$\begin{aligned} P(X \leq 8) &= P(0 \leq X \leq 8) \approx P(-0.5 \leq Y < 8.5) \\ &= P\left(\frac{-0.5-10}{2.24} < \frac{Y-10}{2.24} < \frac{8.5-10}{2.24}\right) \\ &= P(-4.69 < Z < -0.67) \\ &= P(Z < -0.67) - P(Z < -4.69) \\ &= 0.2514 - 0 \\ &= 0.2514 \end{aligned}$$

We find the area under the normal curve to the right of $x - 0.5$ to determine the binomial probability $P(X \geq x)$. To illustrate, the probability that the binomial random variable (with $n = 20$ and $p = 0.5$) is greater than or equal to 14 using a binomial table is

$$P(X \geq 14) = 1 - P(X \leq 13) = 0.058$$

The normal approximation is

$$\begin{aligned} P(X \geq 14) &= P(14 \leq X \leq 20) \approx P(13.5 \leq Y < 20.5) \\ &= P\left(\frac{13.5-10}{2.24} < \frac{Y-10}{2.24} < \frac{20.5-10}{2.24}\right) \\ &= P(1.56 < Z < 4.69) \\ &= P(Z < 4.69) - P(Z < 1.56) \\ &= 1 - 0.9406 \\ &= 0.0594 \end{aligned}$$

8.Ab Omitting the correction factor for continuity

When calculating the probability of individual values of X , as we did when we computed the probability that X equals 10 earlier, the correction factor must be used. If we don't, we are left with finding the area of a line, which is 0. When computing the probability of a range of values of X , we can omit the correction factor. However, the omission of the correction factor will

decrease the accuracy of the approximation. For example, if we approximate $P(X \leq 8)$, as we did previously except without the correction factor, we find

$$P(X \leq 8) \approx P(Y < 8) = P\left(\frac{Y - \mu}{\sigma} < \frac{8 - 10}{2.24}\right) = P(Z < -0.89) = 0.1867$$

The absolute size of the error between the actual cumulative binomial probability and its normal approximation is quite small when the values of x are in the tail regions of the distribution. For example, the probability that a binomial random variable with $n = 20$ and $p = 0.5$ ($\mu = np = 10$; $\sigma^2 = npq = 5$) is less than or equal to 3 is

$$P(X \leq 3) = 0.0013$$

The normal approximation with the correction factor is

$$P(X \leq 3) \approx P(Y < 3.5) = P\left(\frac{Y - \mu}{\sigma} < \frac{3.5 - 10}{2.24}\right) = P(Z < -2.90) = 0.0019$$

The normal approximation without the correction factor (using Excel) is

$$P(X \leq 3) \approx P(Y < 3) = P\left(\frac{Y - \mu}{\sigma} < \frac{3 - 10}{2.24}\right) = P(Z < -3.13) = 0.0009$$

For larger values of n , the differences between the normal approximation with and without the correction factor are small even for values of X near the centre of the distribution. For example, using Excel, the probability that a binomial random variable X with $n = 1000$ and $p = 0.3$ ($\mu = np = 300$; $\sigma^2 = npq = 210$) is less than or equal to 260 is

$$P(X \leq 260) = 0.0029$$

The normal approximation with the correction factor is

$$P(X \leq 260) \approx P(Y < 260.5) = P\left(\frac{Y - \mu}{\sigma} < \frac{260.5 - 300}{14.49}\right) = P(Z < -2.73) = 0.0032$$

The normal approximation without the correction factor is

$$P(X \leq 260) \approx P(Y < 260) = P\left(\frac{Y - \mu}{\sigma} < \frac{260 - 300}{14.49}\right) = P(Z < -2.76) = 0.0029$$

As we pointed out, the normal approximation of the binomial distribution is made necessary by the needs of statistical inference. As you will discover, statistical inference generally involves the use of large values of n , and the part of the sampling distribution that is of greatest interest lies in the tail regions. The correction factor was a temporary tool that allowed us to convince you that a binomial distribution can be approximated by a normal distribution. Now that we have done so, we will use the normal approximation of the binomial distribution to approximate the sampling distribution of a sample proportion in the next chapter, and in such applications the correction factor will be omitted.

EXERCISES

A8.1 Let X be a binomial random variable with $n = 100$ and $p = 0.6$. Approximate the following probabilities, using the normal distribution:

- a $P(X = 65)$
- b $P(X \leq 70)$
- c $P(X > 50)$

A8.2 Companies are interested in the demographics of those who listen to the radio programs they sponsor. A radio station has determined that only 20% of listeners phoning into a morning talk-back program are male. During a particular week, 200 calls are received by this program.

- a What is the probability that at least 50 of these 200 callers are male?
- b What is the probability that more than half of these 200 callers are female?

A8.3 Due to an increasing number of non-performing loans, a bank now insists that several stringent conditions be met before a customer is granted a consumer loan. As a result, 60% of all customers

applying for a loan are *rejected*. If 40 new loan applications are selected at random, what is the probability that:

- a at least 12 are accepted?
- b at least half of them are accepted?
- c no more than 16 are accepted?
- d the number of applications rejected is between 20 and 30, inclusive?

A8.4 Suppose that X is a binomial random variable with $n = 100$ and $p = 0.20$. Use the normal approximation to find the probability that X takes a value between 22 and 25 (inclusive).

A8.5 Venture-capital firms provide financing for small, high-risk enterprises that have the potential to become highly profitable. A successful venture-capital firm notes that it provides financing for only 10% of the proposals it reviews. Of the 200 proposals submitted this year, what is the probability that more than 30 will receive financing?

PART TWO

Statistical inference

CHAPTER 9	Statistical inference and sampling distributions
CHAPTER 10	Estimation: Single population
CHAPTER 11	Estimation: Two populations
CHAPTER 12	Hypothesis testing: Single population
CHAPTER 13	Hypothesis testing: Two populations
CHAPTER 14	Chi-squared tests
CHAPTER 15	Simple linear regression and correlation
CHAPTER 16	Multiple regression

In Part 1, we developed the critical foundation for statistical inference. We introduced descriptive techniques and probability. In the rest of the book we will use these components to develop statistical inference.

Over the next 8 chapters we will present a variety of statistical methods that involve some form of inference. These techniques deal with different types of data and the different kinds of information that we wish to extract from the data. All of these techniques have been proven to be useful to managers, economists and decision makers.

Although these techniques differ widely in the arithmetic needed to produce the results, they are very similar conceptually. In fact, they are so similar that students often encounter difficulty in deciding which technique to use. We will spend a considerable amount of time attempting to ease this difficulty. A major emphasis in this book is on developing technique-recognition skills. One review chapter is provided to assist you in this development.

Statistical inference and sampling distributions

Learning objectives

This chapter presents an introduction to statistical inference, and links numerical descriptive statistics (Chapter 5) and the probability distributions (Chapters 7 and 8) to statistical inference.

At the completion of this chapter, you should be able to:

- L01** explain the importance of statistical inference
- L02** explain how, when and why statistical inference is used
- L03** understand the central limit theorem and the properties of the sampling distribution of the sample mean
- L04** understand the properties of the sampling distribution of the sample proportion
- L05** understand the probability link between sample statistics and population parameters
- L06** apply various sampling distributions in practical applications.

CHAPTER OUTLINE

- Introduction
- 9.1** Data type and problem objective
- 9.2** Systematic approach to statistical inference: A summary
- 9.3** Introduction to sampling distribution
- 9.4** Sampling distribution of the sample mean \bar{X}
- 9.5** Sampling distribution of the sample proportion \hat{p}
- 9.6** From here to inference

SPOTLIGHT ON STATISTICS

Salaries of a business school's MBA graduates

Deans of professional schools within universities often monitor how well the graduates of their programs fare in the job market. Information about the types of jobs graduates secure and their salaries provides a useful gauge of the success of the program.

In advertisements for a large university, the dean of the School of Business claims that the average salary of the school's MBA graduates one year after graduation is \$2000 per week with a standard deviation of \$200. A second-year Business student who has just completed his statistics course would like to check whether the claim is correct. He surveys 25 MBA graduates of



Source: © iStock.com/Alexandr Dubovitskiy

the School of Business who graduated one year ago and determines their weekly salary. He discovers that the sample mean is \$1900. To interpret his finding he needs to calculate the probability that a sample of 25 MBA graduates would have a mean of \$1900 or less when the population mean is \$2000 with a standard deviation is \$200. After calculating the probability, he needs to draw some conclusions. (See pages 363–4 for the answer.)

Introduction

Much of the remainder of this book deals with problems that attempt to say something about the properties of a population. Because populations are generally quite large, the information we usually have available to work with comes from a relatively small sample taken from the population. The process of drawing conclusions about the properties of a population (parameter) based on information obtained from a sample (statistic) is called statistical inference (see Chapter 1).

Examples of statistical inference include:

- estimating the mean monthly expenditure on electricity of households living in houses with solar power
- determining whether the introduction of plain packaging of cigarettes in 2011 has reduced the proportion of smokers in Australia
- determining whether the value of the New Zealand dollar influences the price of petrol in New Zealand
- forecasting the average monthly sales of air conditioners for next year.

In many applications of statistical inference, we draw conclusions about a parameter of a population by using a relevant sample statistic.

In the course of this book, we present a number of different statistical techniques that will be useful to business managers and economists. You will find that the arithmetic needed for each method is quite simple; the only mathematical operations required are addition, subtraction, multiplication, division, and the calculation of squares and square roots. Even these skills may be less in demand if you use a calculator to do much of the work; and if you use a computer, almost no mathematics is needed. In fact, because of the availability of inexpensive computers and software, many students find that they do very few calculations manually. This is certainly true in real-life (defined as anything outside a university or TAFE college) applications of statistics.

In this chapter, we discuss the basic concepts of identifying the data type and problem objective as well as the sampling distribution of a random variable. We also discuss the sampling distribution of the sample mean and sample proportion, and calculate probabilities involving these two random variables. Finally, we link the sampling distributions to statistical inference, which is the focus of most of the remaining chapters of this book.

9.1 Data type and problem objective

The real challenge of the subject of statistics relates to an individual's ability to determine which technique is the most appropriate one to use in answering a given question. Most students who are taking their first course in statistics have some difficulty in recognising the particular kind of statistical problem involved in practice exercises – and hence the appropriate statistical technique to use. This difficulty intensifies when you must apply statistical techniques to practical real-life problems in which the questions to be addressed may themselves be vague and ill-defined. In this book, most of the

exercises and examples depict situations in which the data have already been gathered, and your task is simply to answer a specific question by applying one of the statistical techniques you have studied. In a real-life situation, you will probably have to define the statistical problem, design the experiment, collect the data, and perform and interpret the statistical calculations yourself. The difficulty of determining what to do can be formidable. Because people encounter such difficulty both during and after the study of this subject, we have adopted a systematic approach that is designed to help you identify the statistical problem and the appropriate technique to use.

A number of factors determine which statistical technique should be used, but two of these are especially important: the *type of data* (numerical, nominal or ordinal, see Chapter 2) being measured and the purpose of the statistical inference.

Another key factor in determining the appropriate statistical technique to use is the purpose behind the work. You will find that every statistical method has some specific *objective*. We will now identify and describe four such objectives.

1 Description of a single population

We wish to describe some characteristic of a population of interest. For example, suppose we want to summarise the weekly income of all Australians. The data type in this case is numerical and the problem objective is to estimate the average weekly income of the Australian population.

2 Comparison of two populations

In this case, our goal is to compare a characteristic of one population with the corresponding characteristic of a second population. For example, suppose we wish to compare the average weekly income (in US\$) of Australians and New Zealanders. The type of data is numerical and our problem objective is to compare the mean weekly incomes of two populations.

3 Comparison of more than two populations

Our aim here is to compare the average or the variance of two or more populations. For example, we wish to compare the average weekly income (in US\$) of Australians, New Zealanders and the British. Our data type is numerical and the problem objective involves comparing the mean income of three populations.

4 Analysis of the relationships between two or more variables

Suppose our aim is to know how one variable is related to a number of other variables; that is, forecast one variable (called the *dependent variable*) on the basis of several other variables (called *independent variables*). For example, we want to investigate the effect of family income and family size on a family's expenditure on food. Here the data are numerical and our problem objective is to analyse the relationship between a family's expenditure on food and the family income and family size.

9.2 Systematic approach to statistical inference: A summary

The most difficult issue for students is to determine *when* to apply each technique. Using our systematic approach, in which we detail all the required conditions that must be satisfied in using a method, should alleviate this difficulty.

We propose to deal with the *why* issue in several ways. In most of the worked examples, we set up the problem in a decision context; and even though some are quite simplistic, they should give you some idea of the motivation for using statistics. Many of the exercises also stress the reason for the application. We acknowledge that these reasons are frequently simplified; but as we progress, the assumptions become much more reasonable, and problems that in practice involve the use of several methods can be addressed.

In Chapters 10–16, we develop about 20 statistical techniques, each of which will be identified by problem objective and data type. **Table 9.1** shows the four problem objectives and the two types of data. For each combination of objective and data type, one or more techniques are used to answer questions, and **Table 9.1** identifies the chapter and section in which these techniques are described.

Where possible, we will group the statistical techniques according to their common problem objectives. Because of similarities in some of the techniques, however, this order of presentation cannot always be strictly adhered to. **Table 9.1** should help you to keep track of the order of presentation.

TABLE 9.1 Guide to statistical techniques, showing chapters and sections in which each technique is introduced

Problem objective	Numerical data	Nominal data
Description of a single population	10.2, 10.3, 12.2, 12.4, 14.1	10.4, 12.6, 14.1
Comparison of two populations	11.1, 11.2, 11.3, 13.1, 13.2, 14.2, 14.3	11.4, 13.3, 14.2
Analysis of the relationship between two or more variables	Chs. 15, 16	14.2

9.2a A three-stage solution process

The solution process that we advocate and use to solve problems in statistical inference throughout this book is, by and large, the same one that statistics practitioners use to apply their skills in the real world. The process is divided into three stages. Simply stated, the stages are (1) Identify – the activities we perform before the calculations, (2) Calculate – the calculations, and (3) Interpret – the activities we perform after the calculations.

- **Stage 1: Identify – The activities we perform before the calculations.** We determine the appropriate statistical technique to employ. This stage also addresses the problem of how to gather the data.
- **Stage 2: Calculate – Perform the calculations.** We calculate the statistics. We do this in three ways. To illustrate how the computations are completed, we do the arithmetic manually with the assistance of a calculator. Solving problems manually often provides insights into the statistical inference technique. However, at some point in our journey of discovery of statistical inference, the arithmetic becomes so tedious that we use the computer exclusively.
- The second method is a combination of the Analysis ToolPak (Data menu item *Data Analysis* that is part of Microsoft Excel) and the workbooks that we created. This combination will allow us to compute most (but not all) of the inferential techniques introduced in this book. The rest will have to be done by additional software.
- The third method uses XLSTAT, which is a commercial software add-in. XLSTAT calculates almost all of the techniques covered in this book with the exception of forecasting (Chapter 17). Readers who need these topics in Chapter 17 can use Data Analysis Plus,¹ which can be downloaded from the textbook website; see Section 1.5 Online Resources for further details.
- **Stage 3: Interpret – The activities we perform after the calculations.** We interpret the results and make inferences about the question presented in the problem. To be capable of properly interpreting statistical results, we need to have an understanding of the fundamental principles underlying statistical inference.

¹ Data Analysis Plus is an Excel Add-in, which was available in previous editions of this textbook. It is a collection of macros created to augment Excel's list of statistical procedures provided by Data Analysis. This can be downloaded from the link provided in the textbook website.

EXERCISES

- 9.1** For each of the following, identify the problem objective.
- The Tasmanian department of transport wants to estimate the proportion of school children who use public transport to school. Analysts record the mode of travel to school of a random selection of school children from various private and public schools.
 - A firm in the coal industry wants to estimate the average number of days of sick leave taken annually by its employees. Analysts determine the annual sick days leave taken by a random selection of employees in the coal industry for the past 10 years.
 - A firm wants to determine whether increasing the advertising budget will result in an increase in sales volume. Analysts determine the monthly advertising expenses and the monthly sales volume for the past 12 months.
 - During a recent economic downturn, a number of workers in several large industries had to work reduced hours. An economist wants to determine whether there were differences in the average number of hours of work per week among five industries.
 - In order to design advertising campaigns, a marketing manager needs to know whether different segments of the population prefer her company's product to competing products. She decides to perform a survey that will determine whether different proportions of people in five separate age categories purchase the product.
 - The marketing manager in part (e) also wants to know whether the proportion of men purchasing the product is different from the proportion of women purchasing the product.
 - The production manager of a large plant is contemplating changing the process by which a certain product is produced. As workers in this plant are paid on the basis of their output, it is essential to demonstrate that the rate of assembling a unit will increase under the new system. Ten workers are randomly selected to participate in an experiment in which each worker assembles one unit under the old process and one unit under the new process.

9.3 Introduction to sampling distribution

This and the following sections introduce the sampling distribution, a fundamental element in statistical inference. As discussed in Section 9.1, statistical inference is the process of converting sample data into information to make inferences about the population parameter. Here are the parts of the process we have discussed so far:

- Parameters describe populations.
- Parameters are almost always unknown.
- We take a random sample from the population of interest to obtain the necessary data.
- We calculate one or more statistics from the sample data.

For example, to estimate a population mean, we compute the sample mean. Although there is very little chance that the sample mean and the population mean are identical, we would expect them to be quite close. However, for the purposes of statistical inference, we need to be able to measure how close. We do this by finding the sampling distribution. The sampling distribution plays a crucial role in the process, because the measure of proximity it provides is the key to statistical inference.

sampling distribution of the sample mean

A relative frequency distribution of various values of the sample mean obtained from a number of different samples selected from the same population.

9.4 Sampling distribution of the sample mean \bar{X}

A **sampling distribution of the sample mean** is created by, as the name suggests, sampling. There are two ways to create a sampling distribution. The first is to actually draw samples of the same size from a population, calculate the statistic of interest, and then use descriptive

techniques to learn more about the sampling distribution. The second method relies on the rules of probability and the laws of expected value and variance to derive the sampling distribution. We'll demonstrate the latter approach by developing the sampling distribution of the mean of the throw of two dice.

9.4a Sampling distribution of the sample mean \bar{X}

Consider the population created by throwing a fair die infinitely many times, with the random variable X indicating the number of spots showing on any one throw. The probability distribution of the random variable X is as follows:

x	1	2	3	4	5	6
$p(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

The population is infinitely large, since we can throw the die infinitely many times. From the definitions of expectation and variance presented in Section 7.3, we calculate the population mean, variance and standard deviation.

Population mean:

$$\begin{aligned}\mu &= E(X) \\ &= \sum x p(x) \\ &= 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + \dots + 6\left(\frac{1}{6}\right) \\ &= 3.5\end{aligned}$$

Population variance:

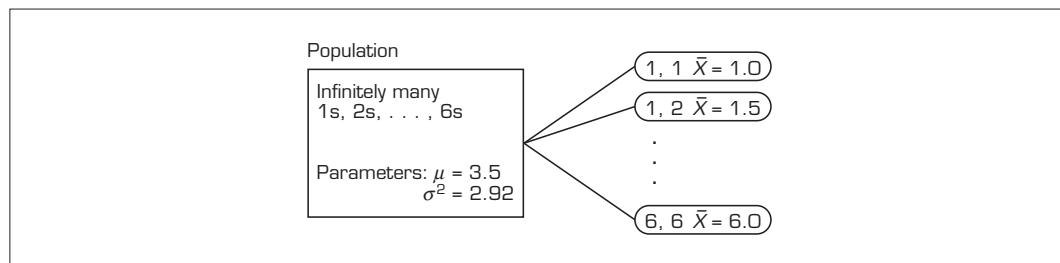
$$\begin{aligned}\sigma^2 &= V(X) = E(X^2) - \mu^2 \\ &= \sum x^2 p(x) - \mu^2 \\ &= \left[(1)^2 \left(\frac{1}{6}\right) + (2)^2 \left(\frac{1}{6}\right) + \dots + (6)^2 \left(\frac{1}{6}\right) \right] - (3.5)^2 \\ &= 2.92\end{aligned}$$

Population standard deviation:

$$\begin{aligned}\sigma &= \sqrt{V(X)} \\ &= \sqrt{2.92} \\ &= 1.71\end{aligned}$$

Now pretend that μ is not known and that we want to estimate its value by using the sample mean \bar{X} , calculated from a sample of size $n = 2$. In actual practice, only one sample would be drawn, and hence there would be only one value of \bar{X} ; but in order to assess how closely \bar{X} estimates the value of μ , we will develop the sampling distribution of \bar{X} by evaluating every possible sample of size $n = 2$.

The sampling distribution is created by drawing samples of size $n = 2$ from the population. In other words, we toss two dice. Consider all the possible different samples of size $n = 2$ that could be drawn from the parent population. Figure 9.1 depicts this process in which we compute the mean for each sample. Because the value of the sample mean varies randomly from sample to sample, we can regard \bar{X} as a new random variable created by sampling. Table 9.2 lists all the possible samples and their corresponding values of \bar{X} .

FIGURE 9.1 Drawing samples of size $n = 2$ from a population**TABLE 9.2** All samples of size $n = 2$ and their means

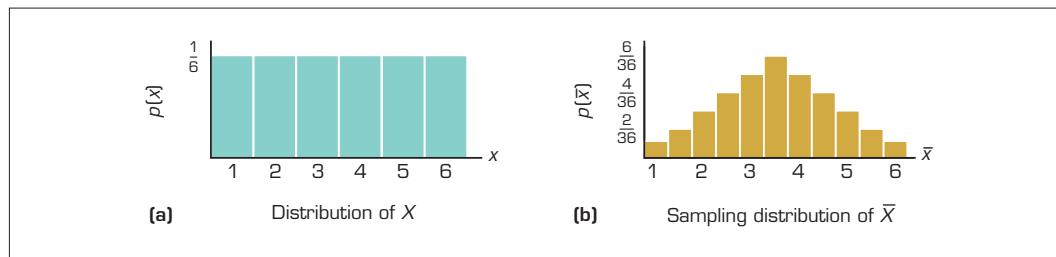
Sample	\bar{x}	Sample	\bar{x}	Sample	\bar{x}
1, 1	1.0	3, 1	2.0	5, 1	3.0
1, 2	1.5	3, 2	2.5	5, 2	3.5
1, 3	2.0	3, 3	3.0	5, 3	4.0
1, 4	2.5	3, 4	3.5	5, 4	4.5
1, 5	3.0	3, 5	4.0	5, 5	5.0
1, 6	3.5	3, 6	4.5	5, 6	5.5
2, 1	1.5	4, 1	2.5	6, 1	3.5
2, 2	2.0	4, 2	3.0	6, 2	4.0
2, 3	2.5	4, 3	3.5	6, 3	4.5
2, 4	3.0	4, 4	4.0	6, 4	5.0
2, 5	3.5	4, 5	4.5	6, 5	5.5
2, 6	4.0	4, 6	5.0	6, 6	6.0

There are 36 different possible samples of $n = 2$; as each sample is equally likely, the probability of any one sample being selected is 1/36. However, \bar{X} can assume only 11 different possible values: 1.0, 1.5, 2.0, ..., 6.0, with certain values of \bar{X} occurring more frequently than others. The value $\bar{X} = 1.0$ occurs only once, so its probability is 1/36. The value $\bar{X} = 1.5$ can occur in two ways; hence, $p(\bar{X} = 1.5) = 2/36$. The probabilities of the other values of \bar{X} are determined in similar fashion, and the sampling distribution of \bar{X} that results is shown in **Table 9.3**.

TABLE 9.3 Sampling distribution of \bar{X}

\bar{x}	$p(\bar{x})$	\bar{x}	$p(\bar{x})$	\bar{x}	$p(\bar{x})$
1.0	$\frac{1}{36}$	3.0	$\frac{5}{36}$	5.0	$\frac{3}{36}$
1.5	$\frac{2}{36}$	3.5	$\frac{6}{36}$	5.5	$\frac{2}{36}$
2.0	$\frac{3}{36}$	4.0	$\frac{5}{36}$	6.0	$\frac{1}{36}$
2.5	$\frac{4}{36}$	4.5	$\frac{4}{36}$		

The most interesting aspect of the sampling distribution of \bar{X} is how different it is from the distribution of X , as can be seen in **Figure 9.2**.

FIGURE 9.2 Distributions of X and \bar{X} 

In this experiment we know the population parameters $\mu = 3.5$ and $\sigma^2 = 2.92$. We can also compute the mean, variance, and standard deviation of the sampling distribution. Once again using the definitions of expected value and variance, we determine the following parameters of the sampling distribution of the sample mean \bar{X} .

Mean of the sampling distribution of \bar{X} :

$$\begin{aligned}\mu_{\bar{X}} &= E(\bar{X}) = \sum \bar{x} p(\bar{x}) \\ &= 1.0 \left(\frac{1}{36} \right) + 1.5 \left(\frac{2}{36} \right) + \dots + 6.0 \left(\frac{1}{36} \right) \\ &= 3.5\end{aligned}$$

Notice that the mean of the sampling distribution of \bar{x} is equal to the mean of the population X of the toss of a die computed previously.

Variance of the sampling distribution of \bar{X} :

$$\begin{aligned}\sigma_{\bar{X}}^2 &= V(\bar{X}) = E(\bar{X}^2) - \mu_{\bar{X}}^2 = \sum \bar{x}^2 p(\bar{x}) - \mu_{\bar{X}}^2 \\ &= \left[(1.0)^2 \left(\frac{1}{36} \right) + (1.5)^2 \left(\frac{2}{36} \right) + \dots + (6.0)^2 \left(\frac{1}{36} \right) \right] - (3.5)^2 \\ &= 1.46\end{aligned}$$

It is no coincidence that the variance of the sampling distribution of \bar{X} is exactly half of the variance of the population X of the toss of a die (computed previously as $\sigma^2 = 2.92$).

Standard deviation of the sampling distribution of \bar{X} :

$$\sigma_{\bar{X}} = \sqrt{1.46} = 1.21$$

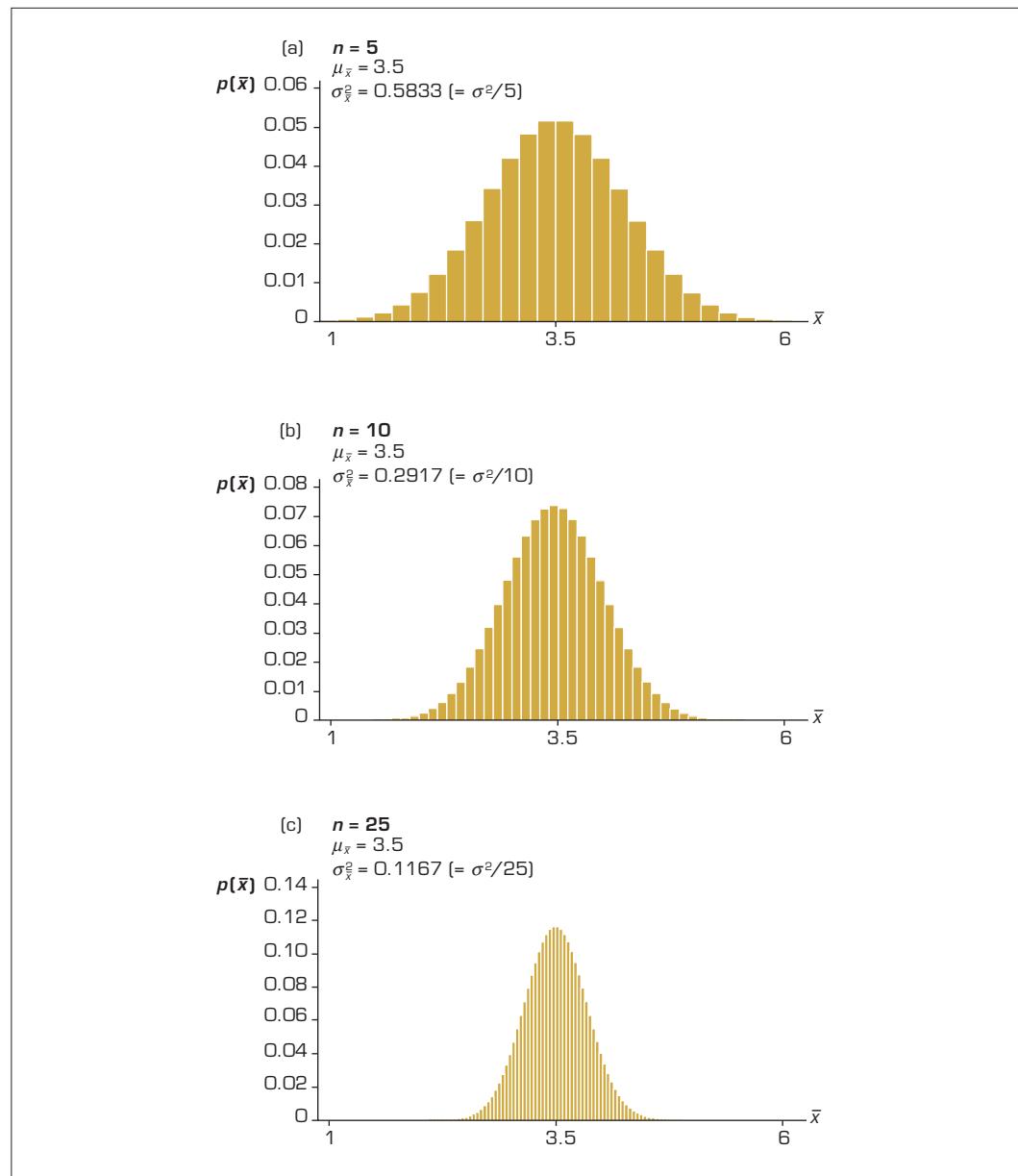
It is important to recognise that the distribution of \bar{X} is different from the distribution of X . **Figure 9.2** shows that the shapes of the two distributions differ. From the above calculations, we see that the mean of the sampling distribution of \bar{X} is equal to the mean of the distribution of X ; that is, $\mu_{\bar{X}} = \mu$. However, the variance of \bar{X} is not equal to the variance of X ; we calculated $\sigma^2 = 2.92$, but $\sigma_{\bar{X}}^2 = 1.46$. It is no coincidence that the variance of \bar{X} is exactly half the variance of X (i.e. $\sigma_{\bar{X}}^2 = \sigma^2 / 2$), as we will see shortly.

Don't get lost in the terminology and notation! Remember that μ and σ^2 are the parameters of the population of X . In order to create the sampling distribution of \bar{X} , we repeatedly drew samples of $n = 2$ from the population and calculated \bar{X} for each sample. Thus, we treat \bar{X} as a brand-new random variable, with its own distribution, mean and variance. The mean is denoted $\mu_{\bar{X}}$ and the variance is denoted $\sigma_{\bar{X}}^2$.

If we now repeat the sampling process with the same population but with other values of n , we produce somewhat different sampling distributions of \bar{X} . **Figure 9.3** shows the sampling distributions of \bar{X} when $n = 5, 10$ and 25 . As n grows larger, the number of possible values of \bar{X} also grows larger; consequently, the histograms depicted in **Figure 9.3** have been smoothed (to avoid drawing a large number of rectangles). Observe that in each case $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}}^2 = \sigma^2/n$.

Notice that in each case the variance of the sampling distribution is less than that of the parent population; that is, $\sigma_{\bar{X}}^2 < \sigma^2$. Given that $\sigma_{\bar{X}}^2 < \sigma^2$, a randomly selected value of \bar{X} (the mean of the number of spots observed in, say, five throws of the die) is likely to be closer to the mean value of 3.5 than is a randomly selected value of X (the number of spots observed in one throw). Indeed, this is what you would expect, because in five throws of the die you are likely to get some 5s and 6s and some 1s and 2s, which will tend to offset one another in the averaging process and produce a sample mean reasonably close to 3.5. As the number of throws of the die increases, the likelihood that the sample mean will be close to 3.5 also increases. Thus, we observe in **Figure 9.3** that the sampling distribution of \bar{X} becomes narrower (or more concentrated about the mean) as n increases.

FIGURE 9.3 Sampling distributions of \bar{X} when $n = 5, 10$ and 25



In summary, **Tables 9.2** and **9.3** and **Figures 9.1**, **9.2** and **9.3** illustrate how the sampling distribution is created, and make four important points about the sampling distribution of the sample mean. Using some basic rules of mathematics, we can prove the following relationships:

- 1 The mean of the sampling distribution of \bar{X} is equal to the mean of the population from which we have sampled. That is,

$$\mu_{\bar{X}} = \mu$$

In our illustration, both means equal 3.5.

- 2 The variance of the sampling distribution of \bar{X} is equal to the variance of the population divided by the sample size. That is,

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

- 3 The standard deviation of the sampling distribution of \bar{X} is called the **standard error of the sample mean** (labelled $\sigma_{\bar{X}}$). That is,

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

The standard error of the sample mean in our example when $n = 2$ is

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1.71}{\sqrt{2}} = 1.21$$

Beneath **Table 9.3**, we calculated the standard deviation of the 36 values of \bar{X} in **Table 9.2** to be 1.21, which is the same as the theoretical standard deviation of \bar{X} of 1.21.

- 4 As the sample size n gets larger, the sampling distribution of \bar{X} becomes increasingly bell shaped, even if the underlying population is not normally distributed or bell shaped. The mathematical phenomenon that produces this discovery regarding the shape of the sampling distribution of \bar{X} is called the **central limit theorem**.

standard error of the sample mean

The standard deviation of the sampling distribution of the sample mean, σ / \sqrt{n} .

central limit theorem

The sampling distribution of the sample mean \bar{X} will be approximately normal when $n > 30$.

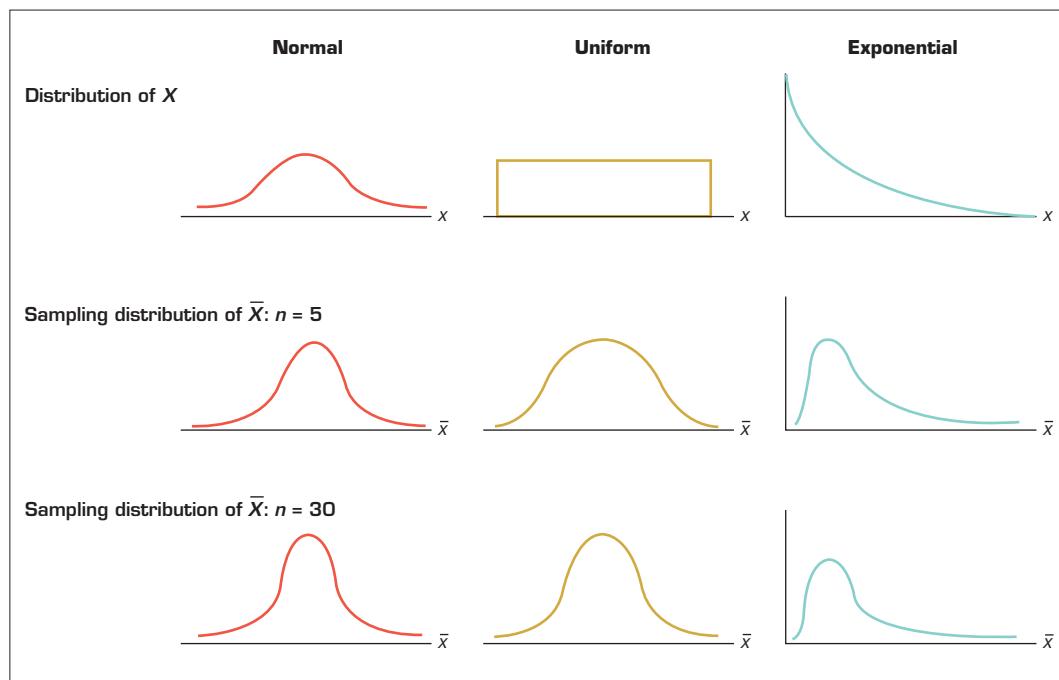
9.4b Sampling distribution of \bar{X}

Central limit theorem

If a random sample is drawn from any population, the sampling distribution of the sample mean (\bar{X}) is approximately normal for a sufficiently large sample size. The larger the sample size, the more closely the sampling distribution of \bar{X} will resemble a normal distribution.

The accuracy of the approximation alluded to in the central limit theorem depends on the probability distribution of the parent population and on the sample size. If the population is normal, then \bar{X} is always normally distributed for all values of n . If the population is non-normal, then \bar{X} is always approximately normal only for larger values of n . In many practical situations, a sample size of $n \geq 30$ may be sufficiently large to allow us to use the normal distribution as an approximation for the sampling distribution of \bar{X} . However, if a population is extremely non-normal (e.g. bimodal and highly skewed distributions), the sampling distribution will also be non-normal – even for moderately large values of n .

Figure 9.4 depicts the sampling distribution of \bar{X} for a variety of populations and sample sizes. Notice that when X is normally distributed, \bar{X} is also normally distributed for both $n = 5$ and $n = 30$ (that is, irrespective of the value of n). When X is uniformly distributed, the shape of the sampling distribution of \bar{X} is much closer to the normal shape when $n = 30$ than when $n = 5$. Obviously, the statement that \bar{X} is approximately normally distributed when X is uniformly distributed and $n = 5$ is quite weak. When $n = 30$, the normal approximation of the sampling distribution of \bar{X} is reasonable. Finally, when X is exponentially distributed, the sampling distribution of \bar{X} for both $n = 5$ and $n = 30$ is clearly non-normal. In order for the normal approximation of the sampling distribution to be valid, n would have to be larger. In general, when X follows a symmetric distribution, \bar{X} is more approximately normal for smaller values of n than it is when X is asymmetric.

FIGURE 9.4 Sampling distributions of \bar{X} from different populations

We can now summarise what we know about the sampling distribution of the sample mean.

Sampling distribution of the sample mean \bar{X}

- 1 $\mu_{\bar{X}} = \mu$
- 2 $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$ or $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ (The standard deviation of \bar{X} is called the standard error of the sample mean.)
- 3 If X is normal, \bar{X} is normal. If X is non-normal, then using the central limit theorem, \bar{X} is approximately normally distributed for sufficiently large sample sizes (at least 30, $n \geq 30$).

9.4c Creating the sampling distribution empirically

In the analysis above, we created the sampling distribution of the mean theoretically. We did so by listing all of the possible samples of $n = 2$ and their probabilities. (They were all equally likely with a probability of 1/36.) From this distribution, we produced the sampling distribution. We could also create the distribution empirically by actually tossing two fair dice repeatedly, calculating the sample mean for each sample, counting the number of times each value of \bar{X} occurs, and calculating the relative frequencies to estimate the theoretical probabilities. If we toss the two dice a large enough number of times, the relative frequencies and theoretical probabilities (calculated above) will be similar. Try it yourself. Toss two dice 500 times, count the number of times each sample mean occurs, and construct the sampling distribution. Obviously, this approach is far from ideal because of the excessive amount of time required to toss the dice enough times to make the relative frequencies good approximations for theoretical probabilities.

EXAMPLE 9.1

LO4

Annual returns of stocks in the construction industry

A stockbroker has observed that the annual return of stocks in the construction industry is actually a normally distributed random variable, with a mean of 12.5% and a standard deviation of 2.5%.

- The stockbroker selects a stock in the construction industry at random. Find the probability that the stock will have an annual return of less than 10.825%.
- The stockbroker then selects four stocks in the construction industry at random. Find the probability that all four of them will have an annual return of less than 10.825%.
- Find the probability that the mean annual return of the four randomly selected stocks will be less than 10.825%.
- Find the probability that the mean annual return of the four randomly selected stocks will be greater than 10.825%.
- Find the probability that the mean annual return of the four randomly selected stocks will be between 10.0% and 16.5%.

Solution

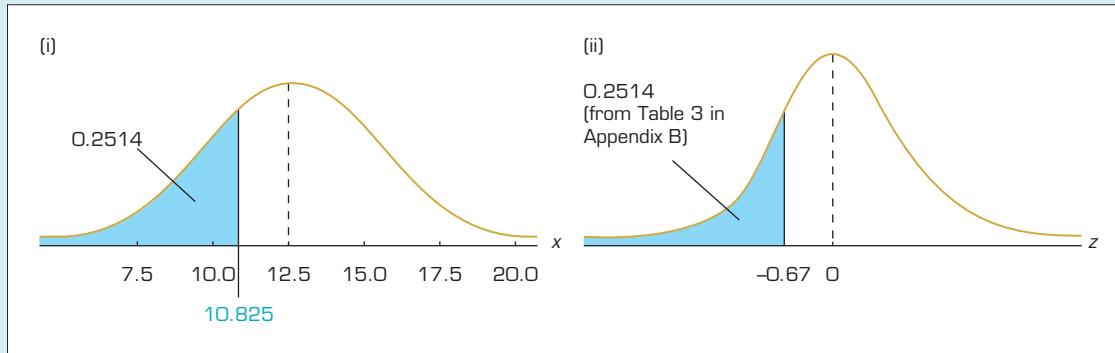
Let X be the annual return of a stock in the construction industry. The random variable X is normally distributed with mean $\mu = 12.5\%$ and standard deviation $\sigma = 2.5\%$.

- We want to find the normal probability $P(X < 10.825)$. To obtain this, we convert the normal variable X with $\mu = 12.5$ and $\sigma = 2.5$ to standard normal variable Z and use the standard normal table in Table 3 of Appendix B:

$$\begin{aligned} P(X < 10.825) &= P\left(\frac{X - \mu}{\sigma} < \frac{10.825 - 12.5}{2.5}\right) \\ &= P(Z < -0.67) \\ &= 0.2514 \end{aligned}$$

Figure 9.5a illustrates this distribution.

FIGURE 9.5A Distributions of X and its standardised value in Example 9.1(a)



- Now we want to find the probability that the annual return of each of the four randomly selected stocks in the construction industry is less than 10.825%. Using the multiplicative rule for independent events,

$$\begin{aligned} P(X < 10.825 \text{ for each of the four stocks}) &= [P(X < 10.825)]^4 \\ &= (0.2514)^4 \\ &= 0.0040 \end{aligned}$$

where we have used the probability $P(X < 10.825) = 0.2514$ obtained in part (a).

- c Now we want to find the probability that the mean annual return of the four randomly selected stocks in the construction industry is less than 10.825%. That is, we want:

$$P(\bar{X} < 10.825)$$

From our previous analysis, we know the following:

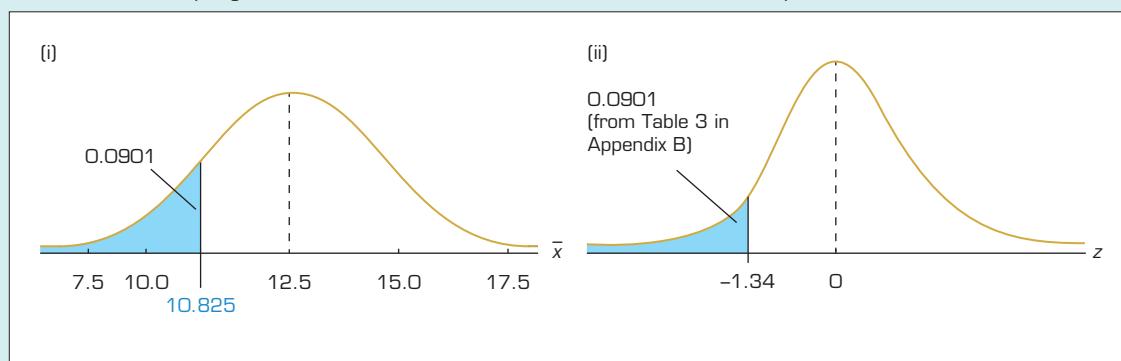
- i \bar{X} is normally distributed, since the population is normal.
- ii $\mu_{\bar{X}} = \mu = 12.5$
- iii $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2.5}{\sqrt{4}} = 1.25$

Hence, to calculate $P(\bar{X} < 10.825)$, we convert the normal variable \bar{X} with $\mu_{\bar{X}} = 12.5$ and $\sigma_{\bar{X}} = 1.25$ to a standard normal variable Z and use the standard normal table in Table 3 of Appendix B:

$$\begin{aligned} P(\bar{X} < 10.825) &= P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} < \frac{10.825 - 12.5}{1.25}\right) \\ &= P(Z < -1.34) \\ &= 0.0901 \end{aligned}$$

Figure 9.5b illustrates this distribution.

FIGURE 9.5B Sampling distribution of \bar{X} and its standardised value in Example 9.1(c)



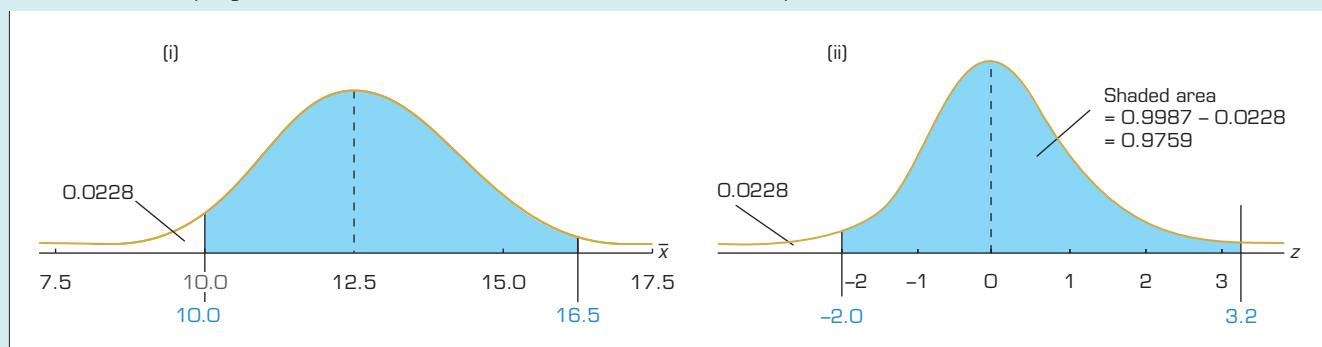
- d Now we want to find the probability that the mean annual return of the four stocks is greater than 10.825%. That is, we want $P(\bar{X} > 10.825)$. Using part (c) and the probability rule for complementary events, we have

$$\begin{aligned} P(\bar{X} > 10.825) &= 1 - P(\bar{X} < 10.825) \\ &= 1 - 0.0901 \\ &= 0.9099 \end{aligned}$$

- e Next we want to find the probability that the mean annual return of the four randomly selected stocks is between 10.0% and 16.5%. That is, $P(10.0 < \bar{X} < 16.5)$. Hence, using the Z transformation and Table 3 in Appendix B,

$$\begin{aligned} P(10.0 < \bar{X} < 16.5) &= P\left(\frac{10.0 - 12.5}{1.25} < \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} < \frac{16.5 - 12.5}{1.25}\right) \\ &= P(-2.0 < Z < 3.2) \\ &= P(Z < 3.2) - P(Z < -2.0) \\ &= 0.9987 - 0.0228 \\ &= 0.9759 \end{aligned}$$

Figure 9.5c illustrates this distribution.

**FIGURE 9.5C** Sampling distribution of \bar{X} and its standardised value in Example 9.1(e)

In Example 9.1 we began with the assumption that both μ and σ were known. Then, using the sampling distribution, we made a probability statement about \bar{X} . This use of the sampling distribution is of no great interest to us, since the values of the parameters μ and σ are generally unknown. We can, however, use the sampling distribution to infer something about the unknown value of μ , on the basis of a sample mean.

Now we are in a position to provide an answer to the opening example in this chapter.

SPOTLIGHT ON STATISTICS

Salaries of a business school's MBA graduates: Solution

We want to find the probability that the sample mean of the weekly salary is less than \$1900. Thus, we seek

$$P(\bar{X} < 1900)$$

The distribution of X , the weekly income, is likely to be positively skewed, but not sufficiently so as to make the distribution of \bar{X} non-normal. As a result, we may assume that \bar{X} is normal with mean $\mu_{\bar{X}} = \mu = 2000$ and standard deviation

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{200}{\sqrt{25}} = 40$$

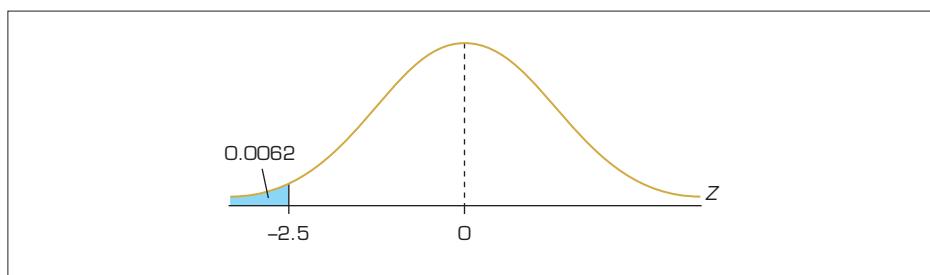
Thus:

$$\begin{aligned} P(\bar{X} < 1900) &= P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} < \frac{1900 - 2000}{40}\right) \\ &= P(Z < -2.5) \\ &= 0.0062 \end{aligned}$$

Figure 9.6 illustrates this distribution.



Source: © iStock.com/Alexandr Dubovitskiy

FIGURE 9.6 $P(\bar{X} < 1900) = P(Z < -2.5)$ 



The probability of observing a sample mean as low as \$1900 when the population mean is \$2000 is extremely small. Because this event is quite unlikely, we would have to conclude that, on the basis of the data collected by the student, the director's claim about the average salary of MBA graduates is not justified.

As you have just seen, the sampling distribution allows us to draw inferences about population parameters. In Chapters 10–16 we will use various sampling distributions to test and estimate several different parameters. In each application, the sampling distribution is a critical component of the technique that is used.

9.4d Sampling from a finite population

The definition of the variance of \bar{X} , $\sigma_{\bar{X}}^2$, is based on the assumption that the population from which we have sampled is very large (in fact, infinite). In most practical situations, this assumption is quite reasonable, since the purpose of sampling is to avoid the cost of investigating large populations. Nonetheless, situations do arise that are marked by a relatively small population. It turns out that \bar{X} is still approximately normally distributed with $\mu_{\bar{X}} = \mu$. The variance of \bar{X} , however, is now defined as

$$\sigma_{\bar{X}}^2 = \left(\frac{\sigma^2}{n} \right) \left(\frac{N-n}{N-1} \right)$$

where N = population size and $(N-n)/(N-1)$ is called the finite population correction factor. With a little algebra, you can see that the finite population correction factor is approximately equal to $1 - (n/N)$. Thus, when the sample size is small relative to the population size, the correction factor is close to 1, and $\sigma_{\bar{X}}^2$ is approximately equal to σ^2/n . For example, if $n/N = 0.05$, then the correction factor is close to 0.95, and $\sigma_{\bar{X}}^2 = 0.95 \sigma^2/n$. As a result, some practitioners use the correction factor only when the population size is small and n/N is relatively large (at least 10%). Since including the correction factor is quite easy, however, we recommend its use whenever the population is finite.

EXERCISES

Learning the techniques

- 9.2** A non-normally distributed population has a mean of 40 and a standard deviation of 12. What does the central limit theorem say about the sampling distribution of the mean if samples of size 100 are drawn from this population?
- 9.3** Refer to Exercise 9.2. Suppose that the population is normally distributed. Does this change your answer? Explain.
- 9.4** A sample of $n = 100$ observations is drawn from a normal population, with $\mu = 1000$ and $\sigma = 200$. Find:
 - a** $P(\bar{X} > 1050)$
 - b** $P(\bar{X} < 960)$
 - c** $P(\bar{X} > 1100)$
- 9.5** Repeat Exercise 9.4 with sample size $n = 16$.

- 9.6** A sample of 50 observations is taken from a normal population, with $\mu = 100$ and $\sigma = 10$. If the population is finite with $N = 250$. Find:
 - a** $P(\bar{X} > 103)$
 - b** $P(98 < \bar{X} < 101)$

- 9.7** Repeat Exercise 9.6 (a) and (b) with:
 - i** $N = 500$
 - ii** $N = 1000$

- 9.8** Table 7 in Appendix B is the random-numbers table. The single-digit numbers that appear there are drawn from a discrete uniform distribution. That is,

$$p(x) = \frac{1}{10}, \text{ where } x = 0, 1, \dots, 9$$

- a** Using your formula for expected value, find the mean of the probability distribution.
- b** Find the variance of this probability distribution.

- c** If a sample of $n = 100$ observations is taken from the discrete uniform population, what are the mean and the variance of the sampling distributions of the sample mean?
- d** Find the following probabilities:
- $P(4.4 < \bar{X} < 4.55)$
 - $P(\bar{X} > 5.0)$
 - $P(\bar{X} < 4.2)$
- 9.9** Consider the following experiment. A container holds six tokens, each marked with a different number between and including 1 and 6. The experiment consists of drawing tokens at random without replacement.
- List all possible samples of size 2 from this population.
 - Find the sampling distribution of \bar{X} .
 - Find the mean and the variance of the sampling distribution of \bar{X} . Check that $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}}^2 = \left(\frac{\sigma^2}{n}\right)\left(\frac{N-n}{N-1}\right)$.
- 9.10** Repeat Exercise 9.9 with the container holding tokens numbered 1 to 5.
- 9.11** Repeat Exercise 9.9 with the container holding tokens numbered 1 to 4.
- 9.12** Given a large population that has a mean of 1000 and a standard deviation of 200, find the probability that a random sample of 400 has a mean that lies between 995 and 1020.
- Applying the techniques**
- 9.13 Self-correcting exercise.** An automatic machine in a manufacturing process is operating properly if the lengths of an important subcomponent are normally distributed, with mean $\mu = 117$ cm and standard deviation $\sigma = 2.1$ cm.
- Find the probability that one randomly selected unit has a length greater than 120 cm.
 - Find the probability that if three units are randomly selected, all three have lengths exceeding 120 cm.
 - Find the probability that if three units are randomly selected, their mean length exceeds 120 cm.
 - Explain the differences between parts (a), (b) and (c).
- 9.14** The manufacturer of cans of salmon that are supposed to have a net weight of 120 g tells you that the net weight is actually a random variable with a mean of 121 g and a standard deviation of 3.6 g. Suppose that you take a random sample of 36 cans.
- a** Find the probability that the sample mean will be less than 119.4 g.
- b** Suppose that your random sample of 36 cans produces a mean of 119 g. Comment on the statement made by the manufacturer.
- 9.15** A shipment of 500 cereal boxes is delivered to a supermarket in Fremantle. The manager is told that the weights of the cereal boxes are normally distributed, with a mean of 165 g and a standard deviation of 5 g.
- Find the probability that a random sample of 200 boxes has a mean weight of between 164.9 g and 165.1 g.
 - Repeat part (a), assuming that the shipment consists of:
 - 1000 boxes
 - 2000 boxes
 - 10000 boxes.
 - Comment on the effect that the population size has on the probability.
- 9.16** The sign on the lift in a building states 'Maximum capacity 1120 kg or 16 persons'. A statistics practitioner wonders what the probability is that 16 people would weigh more than 1120 kg. If the weights of the people who use the lift are normally distributed, with a mean of 68 kg and a standard deviation of 8 kg, what is the probability that the statistics practitioner seeks?
- 9.17** The amount of time that university lecturers devote to their jobs per week is normally distributed, with a mean of 52 hours and a standard deviation of 6 hours.
- What is the probability that a lecturer works for more than 60 hours per week?
 - Find the probability that the mean amount of work per week for three randomly selected lecturers is more than 60 hours.
 - Find the probability that if three lecturers are randomly selected, all three work for more than 60 hours per week.
- 9.18** The time it takes for a statistics lecturer to mark the mid-semester test is normally distributed, with a mean of 4.8 minutes and a standard deviation of 1.3 minutes. There are 60 students in the lecturer's class. What is the probability that he needs more than 5 hours to mark all the mid-semester tests?

- 9.19** Refer to Exercise 9.18. Does your answer change if you discover that the times needed to mark a mid-semester test are not normally distributed?
- 9.20** The supervisor of a chocolate factory has observed that the weight of each '32 g' chocolate bar is actually a normally distributed random variable, with a mean of 32.2 g and a standard deviation of 0.3g.
- Find the probability that, if a customer buys one chocolate bar, that bar will weigh less than 32g.
 - Find the probability that, if a customer buys a pack of four bars, the mean weight of the four bars will be less than 32g.
 - Find the probability that, if a customer buys a pack of four bars, the mean weight of the four bars will be greater than 32g.
- 9.21** The number of pizzas consumed per month by university students is normally distributed with a mean of 10 and a standard deviation of 3.
- What proportion of students consume more than 12 pizzas per month?
 - What is the probability that for a random sample of 25 students more than 275 pizzas are consumed?
- (Hint: What is the mean number of pizzas consumed by the sample of 25 students?)
- 9.22** The marks on a statistics mid-semester exam are normally distributed with a mean of 78 and a standard deviation of 6.
- What proportion of the class has a mid-semester mark of less than 75?
 - What is the probability that a class of 50 has an average mid-semester mark that is less than 75?

9.5 Sampling distribution of the sample proportion \hat{p}

In Chapter 2 we pointed out that when the data type is nominal (categorical), the only calculation permitted is one to determine the proportion of times each value occurs. If the problem objective is to describe a single population, the parameter of interest is the proportion p of times a certain outcome occurs. In keeping with the concepts and notation of the binomial experiment, we label the outcome of interest to us a *success*. Any other outcomes are labelled *failures*. In order to compute binomial probabilities, we assumed that p was known. However, in reality p is unknown, requiring the statistics practitioner to estimate its value from a sample. The estimator of a population proportion of successes is the sample proportion. That is, we count the number of successes in a sample and compute:

$$\hat{p} = \frac{X}{n}$$

This equation shows the result for the sample proportion \hat{p} (pronounced p-hat), where X is the number of successes in the sample and n is the sample size. When we take a sample of size n , we are actually conducting a binomial experiment, and as a result X is binomially distributed. Thus, the probability of any value of \hat{p} can be calculated from its value of X . For example, suppose that we have a binomial experiment with $n = 10$ and $p = 0.4$. To find the probability that the sample proportion \hat{p} is less than or equal to 0.5, we find the probability that X is less than or equal to 5 (because, if $\hat{p} = 0.5$, then $X = np = 10(0.5) = 5$). From Table 1 in Appendix B we find that with $n = 10$ and $p = 0.4$:

$$P(\hat{p} \leq 0.5) = P(X \leq 5) = 0.8338$$

We can similarly calculate the probability associated with other values of \hat{p} . As for the sample mean, the **sampling distribution of the sample proportion** is created by considering all possible sample proportions of the same sample size taken from the population.

sampling distribution of the sample proportion

A relative frequency distribution of various values of the sample proportion using a number of samples from the same population.

9.5a Mean and standard deviation of the sample proportion \hat{p}

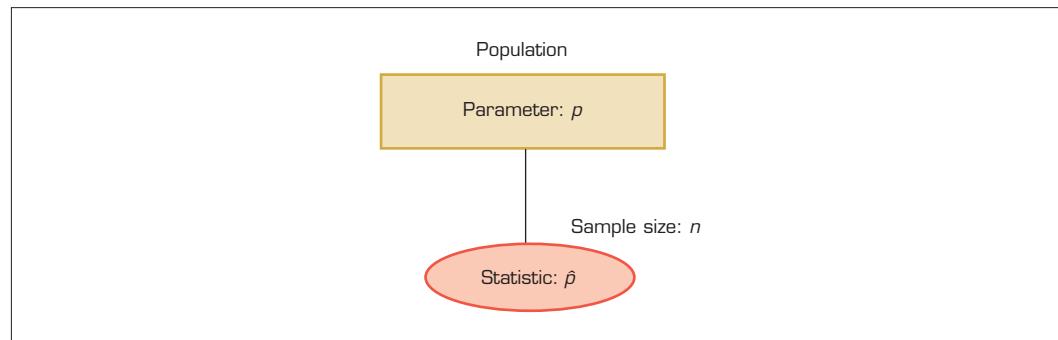
Figure 9.7 describes the sampling process. The proportion of successes in the population is p . A random sample of n observations contains X successes. Hence the proportion of successes in the sample is \hat{p} ($= X/n$). Recall (from Chapter 7) that X is a binomial random variable and the mean of X is

$$E(X) = np$$

and that the standard deviation of X is

$$SD(X) = \sigma_X = \sqrt{npq}, \text{ where } q = 1 - p$$

FIGURE 9.7 Sampling from a population with nominal data



Using the laws of expected value and variance (again, from Chapter 7), we can determine the mean, the variance and the standard deviation of \hat{p} . We will summarise what we have learnt.

$$E(\hat{p}) = \mu_{\hat{p}} = p$$

$$V(\hat{p}) = \sigma_{\hat{p}}^2 = \frac{pq}{n}$$

$$SD(\hat{p}) = \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

where $q = 1 - p$. The standard deviation of \hat{p} is called the **standard error of the sample proportion**. As was the case with the standard error of the sample mean (page 359), the standard error of a sample proportion is $\sqrt{pq/n}$, when sampling from infinitely large populations. When the population is finite, the standard error of the sample proportion must include the finite population correction factor, which can be omitted when the population is large relative to the sample size – a very common occurrence in practice.

standard error of the sample proportion

The standard deviation of the sampling distribution of the sample proportion, $\sqrt{pq/n}$.

9.5b Sampling distribution of \hat{p}

Discrete distributions such as the binomial do not lend themselves easily to the kinds of calculation needed for inference. And inference is the reason we need sampling distributions.

As X is a binomial random variable, we can use the binomial distribution to estimate p . However, the binomial random variable is discrete, making it awkward to use in statistical inferences about p as \hat{p} is continuous. But it can be shown (see Appendix to Chapter 8) that the binomial distribution can be approximated by the normal distribution, provided that n is sufficiently large. (The theoretical sample size requirements are that np and nq are both greater than or equal to 5. We refer to this requirement as *theoretical* because, in practice, much larger sample sizes are needed for the inference to be useful.) As a result, we have the following distribution.

Sampling distribution of the sample proportion \hat{p}

- 1 The sample proportion \hat{p} is approximately normally distributed, provided that n is large such that $np \geq 5$ and $nq \geq 5$, $q = 1 - p$.
- 2 The expected value (or mean): $E(\hat{p}) = p$.
- 3 The variance: $V(\hat{p}) = \sigma_{\hat{p}}^2 = \frac{pq}{n}$
- 4 The standard deviation:² $\sigma_{\hat{p}} = \sqrt{pq/n}$

The standard deviation of \hat{p} is called the **standard error of the sample proportion**. As \hat{p} is approximately normal, it follows that the standardised variable

$$z = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

is approximately distributed as a standard normal distribution, provided that $np \geq 5$ and $nq \geq 5$.

EXAMPLE 9.2

LO4

Will the Liberal–National Coalition candidate win again in the next election?

In the last federal election a Liberal–National Coalition candidate received 45.6% of the two-party preferred votes cast in his electorate. An adviser to the candidate is planning to organise a survey that would ask a random sample of 300 people from that electorate whether they would vote for the candidate again at the next election. If we assume that the popularity of the Coalition candidate has not changed, what is the probability that more than 50% of the sample the adviser is going to use would vote for the Coalition candidate in the next election?

Solution

The number of respondents who would vote for the Coalition candidate is a binomial random variable with $n = 300$ and $p = 0.456$. We want to determine the probability that the sample proportion is greater than 50%. That is, we want to find $P(\hat{p} > 0.50)$.

We now know that the sample proportion \hat{p} is approximately normally distributed with mean $p = 0.456$ and standard deviation

$$\sqrt{pq/n} = \sqrt{(0.456)(0.544)/300} = 0.0288$$

where $q = 1 - p = 1 - 0.456 = 0.544$. Thus, we calculate

$$\begin{aligned} P(\hat{p} > 0.50) &= P\left[\frac{\hat{p} - p}{\sqrt{pq/n}} > \frac{0.50 - 0.456}{0.0288}\right] \\ &= P(Z > 1.53) \\ &= 1 - P(Z < 1.53) \\ &= 1 - 0.9382 \\ &= 0.0618 \end{aligned}$$

If we assume that the level of support remains at 45.6%, the probability that more than 50% of the sample of 300 people would vote for the Coalition candidate again (on a two-party preferred basis) in the next election is about 6.18%.

² As was the case with the standard error of the mean (page 359), the standard error of a sample proportion is $\sqrt{pq/n}$ when sampling from infinitely large populations. When the population is finite, the standard error of the proportion must include the finite population correction factor, which can be omitted when the population is large relative to the sample size, a very common occurrence in practice.

In later chapters, we will develop the sampling distributions of the sample variance and the difference between the sample means and sample proportions, as well as other more complex statistics.

EXERCISES

Learning the techniques

Use the normal approximation without the correction factor to find the probabilities in the following exercises.

- 9.23 a** In a binomial experiment with $n = 300$ and $p = 0.5$, find the probability that \hat{p} is greater than 60%.

- b** Repeat part (a) with $p = 0.55$.
c Repeat part (a) with $p = 0.6$.

- 9.24 a** The probability of success on any trial of a binomial experiment is 25%. Find the probability that the proportion of successes in a sample of 500 is less than 22%.

- b** Repeat part (a) with $n = 800$.
c Repeat part (a) with $n = 1000$.

- 9.25** Determine the probability that in a sample of 100, the sample proportion is less than 0.75 given that $p = 0.8$.

- 9.26** A binomial experiment where $p = 0.4$ is conducted. Find the probability that in a sample of 60, the proportion of successes exceeds 0.35.

Applying the techniques

- 9.27 Self-correcting exercise.** The proportion of eligible voters who will vote for the incumbent in the next election is assumed to be 55%. What is the probability that in a random sample of 500 voters fewer than 49% say they will vote for the incumbent?

- 9.28** The assembly line that produces an electronic component of a missile system has historically resulted in a 2% defect rate. A random sample of 800 components is drawn. What is the probability that the defect rate is greater than 4%? Suppose that in the random sample the defect rate is 4%.

What does that suggest about the defect rate on the assembly line?

- 9.29** The manufacturer of aspirin claims that the proportion of headache sufferers who obtain relief from two aspirin is 53%.

- a** What is the probability that in a random sample of 400 headache sufferers fewer than 50% obtain relief? If 50% of the sample actually obtained relief, what does this suggest about the manufacturer's claim?
b Repeat part (a) using a sample of 1000 headache sufferers.

- 9.30** The manager of a restaurant has determined that the proportion of customers who drink tea is 14%. What is the probability that of the next 100 customers at least 10% will be tea drinkers?

- 9.31** A commercial for a manufacturer of household appliances claims that only 3% of all its products require a service call within the first year after purchase. A consumer protection association wants to check the claim by surveying 400 households that recently purchased one of the company's appliances. What is the probability that more than 5% will require a service call within the first year? What would you say about the commercial's honesty if, in a random sample of 400 households, 5% report at least one service call?

- 9.32** The Laurier Company's brand has a market share of 30%. Suppose that in a survey 1000 consumers of the product are asked which brand they prefer. What is the probability that more than 32% of the respondents will say they prefer the Laurier Company's brand?

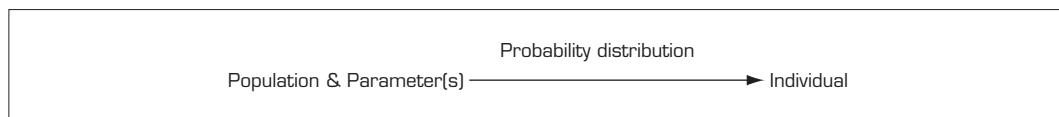
9.6 From here to inference

The primary function of the sampling distribution is statistical inference. To see how the sampling distribution contributes to the development of inferential methods, we need to briefly review how we got to this point.

In Chapters 7 and 8 we introduced probability distributions, which allowed us to make probability statements about values of the random variable. A prerequisite of this calculation is knowledge of the distribution and the relevant parameters.

Figure 9.8 symbolically represents the use of probability distributions. Simply put, knowledge of the population and its parameter(s) allows us to use the probability distribution to make probability statements about individual members of the population. The direction of the arrows indicates the direction of the flow of information.

FIGURE 9.8 Probability distribution



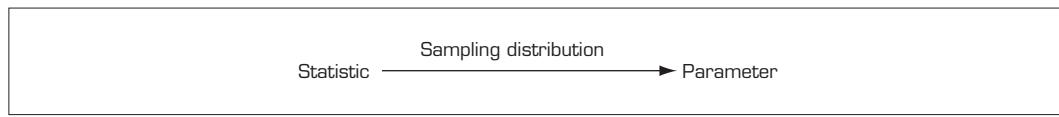
In this chapter we developed the sampling distribution, so that knowledge of the parameter(s) and some information about the distribution allow us to make probability statements about a sample statistic. In the chapter-opening example, knowing the population mean and the standard deviation and assuming that the population is not extremely non-normal enabled us to calculate a probability statement about a sample mean. **Figure 9.9** describes the application of sampling distributions.

FIGURE 9.9 Sampling distribution



Notice that in applying both probability distributions and sampling distributions, we must know the value of the relevant parameters – a highly unlikely circumstance. In the real world, parameters are almost always unknown because they represent descriptive measurements about extremely large populations. Statistical inference addresses this problem. It does so by reversing the direction of the flow of knowledge shown in **Figure 9.9**. In **Figure 9.10** we display the character of statistical inference. Starting from Chapter 10, we assume that most population parameters are unknown. The statistics practitioner will sample from the population and compute the required statistic. The sampling distribution of that statistic will enable us to draw inferences about the parameter.

FIGURE 9.10 Sampling distribution in inference



You may be surprised to learn, by and large, that is all we do in the remainder of this book. Why then, you might think, do we need another 15 chapters? Well, they are necessary because there are many more parameter and sampling distribution combinations that define the inferential procedures to be presented in an introductory statistics course. However, they all work in the same way. If you understand how one procedure is developed, you will likely understand all of them. Our task, and yours, in the next two chapters is to ensure that you understand the first inferential method.

Study Tools

CHAPTER SUMMARY

Because most populations are large, it is extremely costly and impractical to investigate every member of the population to determine the value of the parameters. As a practical alternative, we *sample* the population and use the sample statistics to draw inferences about the population parameters.

The mean and variance of the sampling distribution of the *sample mean* \bar{X} are μ and σ^2/n respectively. If the population is normally distributed, the sampling distribution of \bar{X} is also normally distributed. The *central limit theorem* states that the sampling distribution of \bar{X} is approximately normal (for large n), even if the population is non-normal.

The mean and variance of the sampling distribution of the *sample proportion* \hat{p} are p and $\sqrt{pq/n}$ respectively. If n is large such that $np \geq 5$ and $nq \geq 5$, the sampling distribution of the *sample proportion* \hat{p} is approximately normally distributed.

The sampling distributions of \bar{X} and \hat{p} and other sampling distributions (developed in later chapters) play a critical role in statistics by providing a probability link between the sample statistics and the population parameters.

The data files for Examples and Exercises are provided on the companion website (accessible through <https://login.cengagebrain.com/>).

SYMBOLS

Symbol	Pronounced	Represents
$\mu_{\bar{x}}$	mu-x-bar	Mean of the sampling distribution of the sample mean
$\sigma_{\bar{x}}^2$	Sigma-squared-x-bar	Variance of the sampling distribution of the sample mean
$\sigma_{\bar{x}}$	Sigma-x-bar	Standard deviation (standard error) of the sampling distribution of the sample mean
\hat{p}	p-hat	Sample proportion
$\sigma_{\hat{p}}^2$	Sigma-squared-p-hat	Variance of the sampling distribution of the sample proportion
$\sigma_{\hat{p}}$	Sigma-p-hat	Standard deviation (standard error) of the sampling distribution of the sample proportion

SUMMARY OF FORMULAS

Expected value of the sample mean	$E(\bar{X}) = \mu_{\bar{x}} = \mu$
Variance of the sample mean	$V(\bar{X}) = \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$
Standard error of the sample mean	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
Standardising the sample mean	$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$
Expected value of the sample proportion	$E(\hat{p}) = \mu_{\hat{p}} = p$
Variance of the sample proportion	$V(\hat{p}) = \sigma_{\hat{p}}^2 = \frac{pq}{n}, q = 1 - p$
Standard error of the sample proportion	$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$
Standardising the sample proportion	$Z = \frac{\hat{p} - p}{\sqrt{pq/n}}$

SUPPLEMENTARY EXERCISES

- 9.33** The dean of a business school claims that the average Master of Business Management graduate is offered an annual starting salary of \$109 500. The standard deviation of the offers is \$9900. What is the probability that for a random sample of 38 Master of Business Management graduates, the mean starting annual salary is less than \$105 000?
- 9.34** Refer to Exercise 9.33. Suppose that a random sample of 38 Master of Business Management graduates report that their mean starting salary is \$105 000. What does this tell you about the dean's claim?
- 9.35** A restaurant in a large commercial building provides coffee for the building's occupants. The restaurateur has determined that the mean number of cups of coffee consumed in one day by all the occupants is 2.0, with a standard deviation of 0.6. A new tenant of the building intends to have a total of 125 new employees. What is the probability that the new employees will consume more than 240 cups of coffee per day?
- 9.36** The number of pages photocopied each day by the admin staff in a busy office is normally distributed with a mean of 550 and a standard deviation of 150. Determine the probability that in one business week (i.e. 5 days) more than 3000 pages will be copied.
- 9.37** A university bookstore claims that 50% of its customers are satisfied with the service and the prices.
- If this claim is true, what is the probability that in a random sample of 600 customers less than 45% are satisfied?
 - Suppose that in a random sample of 600 customers, 270 express satisfaction with the bookstore. What does this tell you about the bookstore's claim?
- 9.38** A psychologist believes that 80% of male drivers when lost continue to drive, hoping to find the location they seek, rather than ask for directions. To examine this belief, he took a random sample of 350 male drivers and asked each what they did when lost. If the belief is true, determine the probability that less than 75% said they continue driving.
- 9.39** The Red Lobster restaurant chain regularly surveys its customers. On the basis of these surveys, management claims that 75% of customers rate the food as excellent. A consumer testing service wants to examine the claim by asking 460 customers to rate the food. What is the probability that less than 70% rate the food as excellent?
- 9.40** An accounting professor claims that no more than one-quarter of undergraduate business students will major in accounting.
- What is the probability that in a random sample of 1200 undergraduate business students, 336 or more will major in accounting?
 - A survey of a random sample of 1200 undergraduate business students indicates that there are 336 students who plan to major in accounting. What does this tell you about the professor's claim?
- 9.41** Statisticians determined that the mortgages of homeowners in a city is normally distributed with a mean of \$500 000 and a standard deviation of \$100 000. A random sample of 100 homeowners was drawn. What is the probability that the mean is greater than \$524 000?
- 9.42** Refer to Exercise 9.41. Does your answer change if you discover that mortgages are not normally distributed?
- 9.43** In a survey, Australians were asked how many sources or platforms they use to access news. If 21% of the population report that they use a single source to access news, find the probability that in a sample of 500 at least 22% say that they use a single source to get their news.

Estimation: Single population

Learning objectives

This chapter introduces an important topic in statistical inference: estimation of population parameters. The parameters discussed are the mean and the proportion of a single population.

At the completion of this chapter, you should be able to:

- L01** understand the fundamental concepts of estimation
- L02** understand the difference between a point estimator and an interval estimator
- L03** understand the *t* distribution
- L04** determine when to use the *z* distribution and the *t* distribution in estimation
- L05** construct an interval estimate of a population mean when the population variance is known
- L06** construct an interval estimate of a population mean when the population variance is unknown
- L07** construct an interval estimate of a population proportion
- L08** interpret interval estimates
- L09** determine the minimum sample size required for estimating a population mean and a population proportion.

CHAPTER OUTLINE

Introduction

10.1 Concepts of estimation

10.2 Estimating the population mean μ when the population variance σ^2 is known

10.3 Estimating the population mean μ when the population variance σ^2 is unknown

10.4 Estimating the population proportion p

10.5 Determining the required sample size

10.6 Applications in marketing: Market segmentation

SPOTLIGHT ON STATISTICS

Segmentation of the breakfast cereal market

Product segmentation is commonly used in marketing products such as breakfast cereals. A particular food manufacturer uses health and diet consciousness as the segmentation variable, and the following four segments to market its breakfast cereal:

- 1** Concerned about eating healthy foods
- 2** Concerned primarily about weight
- 3** Concerned about health because of illness
- 4** Unconcerned



Source: Shutterstock.com/Pogorelova Olga

A questionnaire is used to categorise people as belonging to one of these groups. In a recent survey, a random sample of 1250 Australian adults (20 years and over) were asked to complete the questionnaire. The results were recorded and stored in **CH10:XM10-00**. The ABS statistics reveal that there were 18798064 Australians in the 20 and over age group in 2018 (ABS Australian Demographic Statistics, cat. no. 3101.0, March 2019). Estimate with 95% confidence the number of Australian adults who are concerned about eating healthy foods. See pages 419–20 for the solution.

Introduction

Having discussed descriptive statistics (Chapters 3, 4 and 5), probability and probability distributions (Chapters 6, 7 and 8) and sampling distributions (Chapter 9), we are ready to tackle statistical inference. As we explained in Chapter 1, statistical inference is the process by which we acquire information and draw conclusions about populations from samples.

In Chapter 9, in which we presented a central concept of statistical inference, the sampling distribution, we pointed out that we can make inferences about populations in two ways: by estimating the unknown population parameter or by testing its value. In this chapter we will deal with the problem of estimating parameters that describe single populations and demonstrate them with simple examples.

In Chapter 12, we describe the fundamentals of hypothesis testing. Because most of what we do in the remainder of this book applies the concepts of estimation and hypothesis testing, an understanding of Chapters 10–13 is vital to your development as a statistics practitioner. The actual parameter of interest will be determined by the type of data. If the data type is numerical (quantitative), we will be interested in estimating the population mean. If the data type is nominal, however, the parameter is the population proportion p .

Here are some examples illustrating the estimation of parameters of a single population:

- 1** An inspector from the Department of Consumer Affairs wanted to know whether the actual weight of cans of tuna was at least as large as the weight shown on the label. Since she cannot weigh every single can of tuna, she draws a random sample of cans and uses the sample data to estimate the mean of all cans of tuna.
- 2** One week before election day, an incumbent member of parliament (MP) wants to know if he will win the forthcoming election. A survey of 500 registered voters is conducted in which voters are asked whom they will vote for. The results allow the MP to estimate the proportion of all voters who support him.

Chapters 10–13 use our usual approach in solving problems manually and using Excel on the computer. Some of the computations are not available in the Data Analysis tool available in Excel as a standard add-in. In these cases, we have created workbooks in Excel which can be used for solving problems with ease for small as well as large sample data. The workbooks are available in the Workbooks tab on the companion website. Wherever required, in this chapter, we will utilise the **Estimators** workbook, which is provided on the companion website (accessible through <https://login.cengagebrain.com/>). There are nine worksheets in the Estimators workbook. Each is associated with confidence interval estimators that will be introduced later in this chapter and in Chapter 11. This workbook also contains two more worksheets that can be used to obtain the required sample sizes for estimating the mean and proportion. The spreadsheets also allow us to perform the ‘what if’ analyses discussed in these chapters. Additionally, XLSTAT* commands are provided for those examples that cannot be completed using the *Data Analysis* tool in Excel.

* XLSTAT is a commercially created add-in that can be loaded onto your computer to enable you to use Excel for almost all statistical procedures introduced in this book.

10.1 Concepts of estimation

As its name suggests, the objective of estimation is to determine the approximate value of a population parameter on the basis of a sample statistic. For example, the sample mean is employed to estimate the population mean. We refer to the sample mean as the *estimator* of the population mean. Once the sample mean has been calculated, its value is called the *estimate*. In this chapter, we will introduce the statistical process in which we estimate a population mean and population proportion using sample data. In the rest of the book, we use the concepts and techniques introduced here for other parameters.

10.1a Point and interval estimators

We can use sample data to estimate a population parameter in two ways. First, we can compute the value of the estimator and consider that value as the estimate of the parameter. Such an estimator is called a **point estimator**.

Point estimator

A point estimator draws inferences about a population by estimating the value of an unknown parameter using a single value or point.

point estimator

Estimates the value of an unknown parameter using a single value.

There are two drawbacks to using point estimators. First, we often need to know how close the estimator is to the parameter. Second, in drawing inferences about a population, it is intuitively reasonable to expect that a large sample will produce more accurate results than a smaller sample because it contains more information. But point estimators don't have the capacity to reflect the effects of larger sample sizes. As a consequence, we use the second method of estimating a population parameter, the **interval estimator**.

Interval estimator

An interval estimator draws inferences about a population by estimating the value of an unknown parameter using an interval.

interval estimator

Estimates the value of an unknown parameter which reflects the variability in the sample, sample size and the level of confidence using an interval.

Thus, we may estimate that the mean annual income for first-year accountants lies in the range \$68 000 to \$78 000. This interval estimate provides us with additional information about how low and how high the true population mean income is likely to be. As you will see, this additional information is considerably more useful than the information supplied by a point estimate alone. The interval estimator is also affected by the sample size.

To illustrate the difference between point and interval estimators, suppose that a statistics lecturer wants to estimate the mean summer income of his second-year business students. He selects 25 students at random and calculates the sample mean weekly income to be \$550. The point estimate is the sample mean. That is, he estimates the mean weekly summer income of *all* second-year business students to be \$550. Using the technique described in Section 10.3, he could instead use an interval estimate and estimate that the average second-year business student earns, for example, between \$530 and \$570 each week during the summer.

In the preceding discussion, we used the terms *estimator* and *estimate*. We distinguish between them by noting that an estimate is the calculation of a specific value of the estimator. For example, we say that the sample mean is an estimator of a population mean. However, once we calculate the value of the sample mean, that value is an estimate of the population mean.

10.1b Quality of an estimator

Numerous applications of estimation occur in the real world. For example, television network executives want to know the proportion of television viewers who are tuned into their network; a production manager wishes to know the average daily production in the plant; a union negotiator would like to know the average annual income of Australian blue-collar workers. In each of these cases, in order to accomplish the objective exactly, the interested party would have to examine each member of the population and then calculate the parameter of interest. For instance, the union negotiator would have to ask every Australian blue-collar worker what his or her annual income is, and then calculate the average of these values – a task that is both impractical and prohibitively expensive. An alternative would be to take a random sample from this population, calculate the sample mean, and use that as an estimate of the population mean.

The use of the sample mean to estimate the population mean seems logical. The selection of the sample statistic to be used as an estimator, however, depends on the characteristics of that statistic. Naturally, we want to use the statistic with the most desirable qualities for our purposes. One such desirable quality of an estimator is for it to be an unbiased estimator. An **unbiased estimator** of a population parameter is one whose expected value is equal to that parameter. This means that, if you were to take an infinite number of samples, calculate the value of the estimator in each sample, and then average these values, the average value would equal the parameter. Essentially, this amounts to saying that, on average, the sample statistic is equal to the population parameter.

unbiased estimator

Has an expected value that equals the parameter being estimated.

Unbiased estimator

An **unbiased estimator** of a population parameter is an estimator whose expected value is equal to that parameter.

The *bias* of an estimator is the absolute value of the difference between the expected value of the estimator and the population parameter. For example, the bias of the sample mean \bar{X} is $\text{Bias}(\bar{X}) = |E(\bar{X}) - \mu|$.

In Section 9.4, we established that $\mu_{\bar{X}} = E(\bar{X}) = \mu$. That is, $\text{Bias}(\bar{X}) = |E(\bar{X}) - \mu| = 0$. Hence the sample mean \bar{X} is an unbiased estimator of the population mean μ . **Figure 10.1** depicts the sampling distribution of an unbiased estimator \bar{X} of the population mean μ , and **Figure 10.2** presents the sampling distribution of a biased estimator, $\hat{\mu}$, of μ .

We also know from Section 9.5 that $\mu_{\hat{p}} = E(\hat{p}) = p$. This means that the sample proportion \hat{p} is an unbiased estimator of the population proportion p .

FIGURE 10.1 Sampling distribution of \bar{X}

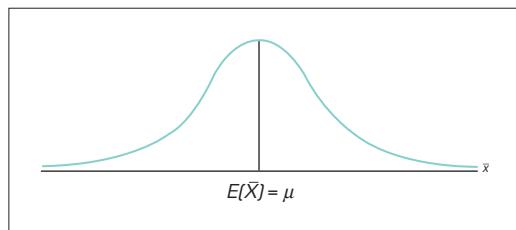
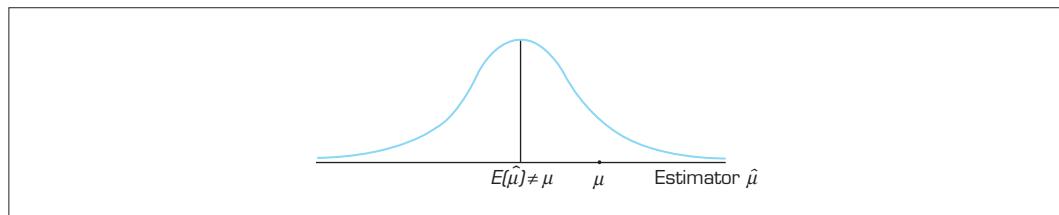


FIGURE 10.2 Sampling distribution of a biased estimator $\hat{\mu}$ of μ



As a second illustration of quality, recall that in Chapter 5 we defined the sample variance s^2 to be $\sum(x_i - \bar{X})^2 / (n-1)$. At the time, it seemed odd that we divided by $(n-1)$ rather than by n . The reason for choosing $(n-1)$ was to make $\sum(s^2) = \sigma^2$, so that the definition of s^2 produces an unbiased estimator of σ^2 . Defining s^2 as $\sum(x_i - \bar{X})^2 / n$ would have resulted in a biased estimator of σ^2 ; one that produced an average s^2 that was smaller than the true value of σ^2 .

Knowing that an estimator is unbiased only assures us that its expected value equals the parameter; it does not tell us how close the estimator is to the parameter. Another desirable quality is for the estimator to be as close to its parameter as possible; and certainly, as the sample size grows larger, the sample statistic should come closer to the population parameter. This quality is called *consistency*. A **consistent estimator** of a proportion parameter is one that approaches the value of the parameter as the sample size increases.

Consistent estimator

An unbiased estimator is said to be **consistent** if the difference between the estimator and the parameter grows smaller as the sample size grows larger.

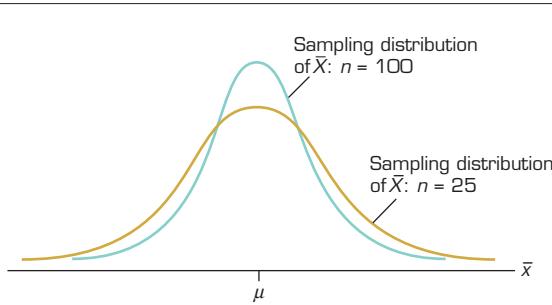
consistent estimator

Approaches the value of the parameter it is estimating as the sample size increases.

The measure we use to gauge closeness (or consistency) between an estimator and the parameter is called *mean square error*, which is the sum of (bias)² and variance. Since \bar{X} is an unbiased estimator of μ , the bias is zero. Furthermore, the variance of \bar{X} ($= \sigma^2/n$) tends to zero as the sample size n becomes large. Therefore, the mean square error tends to zero as n becomes large. That is, the difference between the estimator \bar{X} and the parameter μ becomes smaller as n becomes large and therefore \bar{X} is a consistent estimator of μ . As a consequence, this implies that, as n grows larger, an increasing proportion of the statistic \bar{X} falls close to μ .

Figure 10.3 depicts two sampling distributions of \bar{X} when samples are drawn from a population whose mean is 0 and whose standard deviation is 10. One sampling distribution is based on samples of size 25, and the other on samples of size 100. The former is more spread out than the latter.

FIGURE 10.3 Sampling distributions of \bar{X} with $n = 25$ and $n = 100$



Similarly, \hat{p} is a consistent estimator of p because it is unbiased and the variance of \hat{p} is $p(1-p)/n$, which becomes smaller as n grows larger.

A third desirable quality is **relative efficiency**, which compares two unbiased estimators of a parameter.

We have already seen that the sample mean is an unbiased estimator of the population mean and that its variance is σ^2/n . Statisticians have established that the sample median is

Relative efficiency

If there are two unbiased estimators of a parameter, the one whose variance is smaller is said to have **relative efficiency**.

also an unbiased estimator, but that its variance is greater than that of the sample mean (when the population is normal). As a consequence, the sample mean is relatively more efficient than the sample median when estimating the population mean.

In the remaining chapters of this book, we will present the statistical inference of a number of different population parameters. In each case, we will select a sample statistic that is unbiased and consistent. When there is more than one such statistic, we will choose the one that is relatively efficient to serve as the estimator.

10.1c Developing an understanding of statistical concepts

In this section we described three desirable characteristics of estimators: unbiasedness, consistency and relative efficiency. An understanding of statistics requires that you know that there are several potential estimators for each parameter, but that we choose the estimators used in this book because they possess these characteristics.

EXERCISES

- 10.1** How do point estimators and interval estimators differ?
- 10.2** Define ‘unbiasedness’.
- 10.3** Draw a sampling distribution of an unbiased estimator.
- 10.4** Draw a sampling distribution of a biased estimator.
- 10.5** Define ‘consistency’.
- 10.6** Draw diagrams representing what happens to the sampling distribution of a consistent estimator when the sample size increases.
- 10.7** Define ‘relative efficiency’.
- 10.8** Draw a diagram that shows the sampling distribution representing two unbiased estimators, one of which is relatively efficient.

10.2 Estimating the population mean μ when the population variance σ^2 is known

In this section we discuss how to estimate an unknown population mean μ when the population variance σ^2 is known. We admit that this is quite unrealistic; if the population mean is unknown, it is quite unlikely that we would know the value of the population variance. However, this approach allows us to introduce the subject and then progress to more realistic situations later.

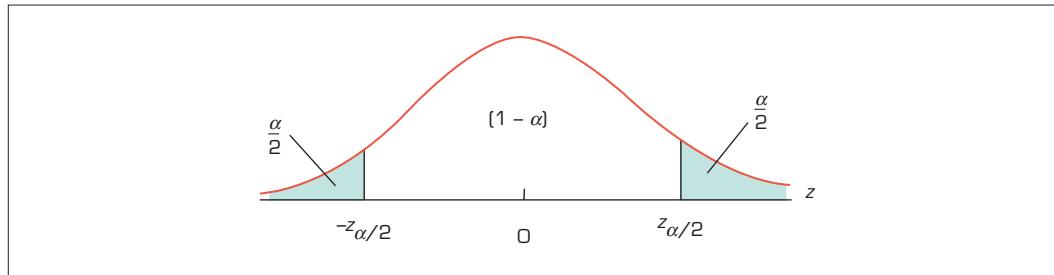
In order to proceed, recall our discussion of sampling distributions in Chapter 9. At that time we showed that if we repeatedly draw samples of size n from a population whose mean and variance are μ and σ^2 respectively, then the sample mean \bar{X} will be normally distributed (or approximately so, using the central limit theorem), with mean μ and variance σ^2/n . (Recall

also that the population is assumed to be normal; and if that is not the case, n must be sufficiently large; that is, n is at least ≥ 30 .) This means that the variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

follows a *standard normal distribution* (or approximately so). **Figure 10.4** describes this distribution. (Recall from Chapter 8 that $z_{\alpha/2}$ represents the point such that the area to its right under the standard normal curve is equal to $\alpha/2$. For example, for $z_{0.025}$, this area will be 0.025.)

FIGURE 10.4 Sampling distribution of $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$



10.2a Constructing a confidence interval estimate for μ when σ^2 is known

When a sample is drawn, the value of \bar{X} calculated from the sample is such that

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

can be anywhere along the horizontal axis (theoretically, from $-\infty$ to ∞).

We know, however, that the probability that z falls between $-z_{\alpha/2}$ and $+z_{\alpha/2}$ is equal to $(1 - \alpha)$. (For example, when $1 - \alpha = 0.95$, we have $z_{\alpha/2} = z_{0.025} = 1.96$.) This can be expressed as

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

With some algebraic manipulation, we can write this in the following form:

$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Notice that in this form the population mean is in the centre of the interval created by adding and subtracting $z_{\alpha/2}$ standard errors to the sample mean. It is important for you to understand that this is merely another form of probability statement about the sample mean. This equation says that, with repeated sampling from this population, the proportion of values of \bar{X} for which the interval

$$\left[\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right]$$

includes the population mean μ is equal to $(1 - \alpha)$. However, this form of probability statement is very useful to us because it is the **confidence interval estimator** of μ .

confidence interval estimator

An interval estimator in which we have a certain degree of confidence that it contains the value of the parameter.

Confidence interval estimator of μ

$$\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

The probability $(1 - \alpha)$ is called the confidence level.

$\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ is called the **lower confidence limit (LCL)**.

$\bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ is called the **upper confidence limit (UCL)**.

We often represent the confidence interval estimator as

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where the minus sign defines the lower confidence limit and the plus sign defines the upper confidence limit.

lower confidence limit (LCL)

The lower bound of the confidence interval.

upper confidence limit (UCL)

The upper bound of the confidence interval.

To apply this formula, we specify the level of confidence $(1 - \alpha)$, from which we determine α , $\alpha/2$ and $z_{\alpha/2}$ (from Table 3 in Appendix B). Because the level of confidence is the probability that the interval includes the actual value of μ , we generally set $(1 - \alpha)$ close to 1 (usually between 0.90 and 0.99).

In **Table 10.1**, we list three commonly used confidence levels and their associated values of $z_{\alpha/2}$. For example, if the level of confidence is $(1 - \alpha) = 0.95$, $\alpha = 0.05$, $\alpha/2 = 0.025$, and $z_{\alpha/2} = z_{0.025} = 1.96$. The resulting confidence interval estimator is then called the **95% confidence interval estimator of μ** .

TABLE 10.1 Four commonly used confidence levels and $z_{\alpha/2}$

Confidence level $(1 - \alpha)$	α	$\alpha/2$	$z_{\alpha/2}$
0.90	0.10	0.05	$z_{0.05} = 1.645$
0.95	0.05	0.025	$z_{0.025} = 1.96$
0.98	0.02	0.01	$z_{0.01} = 2.33$
0.99	0.01	0.005	$z_{0.005} = 2.575$

The following example illustrates how statistical techniques are applied using the three-stage solution process outlined in Section 9.3. It also illustrates how we intend to solve problems in the rest of this book. The process is divided into three stages: (1) *identify* the appropriate statistical techniques, (2) *calculate* the necessary statistics manually or using computer software, and (3) *interpret* the results and answer the question presented in the problem.

EXAMPLE 10.1

L04 L05

Number of hours children spend watching television

XM10-01 The sponsors of television shows targeted at children wanted to know the amount of time children spend watching television, since the types and number of programs and commercials presented are greatly influenced by this information. As a result, a survey was conducted to estimate the average number of hours Australian children spend watching television per week. From past experience, it is known that the population standard deviation σ is 8.0 hours. The following are the data gathered from a sample of 100 children. Find the 95% confidence interval estimate of the average number of hours Australian children spend watching television.

Amount of time spent watching television each week

39.7	21.5	40.6	15.5	43.9	33.0	21.0	15.8	27.1	23.8	18.3	23.4	20.6
28.4	29.8	41.3	36.8	35.5	27.2	21.0	19.7	22.8	30.0	22.1	30.8	34.7
15.0	23.6	38.9	29.1	28.7	29.3	20.3	36.1	21.6	15.1	43.8	29.0	30.2
26.5	20.5	24.1	29.3	14.7	13.9	37.1	32.5	24.4	22.9	24.5	19.5	29.9
46.4	31.6	20.6	38.0	21.8	23.2	22.0	35.3	17.0	24.4	34.9	24.0	32.9
15.1	23.4	19.5	26.5	42.4	38.6	23.4	37.8	26.5	22.7	27.0	16.4	39.4
38.7	9.5	20.6	21.3	33.5	23.0	35.7	23.4	30.8	27.7	25.2	50.3	31.3
28.9	31.2	15.6	32.8	17.0	11.3	26.9	26.9	21.9				

Solution

Identifying the technique

We recognise that the problem objective is to describe the population of time spent by Australian children watching television per week (X) where the data type is numerical.

The parameter to be estimated is μ , the average (mean) number of hours of television watched by all Australian children. The population standard deviation σ is known ($\sigma = 8$). Since $n = 100 > 30$, using the central limit theorem, the sample mean \bar{X} is approximately normal. Therefore, the confidence interval estimator of a population mean μ is

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

The next step is to perform the calculations. As discussed above, we will perform the calculations in two ways: manually and using Excel.

Calculating manually

We need the following four values to calculate the confidence interval estimate of μ :

$$\bar{X}, z_{\alpha/2}, \sigma, n$$

Using the calculator, $\sum x_i = 2719.1$. Therefore,

$$\bar{X} = \frac{\sum x_i}{n} = \frac{2719.1}{100} = 27.191$$

The confidence level is set at 95%; thus,

$$1 - \alpha = 0.95; \alpha = 1 - 0.95 = 0.05 \text{ and } \alpha/2 = 0.025$$

From Table 3 in Appendix B, we obtain

$$z_{\alpha/2} = z_{0.025} = 1.96$$

The population standard deviation is $\sigma = 8$ and the sample size is $n = 100$. Substituting the values for \bar{X} , $z_{\alpha/2}$, σ and n in the confidence interval estimator formula, the 95% confidence interval estimate is

$$\begin{aligned} \bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} &= 27.191 \pm 1.96 \frac{8.0}{\sqrt{100}} \\ &= 27.191 \pm 1.568 \\ &= [25.623, 28.759] \end{aligned}$$

We therefore estimate that the average number of hours children spend watching television each week lies somewhere between LCL = 25.62 hours and UCL = 28.76 hours.

Interpreting the results

From this estimate, a network executive may decide (for example) that, as the average child watches at least 25.62 hours of television per week, the number of commercials children see is sufficiently high to satisfy the program sponsors. A number of other decisions may follow.

Of course, the point estimate ($\bar{X} = 27.191$ hours per week) alone would not provide enough information to the executive. He would also need to know how low the population mean is likely to be; and for other decisions, he might need to know how high the population mean is likely to be. A confidence interval estimate gives him that information.

Using the computer

EXCEL Workbook

Data Analysis cannot be used to produce interval estimates. To produce the interval estimate for this problem, if you can calculate the sample mean and know the sample size and the population standard deviation, you can use the **Estimators** workbook, which is available in the Workbooks tab on the companion website (accessible through <https://login.cengagebrain.com/>). In this example, as the population standard deviation is known, we use the **z-Estimate_Mean** worksheet.

Excel output for Example 10.1

	A	B	C	D	E
1	z-Estimate of a Mean				
2					
3	Sample mean	27.19	Confidence Interval Estimate		
4	Population standard deviation	8.00		27.19	\pm 1.57
5	Sample size	100		Lower confidence limit	25.62
6	Confidence level	0.95		Upper confidence limit	28.76

COMMANDS

- Type the data into one column or open the data file (**XM10-01**). In any empty cell, calculate the sample mean (**=AVERAGE(A1:A101)**).
- Open the **Estimators** workbook (Estimators.xlsx) and click the **z-Estimate_Mean** worksheet. In cell B3, type or copy the value of the sample mean. If you use **Copy** also use **Paste Special** and **Values**. In cells B4–B6, type the value of σ (**8.0**), the value of n (**100**), and the confidence level (**0.95**) respectively.

In addition to providing a method of using Excel, this spreadsheet allows you to perform a ‘what-if’ analysis. That is, this worksheet provides you with the opportunity to learn how changing some of the inputs affects the estimate. For example, type 0.99 in cell B6 to see what happens to the size of the interval when you increase the confidence level. Then type 1000 in cell B5 to examine the effect of increasing the sample size, and type 4 in cell B4 to see what happens when the population standard deviation is smaller.

XLSTAT

XLSTAT output for Example 10.1

	A	B	C	D
1	95% confidence interval on the mean:			
2	[25.623, 28.759]			

(Note: This is only a partial output).

COMMANDS

- Type the data into one column or open the data file (**XM10-01**).
- Click **XLSTAT** and **One-sample t-test and z-test**.
- In the **Data**: dialog box type the input range (**A1:A100**). Click **Column labels** if the first row contains the name of the variable (as in this example). Check **z-test** but do not check Student’s t-test.
- Click the **Options** tab and choose **Mean ≠ Theoretical mean** in the **Alternative hypothesis**: box. Type the value of α (in per cent) in the **Significance**: box (**5**). If there are blanks in the column (usually used to represent missing data) click **Missing data**, Remove the observations. For the Variance for z-test: check **User defined: Variance**: and type the value of σ^2 (**64.0**). Click **OK** and then **Continue**.

Reconsider Example 10.1. Recall that the objective was to describe the population of hours that children spend watching television per week. The type of data is numerical, since what is being measured is the amount of time. To understand this point more fully, imagine that the 100 children in the sample were also asked whether or not they believed the message being conveyed by the commercials. Their responses to this question could be yes or no, and such responses produce nominal data because the results are not numerical. Therefore, some other technique (to be discussed in Section 10.4) would have to be used to assess the data.

We can summarise the factors that determine when to use the z -interval estimator of μ as follows.

IN SUMMARY

Factors that identify the z -interval estimator of μ

- 1 Problem objective:** to describe a single population
- 2 Data type:** numerical (quantitative)
- 3 Type of descriptive measurement:** central location
- 4 Population variance:** known

Before doing another example, let us review the steps used to estimate a population mean when the population variance σ is known.

IN SUMMARY

Steps in estimating a population mean when the population variance is known

- 1** Determine the sample mean \bar{X} .
- 2** Determine the desired confidence level $(1 - \alpha)$, which in turn specifies α . From Table 3 in Appendix B, find $z_{\alpha/2}$.
- 3** Calculate $LCL = \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ and $UCL = \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$.

10.2b Interpreting the confidence interval estimate

In Example 10.1, we found the 95% confidence interval estimate of the average (mean) number of hours that children spend watching television per week to be $LCL = 25.623$ and $UCL = 28.759$. Some people erroneously interpret this interval to mean that there is a 95% probability that the population mean lies between 25.623 and 28.759. This interpretation is incorrect because it implies that the population mean is a variable about which we can make probability statements. In fact, the population mean is a fixed but unknown quantity. Consequently, we cannot interpret the confidence interval estimate of μ as a probability statement about μ .

To translate the interval estimate properly, we must recall that the interval estimator was derived from the sampling distribution, which enables us to make probability statements about the sample mean. Thus, we say that the 95% confidence interval estimator of μ implies that 95% of the values of \bar{X} will create intervals that will contain the true value of the population mean. The other 5% of sample means will create intervals that do not include the population mean. (That is, 95% of the interval estimates will be right and 5% will be wrong.)

As an illustration, suppose we want to estimate the mean value of the distribution resulting from the throw of a die. Because we know the distribution, we also know that $\mu = 3.5$ and $\sigma = 1.71$ (see page 355, Section 9.4a). Pretend now that we only know that $\sigma = 1.71$ and the population mean μ is unknown, and that we want to estimate its value.

In order to estimate μ , we draw a sample of size $n = 100$ and calculate \bar{X} . The 90% confidence interval estimator of μ is:

$$\bar{X} \pm z_{0.05} \frac{\sigma}{\sqrt{n}}$$

The 90% confidence interval estimator is:

$$\bar{X} \pm z_{0.05} \frac{\sigma}{\sqrt{n}} = \bar{X} \pm 1.645 \frac{1.71}{\sqrt{100}} = \bar{X} \pm 0.28$$

This interval estimator means that, if we repeatedly draw samples of size 100 from this population, 90% of the values of \bar{X} will be such that μ would lie somewhere between $\bar{X} - 0.28$ and $\bar{X} + 0.28$, and 10% of the values of \bar{X} will produce intervals that would not include μ . To illustrate this point, imagine that we draw 40 samples of 100 observations each. The values of \bar{X} and the resulting confidence interval estimates of μ are shown in **Table 10.2**. Notice that not all the intervals include the true value of the parameter $\mu = 3.5$. Samples 5, 16, 22 and 34 produce values of \bar{X} that, in turn, produce confidence intervals that exclude μ .

TABLE 10.2 90% confidence interval estimates of μ

Sample	\bar{X}	LCL = $\bar{X} - 0.28$	UCL = $\bar{X} + 0.28$	Does interval include $\mu = 3.5$?	Sample	\bar{X}	LCL = $\bar{X} - 0.28$	UCL = $\bar{X} + 0.28$	Does interval include $\mu = 3.5$?
1	3.55	3.27	3.83	Yes	21	3.40	3.12	3.68	Yes
2	3.61	3.33	3.89	Yes	22	3.88	3.60	4.16	No
3	3.47	3.19	3.75	Yes	23	3.76	3.48	4.04	Yes
4	3.48	3.20	3.76	Yes	24	3.40	3.12	3.68	Yes
5	3.80	3.52	4.08	No	25	3.34	3.06	3.62	Yes
6	3.37	3.09	3.65	Yes	26	3.65	3.37	3.93	Yes
7	3.48	3.20	3.76	Yes	27	3.45	3.17	3.73	Yes
8	3.52	3.24	3.8	Yes	28	3.47	3.19	3.75	Yes
9	3.74	3.46	4.02	Yes	29	3.58	3.30	3.86	Yes
10	3.51	3.23	3.79	Yes	30	3.36	3.08	3.64	Yes
11	3.23	2.95	3.51	Yes	31	3.71	3.43	3.99	Yes
12	3.45	3.17	3.73	Yes	32	3.51	3.23	3.79	Yes
13	3.57	3.29	3.85	Yes	33	3.42	3.14	3.7	Yes
14	3.77	3.49	4.05	Yes	34	3.11	2.83	3.39	No
15	3.31	3.03	3.59	Yes	35	3.29	3.01	3.57	Yes
16	3.10	2.82	3.38	No	36	3.64	3.36	3.92	Yes
17	3.50	3.22	3.78	Yes	37	3.39	3.11	3.67	Yes
18	3.55	3.27	3.83	Yes	38	3.75	3.47	4.03	Yes
19	3.65	3.37	3.93	Yes	39	3.26	2.98	3.54	Yes
20	3.28	3.00	3.56	Yes	40	3.54	3.26	3.82	Yes

Students often react to this situation by asking, what went wrong with samples 5, 16, 22 and 34? The answer is nothing. Statistics does not promise 100% certainty. In fact, in this illustration, we expected 90% of the 40 intervals to include μ and 10% to exclude μ . Since we produced 40 confidence intervals, we expected that 4 (10% of 40) intervals would not contain $\mu = 3.5$.¹ It is important to understand that, even when the statistics practitioner performs

¹ In this illustration, exactly 10% of the 40 sample means produced interval estimates that excluded the value of μ , but this will not always be the case. Remember, we expect 10% of the sample means in the long run to result in intervals excluding μ . This group of 40 sample means does not constitute ‘the long run’.

experiments properly, a certain proportion of the experiments (in this example, 10%) will produce incorrect estimates by random chance.

We can improve the confidence associated with the interval estimate. If we let the confidence level $(1 - \alpha)$ equal 0.95, the confidence interval estimator is

$$\bar{X} \pm z_{0.025} \frac{\sigma}{\sqrt{n}} = \bar{X} \pm 1.96 \frac{1.71}{\sqrt{100}} = \bar{X} \pm 0.34$$

Because this interval is wider, it is more likely to include the value of μ . If you reconstruct the intervals in **Table 10.2**, this time using a 95% confidence interval estimator, only samples 16, 22 and 34 will produce intervals that do not include μ . (Notice that we expected 5% of the intervals to exclude μ and that we actually observed $3/40 = 7.5\%$.) The 99% confidence interval estimate is

$$\bar{X} \pm z_{0.005} \frac{\sigma}{\sqrt{n}} = \bar{X} \pm 2.575 \frac{1.71}{\sqrt{100}} = \bar{X} \pm 0.44$$

Applying this confidence interval to the sample means listed in **Table 10.2** would result in having all 40 interval estimates include the population mean $\mu = 3.5$. (We expected 1% of the intervals to exclude μ ; we observed $0/40 = 0\%$.)

In actual practice only one sample will be drawn and thus only one value of \bar{X} will be calculated. The resulting interval estimate will either correctly include the parameter or incorrectly exclude it. Unfortunately, statistics practitioners do not know whether they are correct in each case; they know only that, in the long run, they will incorrectly estimate the parameter some of the time. Statistics practitioners accept that as a fact of life.

We summarise our calculations in Example 10.1 as follows. We estimate that the mean number of hours Australian children spend watching television falls between 25.6 and 28.8 hours, and this type of estimator is correct 95% of the time. Thus, the confidence level applies to the estimation procedure and not to any one interval. Incidentally, the media often refer to the 95% figure as '19 times out of 20', which emphasises the long-run aspect of the confidence level.

10.2c Information and the width of the interval

Interval estimation, like all other statistical techniques, is designed to convert data into information. However, a wide interval provides little information. For example, suppose that as a result of a statistical study, we estimate with 95% confidence that the mean starting salary of an accountant lies between \$44 000 and \$100 000. This interval is so wide that very little information is derived from the data. Suppose, however, that the interval estimate was \$66 000 to \$77 000. This interval is much narrower, providing accounting students more precise information about starting salaries.

The width of the confidence interval estimate is a function of the population standard deviation, the confidence level and the sample size. Consider Example 10.1, in which σ was assumed to be 8. The 95% interval estimate was 27.191 ± 1.568 . If σ equalled 16, the 95% confidence interval estimate would become

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 27.191 \pm z_{0.025} \frac{16}{\sqrt{100}} = 27.191 \pm 1.96 \frac{16}{\sqrt{100}} = 27.191 \pm 3.136$$

Thus, doubling the population standard deviation has the effect of doubling the width of the confidence interval estimate. This result is quite logical. If there is a great deal of variation in the random variable (reflected by a large standard deviation), it is more difficult to accurately estimate the population mean. That difficulty is translated into a wider interval.

Although we have no control over the value of σ , we do have the power to select values for the other two elements. In Example 10.1, we chose a 95% confidence level. If we had chosen 99% instead, the interval estimate would have been

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 27.191 \pm z_{0.005} \frac{8}{\sqrt{100}} = 27.191 \pm 2.575 \frac{8}{\sqrt{100}} = 27.191 \pm 2.06$$

A 90% confidence level results in this interval estimate:

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 27.191 \pm z_{0.05} \frac{8}{\sqrt{100}} = 27.191 \pm 1.645 \frac{8}{\sqrt{100}} = 27.191 \pm 1.316$$

As you can see, increasing the confidence level will widen the interval. Similarly, decreasing the confidence level will narrow the interval. However, a large confidence level is generally desirable since that means a larger proportion of confidence interval estimates that will be correct in the long run. There is a direct relationship between the confidence level and the width of the interval. This is because in order to be more confident in the estimate we need to widen the interval. (The analogy is that to be more likely to capture a butterfly, we need a larger butterfly net.) The trade-off between increased confidence and the resulting wider confidence interval estimates must be resolved by the statistics practitioner. As a general rule, however, the 95% confidence level is considered 'standard'.

The third element is the sample size. Had the sample size been 400 instead of 100, the 95% confidence interval estimate would become:

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 27.191 \pm z_{0.025} \frac{8}{\sqrt{400}} = 27.191 \pm 1.96 \frac{8}{\sqrt{400}} = 27.191 \pm 0.784$$

Increasing the sample size fourfold decreases the width of the interval by half. A larger sample size provides more potential information. The increased amount of information is reflected in a narrower interval. However, there is another trade-off: increasing the sample size increases the sampling cost. We will discuss these issues when we present sample-size selection in Section 10.5.

10.2d (Optional) Estimating the population mean using the sample median

To understand why the sample mean is most often used to estimate a population mean, let's examine the properties of the sampling distribution of the sample median (denoted here as m). The sampling distribution of a sample median is normally distributed provided that the population is normal. Its mean and standard deviation are

$$\mu_m = \mu$$

and

$$\sigma_m = \frac{1.2533\sigma}{\sqrt{n}}$$

Using the same algebraic steps as used above, we derive the confidence interval estimator of a population mean using the sample median:

$$m \pm z_{\alpha/2} \frac{1.2533\sigma}{\sqrt{n}}$$

To illustrate, suppose that we have drawn the following random sample from a normal population whose standard deviation is 2.

1 1 1 3 4 5 6 7 8

The sample mean is $\bar{X} = 4$, and the median is $m = 4$.

The 95% confidence interval estimates using the sample mean and the sample median are

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 4.0 \pm 1.96 \frac{2}{\sqrt{9}} = 4.0 \pm 1.307$$

$$m \pm z_{\alpha/2} \frac{1.2533\sigma}{\sqrt{n}} = 4.0 \pm 1.96 \frac{(1.2533)(2)}{\sqrt{9}} = 4.0 \pm 1.638$$

As you can see, the interval based on the sample mean is narrower; as we pointed out previously, narrower intervals provide more precise information. To understand why the sample mean produces better estimators than the sample median, recall how the median is calculated. We simply put the data in order and select the observation that falls in the middle. Thus, as far as the median is concerned the data appear as

1 2 3 4 5 6 7 8 9

By ignoring the actual observations and using their ranks instead, we lose information. With less information, we have less precision in the interval estimators and so ultimately make poorer decisions.

EXERCISES

Learning the techniques

- 10.9** In a random sample of 400 observations from a population whose variance is $\sigma^2 = 100$, we calculated $\bar{X} = 75$. Find the 95% confidence interval estimate of the population mean μ .

- 10.10** Suppose that a random sample of five observations was taken from a normal population whose variance is 25. The results are 8, 15, 12, 6, 7. Find the 99% confidence interval estimate of the population mean.

- 10.11** A random sample of 400 observations from a population whose standard deviation is 90 produced $\bar{X} = 1500$. Find the 90% confidence interval estimate of μ .

- 10.12** The following observations were drawn from a normal population whose variance is 100:

12 8 22 15 30 6 39 48

Determine the 90% confidence interval of the population mean.

- 10.13** Describe what happens to the width of a confidence interval estimate of μ when each of the following occurs:
- The confidence level increases from 95% to 99%.
 - The sample size decreases.
 - The value of σ increases.

Exercises 10.14–10.21 are ‘what-if analyses’ designed to determine what happens to the interval estimate when the confidence level, the sample size and the standard

deviation change. These problems can be solved manually or by using the Excel **z-Estimate_Mean** worksheet in the **Estimators** workbook.

- 10.14** **a** A statistics practitioner took a random sample of 50 observations from a population whose standard deviation is 25 and computed the sample mean to be 100. Estimate the population mean with 90% confidence.
- b** Repeat part (a) using a 95% confidence level.
- c** Repeat part (a) using a 99% confidence level.
- d** Describe the effect on the confidence interval estimate of increasing the confidence level.
- 10.15** **a** A random sample of 25 observations was drawn from a normal population whose standard deviation is 50. The sample mean was 200. Estimate the population mean with 95% confidence.
- b** Repeat part (a) changing the population standard deviation to 25.
- c** Repeat part (a) changing the population standard deviation to 10.
- d** Describe what happens to the confidence interval estimate when the standard deviation is decreased.

- 10.16** **a** A random sample of 25 was drawn from a normal distribution whose standard deviation is 5. The sample mean was 80. Determine the 95% confidence interval estimate of the population mean.

- b** Repeat part (a) with a sample size of 100.
c Repeat part (a) with a sample size of 400.
d Describe what happens to the confidence interval estimate when the sample size increases.
- 10.17 a** Given the following information, determine the 98% confidence interval estimate of the population mean:
- $$\bar{X} = 500 \quad \sigma = 12 \quad n = 50$$
- b** Repeat part (a) using a 95% confidence level.
c Repeat part (a) using a 90% confidence level.
d Review parts (a) to (c) and discuss the effect on the confidence interval estimator of decreasing the confidence level.
- 10.18 a** The mean of a sample of 25 was calculated as $\bar{X} = 500$. The sample was randomly drawn from a normal population whose standard deviation is 15. Estimate the population mean with 99% confidence.
- b** Repeat part (a) changing the population standard deviation to 30.
c Repeat part (a) changing the population standard deviation to 60.
d Describe what happens to the confidence interval estimate when the standard deviation is increased.
- 10.19 a** A statistics practitioner randomly sampled 100 observations from a normal population whose standard deviation is 5 and found that $\bar{X} = 10$. Estimate the population mean with 90% confidence.
- b** Repeat part (a) with a sample size of 25.
c Repeat part (a) with a sample size of 10.
d Describe what happens to the confidence interval estimate when the sample size decreases.
- 10.20 a** From the information given here for a sample taken from a normal population, determine the 95% confidence interval estimate of the population mean:
- $$\bar{X} = 100 \quad \sigma = 20 \quad n = 25$$
- b** Repeat part (a) with $\bar{X} = 200$.
c Repeat part (a) with $\bar{X} = 500$.
d Describe what happens to the width of the confidence interval estimate when the sample mean increases.
- 10.21 a** A random sample of 100 observations was randomly drawn from a population whose standard deviation is 5. The sample mean was

calculated as $\bar{X} = 400$. Estimate the population mean with 99% confidence.

b Repeat part (a) with $\bar{X} = 200$.
c Repeat part (a) with $\bar{X} = 100$.
d Describe what happens to the width of the confidence interval estimate when the sample mean decreases.

Exercises 10.22 –10.26 are based on the optional subsection ‘Estimating the population mean using the sample median’ on pages 386–7. All exercises assume that the population is normal.

- 10.22** Is the sample median an unbiased estimator of the population mean? Explain.
- 10.23** Is the sample median a consistent estimator of the population mean? Explain.
- 10.24** Show that the sample mean is relatively more efficient than the sample median when estimating the population mean.
- 10.25 a** Given the following information, determine the 90% confidence interval estimate of the population mean using the sample median.

$$\text{Sample median} = 500 \quad \sigma = 12 \quad n = 50$$

- b** Compare your answer in part (a) with that produced in part (c) of Exercise 10.17. Why is the confidence interval estimate based on the sample median wider than that based on the sample mean?

Applying the techniques

- 10.26 Self-correcting exercise.** In a survey conducted to determine, among other things, the cost of holidays, 164 individuals were randomly sampled. Each was asked to assess the total cost of his or her most recent holiday. The average cost was \$2772. Assuming that the population standard deviation was \$800, estimate the population mean cost of holidays with 99% confidence.
- 10.27** A survey of 20 Australian companies indicated that the average annual income of company secretaries was \$120 000. Assuming that the population standard deviation is \$7500 and that the annual incomes are normally distributed, calculate the 90% confidence interval estimate of the average annual income of all company secretaries.

- 10.28** In a random sample of 70 students in a large university, a dean found that the mean weekly

time devoted to homework was 14.3 hours. If we assume that homework time is normally distributed, with a population standard deviation of 4.0 hours, find the 99% confidence interval estimate of the weekly time spent doing homework for all the university's students.

- 10.29** To determine the mean waiting time for his customers, a bank manager took a random sample of 50 customers and found that the mean waiting time was 7.2 minutes. Assuming that the population standard deviation is known to be 5 minutes, find the 90% confidence interval estimate of the mean waiting time for all of the bank's customers.

Computer/manual applications

The following exercises may be answered manually or with the assistance of a computer and software. Excel's **Estimators** workbook can also be used for this purpose.

- 10.30 XR10-30** The following data represent a random sample of 9 marks (out of 10) on a statistics quiz. The marks are normally distributed with a population standard deviation of 2. Estimate the population mean with 90% confidence.

7	9	7	5	4	8	3	10	9
---	---	---	---	---	---	---	----	---

- 10.31 XR10-31** The following observations are the ages of a random sample of eight men in a bar. It is known that the ages are normally distributed with a population standard deviation of 10. Determine the 95% confidence interval estimate of the population mean. Interpret the interval estimate.

52	68	22	35	30	56	39	48
----	----	----	----	----	----	----	----

- 10.32 XR10-32** How many rounds of golf do physicians (who play golf) play per year? A survey of 12 physicians revealed the following numbers:

3	41	17	1	33	37	18	15	17	12	29	51
---	----	----	---	----	----	----	----	----	----	----	----

Estimate with 95% confidence the mean number of rounds per year played by these physicians, assuming that the number of rounds is normally distributed with a population standard deviation of 12.

- 10.33 XR10-33** Among the most exciting aspects of a university lecturer's life are the departmental meetings where such critical issues as the colour the walls will be painted and who gets a new desk are decided. A sample of 20 lecturers was asked how many hours per year are devoted to

these meetings. The responses are listed here. Assuming that hours spent on meetings is normally distributed with a standard deviation of 8 hours, estimate the mean number of hours spent at departmental meetings by all lecturers. Use a confidence level of 90%.

14	17	3	6	17	3	8	4	20	15
7	9	0	5	11	15	18	13	8	4

- 10.34 XR10-34** The number of used cars sold annually by salespeople is normally distributed with a standard deviation of 15. A random sample of 15 salespeople was taken and the number of cars each sold is listed here. Find the 95% confidence interval estimate of the population mean. Interpret the interval estimate.

79	43	58	66	101	63	79	33
58	71	60	101	74	55	88	

- 10.35 XR10-35** It is known that the amount of time needed to change the oil in a car is normally distributed with a standard deviation of 5 minutes. The amount of time (in minutes) to complete a random sample of 10 oil changes was recorded and listed here. Compute the 99% confidence interval estimate of the mean of the population.

11	10	16	15	18	12	25	20	18	24
----	----	----	----	----	----	----	----	----	----

- 10.36 XR10-36** Suppose that the amount of time teenagers spend weekly working at part-time jobs is normally distributed with a standard deviation of 40 minutes. A random sample of 15 teenagers was drawn and each reported the amount of time (in minutes) spent at part-time jobs. These are listed here. Determine the 95% confidence interval estimate of the population mean.

180	130	150	165	90	130	120	60
200	180	80	240	210	150	125	

- 10.37 XR10-37** One of the few negative side effects of quitting smoking is weight gain. Suppose that the weight gain in the 12 months following a cessation in smoking is normally distributed with a standard deviation of 3 kg. To estimate the mean weight gain, a random sample of 13 quitters was drawn and their weight gains recorded and listed here. Determine the 90% confidence interval estimate of the mean 12-month weight gain for all quitters.

8	12	4	1	7	11	9	6	5	9	3	4	7
---	----	---	---	---	----	---	---	---	---	---	---	---

- 10.38 XR10-38** Because of different sales ability, experience and devotion, the incomes of real estate agents vary considerably. Suppose that in a large city the annual income is normally distributed with a standard deviation of \$15000. A random sample of 16 real estate agents was asked to report their annual income (in \$'000). The responses are recorded and listed here. Determine the 99% confidence interval estimate of the mean annual income of all real estate agents in the city.

65	94	57	111	83	61	50	73
68	80	93	84	113	41	60	77

The following exercises require the use of a computer and software. The answers may also be calculated manually using the sample statistics based on the data provided.

- 10.39 XR10-39** A random sample of 400 observations was drawn from a population whose standard deviation is 90. Some of the observations are shown below. Estimate the population mean with 95% confidence.

895	961	1007	...	871	1132	906
-----	-----	------	-----	-----	------	-----

Sample statistics: $\bar{X} = 1010$; $\sigma = 90$; $n = 400$.

- 10.40 XR10-40** In an article about disinflation, various investments were examined. The investments included shares, bonds and real estate. Suppose that a random sample of 200 rates of return on real estate investments were calculated and stored. Some of these data are shown below. Assuming that the standard deviation of all rates of return on real estate investments is 2.1%, estimate the mean rate of return on all real estate investments with 90% confidence. Interpret the estimate.

11.63	10.43	14.92	...	10.58	12.79
-------	-------	-------	-----	-------	-------

Sample statistics: $\bar{X} = 12.1$; $\sigma = 2.1$; $n = 200$.

- 10.41 XR10-41** A statistics lecturer is in the process of investigating how many classes university students miss each semester. To help answer this question, she took a random sample of 100 university students and asked each to report how many classes he or she had missed in the previous semester. These data are recorded. (Some of these data are listed below.) Estimate the mean number of classes missed by all students at the university. Use a 99% confidence level and assume that the population standard deviation is known to be 2.2 classes.

4	0	1	6	...	3	5	4
---	---	---	---	-----	---	---	---

Sample statistics: $\bar{X} = 3.88$; $\sigma = 2.2$; $n = 100$.

- 10.42 XR10-42** The image of the Japanese manager is that of a workaholic with little or no leisure time. In a survey, a random sample of 250 Japanese middle managers was asked how many hours per week they spent in leisure activities (e.g. sports, movies, television). The results of the survey are recorded and stored. Assuming that the population standard deviation is six hours, estimate with 90% confidence the mean leisure time per week for all Japanese middle managers. What do these results tell you?

Sample statistics: $\bar{X} = 19.28$; $\sigma = 6$; $n = 250$.

- 10.43 XR10-43** One measure of physical fitness is the amount of time it takes for the pulse rate to return to normal after exercise. A random sample of 100 women aged 40–50 exercised on stationary bicycles for 30 minutes. The amount of time it took for their pulse rates to return to pre-exercise levels was measured and recorded. If the times are normally distributed with a standard deviation of 2.3 minutes, estimate with 99% confidence the true mean pulse-recovery time for all women aged 40–50. Interpret the results.

Sample statistics: $\bar{X} = 15.00$; $\sigma = 2.3$; $n = 100$.

- 10.44 XR10-44** A survey of 80 randomly selected companies asked them to report the annual income of their chief executives. Assuming that incomes are normally distributed with a standard deviation of \$30000, determine the 90% confidence interval estimate of the mean annual income of all company chief executives. Interpret the statistical results.

Sample statistics: $\bar{X} = \$585\,063$; $\sigma = \$30\,000$; $n = 80$.

- 10.45 XR10-45** The supervisor of a production line that assembles computer keyboards has been experiencing problems since a new process was introduced. He notes that, when the productivity of one station does not match that of the others, this results in an increase in the number of defective units and of backlogs. To make the assembly line run more efficiently, he needs a 90% confidence interval estimate of the mean assembly time for all stations. He starts by drawing a sample of 75 completion times (in seconds) of the operation at the point where problems have been occurring. Based on historical experience, the supervisor knows that the population standard deviation of assembly times is 10 seconds. Interpret your results.

Sample statistics: $\bar{X} = 91.47$; $\sigma = 10$; $n = 75$.

10.3 Estimating the population mean μ when the population variance σ^2 is unknown

In the preceding section, the confidence interval estimator was constructed on the basis of knowing the value for σ^2 so that $Z = (\bar{X} - \mu) / (\sigma / \sqrt{n})$ is a standard normal random variable. This assumption is quite unrealistic because, if the population mean μ is unknown, it is unreasonable to believe that the population variance σ^2 would be known. As a result, $Z = (\bar{X} - \mu) / (\sigma / \sqrt{n})$ cannot be the basis of the confidence interval estimator. Because σ^2 is usually unknown, we estimate it by the sample variance s^2 , and we must substitute the sample standard deviation, s , in place of σ ; however, $(\bar{X} - \mu) / (s / \sqrt{n})$ is not normally distributed. In a 1908 paper, William Sealy Gosset (1876–1937), who used to publish under the pseudonym ‘Student’, showed that $(\bar{X} - \mu) / (s / \sqrt{n})$ has a particular distribution, called the **Student t distribution**, or more simply the **t distribution**, when the population sampled is normally distributed. The quantity $(\bar{X} - \mu) / (s / \sqrt{n})$ is called the **t-statistic**.

Student t distribution, or t distribution

A continuous distribution used in statistical inference when the population variance is not known.

t-statistic for μ

Standardised value of the sample mean when σ is unknown and replaced by s .

10.3a Student t distribution

Figure 10.5 depicts a Student *t* distribution. The mean and variance of a *t* random variable are

$$E(t) = 0$$

$$V(t) = \frac{n-1}{n-3} \quad \text{for } n > 3$$

As $(n-1) > (n-3)$, $V(t)$ is always greater than 1. As you can see, the Student *t* distribution is similar to the standard normal distribution. Like the standard normal distribution, the Student *t* distribution is symmetrical about zero. It is mound shaped, whereas the normal distribution is bell shaped. **Figure 10.6** shows both a Student *t* and a standard normal distribution. Although the variance of a standard normal distribution is 1, the variance of a *t* distribution is always greater than 1. Therefore, the *t* distribution is more widely dispersed than the standard normal distribution. The extent to which the Student *t* distribution is more spread out than the standard normal distribution is determined by a function of the sample size called the degrees of freedom (abbreviated as d.f. and denoted by v), which varies by the *t*-statistic. For this application, the degrees of freedom equals the sample size minus 1. That is, the degrees of freedom $v = n - 1$. **Figure 10.7** depicts Student *t* distributions with several different degrees of freedom. Notice that as the degrees of freedom grows larger, the dispersion of the Student *t* distribution becomes smaller.

FIGURE 10.5 Student *t* distribution

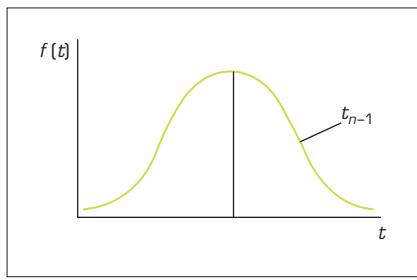


FIGURE 10.6 Student *t* and standard normal distribution

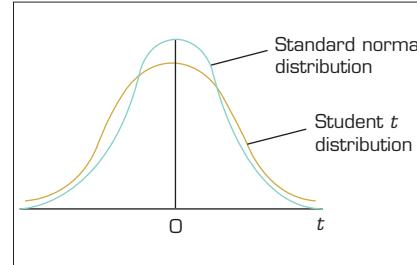
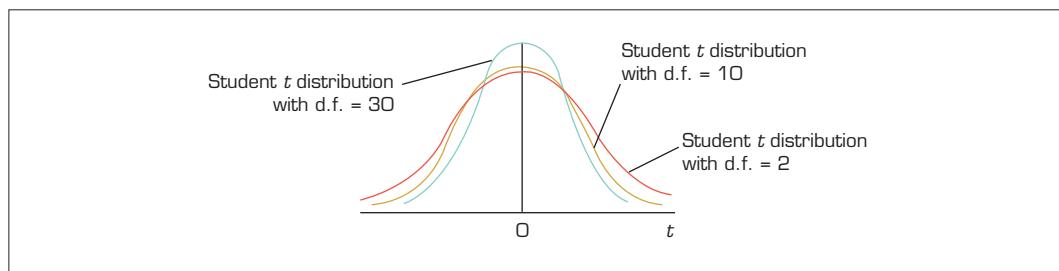
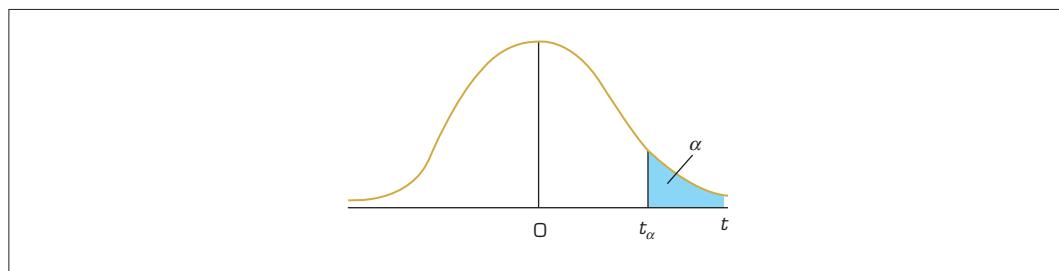


FIGURE 10.7 Student *t* distributions

For the purpose of calculating confidence interval estimates, as with $z_{\alpha/2}$ values, we need to be able to determine $t_{\alpha/2}$ values. Table 4 in Appendix B specifies values of t_α , where t_α equals the value of t for which the area to its right under the Student *t* curve is equal to α (see **Figure 10.8**). That is, $P(t > t_\alpha) = \alpha$. This table is reproduced as **Table 10.3**.

FIGURE 10.8 Student *t* value such that the area to its right under the curve is α 

Observe that in **Table 10.3**, t_α is provided for degrees of freedom (v) ranging from 1 to 200 and ∞ (infinity). To read this table, simply identify α (alpha) and the degrees of freedom and find that value or the closest number to it. Then locate the column representing the t_α value you want. We denote this value as $t_{\alpha,v}$. For example, if we want the value of t such that the right tail area under the Student *t* curve is 0.05 and the degrees of freedom is $v = 4$, we locate 4 in the first column and move across this row until we locate the value under the heading $t_{0.05}$. We find (see table below)

$$t_{0.05, 4} = 2.132$$

FINDING $t_{0.05,4}$

Degrees of freedom	$t_{0.10}$	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032

If the degrees of freedom is 25, we find (see **Table 10.3**)

$$t_{0.05, 25} = 1.708$$

TABLE 10.3 Reproduction of Table 4 in Appendix B: t values

v	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.010}$	$t_{0.005}$	v	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.010}$	$t_{0.005}$
1	3.078	6.314	12.71	31.82	63.66	29	1.311	1.699	2.045	2.462	2.756
2	1.886	2.920	4.303	6.965	9.925	30	1.310	1.697	2.042	2.457	2.750
3	1.638	2.353	3.182	4.541	5.841	35	1.306	1.690	2.030	2.438	2.724
4	1.533	2.132	2.776	3.747	4.604	40	1.303	1.684	2.021	2.423	2.704
5	1.476	2.015	2.571	3.365	4.032	45	1.301	1.679	2.014	2.412	2.690
6	1.440	1.943	2.447	3.143	3.707	50	1.299	1.676	2.009	2.403	2.678
7	1.415	1.895	2.365	2.998	3.499	55	1.297	1.673	2.004	2.396	2.668
8	1.397	1.860	2.306	2.896	3.355	60	1.296	1.671	2.000	2.390	2.660
9	1.383	1.833	2.262	2.821	3.250	65	1.295	1.669	1.997	2.385	2.654
10	1.372	1.812	2.228	2.764	3.169	70	1.294	1.667	1.994	2.381	2.648
11	1.363	1.796	2.201	2.718	3.106	75	1.293	1.665	1.992	2.377	2.643
12	1.356	1.782	2.179	2.681	3.055	80	1.292	1.664	1.990	2.374	2.639
13	1.350	1.771	2.160	2.650	3.012	85	1.292	1.663	1.988	2.371	2.635
14	1.345	1.761	2.145	2.624	2.977	90	1.291	1.662	1.987	2.368	2.632
15	1.341	1.753	2.131	2.602	2.947	95	1.291	1.661	1.985	2.366	2.629
16	1.337	1.746	2.120	2.583	2.921	100	1.290	1.660	1.984	2.364	2.626
17	1.333	1.740	2.110	2.567	2.898	110	1.289	1.659	1.982	2.361	2.621
18	1.330	1.734	2.101	2.552	2.878	120	1.289	1.658	1.980	2.358	2.617
19	1.328	1.729	2.093	2.539	2.861	130	1.288	1.657	1.978	2.355	2.614
20	1.325	1.725	2.086	2.528	2.845	140	1.288	1.656	1.977	2.353	2.611
21	1.323	1.721	2.080	2.518	2.831	150	1.287	1.655	1.976	2.351	2.609
22	1.321	1.717	2.074	2.508	2.819	160	1.287	1.654	1.975	2.350	2.607
23	1.319	1.714	2.069	2.500	2.807	170	1.287	1.654	1.974	2.348	2.605
24	1.318	1.711	2.064	2.492	2.797	180	1.286	1.653	1.973	2.347	2.603
25	1.316	1.708	2.060	2.485	2.787	190	1.286	1.653	1.973	2.346	2.602
26	1.315	1.706	2.056	2.479	2.779	200	1.286	1.653	1.972	2.345	2.601
27	1.314	1.703	2.052	2.473	2.771						
28	1.313	1.701	2.048	2.467	2.763	∞	1.282	1.645	1.960	2.326	2.576

If the degrees of freedom is 74, we find the degrees of freedom closest to 74 listed in the table, which is 75. We then find (see **Table 10.3**)

$$t_{0.05,74} \approx t_{0.05,75} = 1.665$$

10.3b Using the computer

To calculate Student t probabilities in Excel, proceed as follows.

COMMANDS

To compute the Student t probabilities, type into any cell

=TDIST([x], [v], [Tails])

where x must be positive, v is the degrees of freedom, and Tails is 1 or 2. Typing 1 for Tails (one tail) produces the area to the right of x . Typing 2 for Tails (two tails) produces the area to the right of x plus the area to the left of $-x$.

For example, to calculate the $P[t > 2]$ and $P[|t| > 2]$ respectively, where the degrees of freedom $v = 50$,

$\text{TDIST}(2, 50, 1) = 0.025474$

and

$\text{TDIST}(2, 50, 2) = 0.050947$

To determine a value of a Student t random variable follow these instructions.

COMMANDS

To compute the t value corresponding to area A in the right tail, type into any cell

=TINV([2A],[v])

where A is the area to the right of t and v is the degrees of freedom. The result is the value of t such that the area to its right is A . The other half of the probability is located to the left of $-t$.

For example, to find the value of t_0 such that $P[t > t_0] = 0.025$, where $v = 200$, we type **=TINV(0.05, 200)**, which produces $t_0 = 1.972$. This means that $P(t > 1.972) + P(t < -1.972) = 0.025 + 0.025 = 0.05$.

In his 1908 article, Gosset showed that, when the degrees of freedom is infinitely large, t is equal to z . (As the sample size n increases, $\text{var}(t)$ approaches 1, and hence t approaches z .) That is, the Student t distribution is identical to the standard normal distribution for large values of n . As you can see, the last row in the Student t table shows values of $t_{\alpha,\text{d.f.}}$ with $\text{d.f.} = \infty$ that are equal to the z_α values we used in the previous section. For example, $t_{0.05,\infty} = z_{0.05} = 1.645$. Notice the similarity between the values of $t_{\alpha,\text{d.f.}}$ with 200 degrees of freedom and those with an infinite number of degrees of freedom. Consequently, when we have a Student t distribution with degrees of freedom greater than 200, we will approximate it by a Student t distribution with an infinite number of degrees of freedom (which is the same as the standard normal distribution). (Note: Some statistics practitioners consider $\text{d.f.} = 30$ as large enough to make the approximation $t \approx z$. However, in this book, we use $\text{d.f.} \approx 200$ to make such approximations.)

It should be noted that the **statistic $(\bar{X} - \mu)/(s / \sqrt{n})$ has the t distribution only if the sample is drawn from a normal population**. Such an application of the t distribution is said to be *robust*; this means that the t distribution also provides an adequate approximate sampling distribution of the t -statistic for moderately non-normal populations. Thus, the statistical inference techniques that follow are valid except when applied to distinctly non-normal populations.

In actual practice, some statistics practitioners ignore the preceding requirement or blindly assume that the population is normal or only somewhat non-normal. We urge you not to be one of them. Since we seldom get to know the true value of the parameter in question, our only way of knowing whether the statistical technique is valid is to be certain that the requirements underlying the technique are satisfied. At the very least, you should draw the histogram of any random variable that you are assuming is normal, to ensure that the assumption is not badly violated.

Using the same logic that produced the confidence interval estimator of the population mean μ when the variance σ^2 is known, we can develop the confidence interval estimator of the population mean when the variance σ^2 is unknown.

10.3c Estimating μ with σ^2 unknown

Using the same algebraic manipulations we employed to produce the confidence interval estimator of μ with σ^2 known, we develop the following estimator.

Confidence interval estimator of μ , with σ^2 unknown

$$\bar{X} \pm t_{\sigma/2,n-1} \frac{s}{\sqrt{n}}$$

provided that the sample comes from a normal population.

It should be noted that this estimator can only be used if the random variable X is normally distributed. In practice, this means that we can use this formula as long as X is not extremely non-normal.

When d.f. > 200, $t_{\alpha/2, \text{d.f.}}$ is approximately equal to $z_{\alpha/2}$. If σ^2 is unknown and d.f. > 200, then $t_{\alpha/2, n-1}$ can be determined from the standard normal table. Bear in mind, however, that the interval estimate is still $\bar{X} \pm t_{\alpha/2, n-1}(s/\sqrt{n})$ with $t_{\alpha/2, n-1}$ being approximated by $z_{\alpha/2}$. It should be noted that in most realistic applications where we wish to estimate a population mean, the population variance is unknown. Consequently, the interval estimator of the population mean that will be used most frequently in real life is $\bar{X} \pm t_{\alpha/2, n-1}(s/\sqrt{n})$ no matter what the sample size. However, it can be approximated by

$$\bar{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

when n is sufficiently large (i.e. $n > 200$), where $t_{\alpha/2, n-1}$ is approximated by $z_{\alpha/2}$.

Confidence interval estimators of μ

We now have two different interval estimators of the population mean. The basis for deciding which one to use is quite clear.

- 1 If the population variance σ^2 is known and the population is normally distributed or $n \geq 30$, the confidence interval estimator of the population mean μ is

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- 2 If the population variance σ^2 is unknown and the population is normally distributed, the confidence interval estimator of the population mean μ is

$$\bar{X} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$$

[Note that we can approximate $t_{\alpha/2, n-1}$ by $z_{\alpha/2}$ for large values of n (i.e. $n > 200$).]

EXAMPLE 10.2

LO4 LO6

Distance travelled by a fleet of taxis

XM10-02 As you are probably aware, a taxi fare is determined by distance travelled as well as the amount of time taken for the trip. In preparing to apply for a rate increase, the general manager of a fleet of taxis wanted to know the distance customers travel by taxi on an average trip. She organised a survey in which she asked taxi drivers to record the number of kilometres (to the nearest one-tenth) travelled by randomly selected customers. A sample of 41 customers was produced. The results appear below. The general manager wants to estimate the mean distance travelled with 95% confidence.

Distance travelled by taxi (km)

8.2	9.1	11.2	5.0	6.4	9.5	10.1	7.9	8.3	6.8	6.9	7.9	1.1	6.7	11.4	6.9
6.5	8.0	1.5	8.2	7.6	14.1	7.0	10.0	7.1	8.0	8.1	4.4	5.9	2.3	13.3	9.2
2.8	13.0	8.3	10.4	9.0	3.5	9.8	6.5	7.7							

Solution

Identifying the technique

The problem objective is to describe a single population, the distance travelled by taxi customers. The data are numerical (kilometres travelled). The parameter to be estimated is the population mean μ , the mean distance



(kilometres) travelled by all taxi customers, and the population variance σ^2 is unknown. If we assume that the distance travelled by taxi customers is normally distributed, then the confidence interval estimator for μ is

$$\bar{X} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$$

Calculating manually

From the sample of 41 observations, we find $\sum x_i = 315.6$ and $\sum x_i^2 = 2772.0$. Therefore,

$$\bar{X} = \frac{\sum x_i}{n} = \frac{315.6}{41} = 7.70 \text{ km}$$

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

$$= \frac{2772.0 - \frac{(315.6)^2}{41}}{40}$$

$$= 8.57 \text{ (km)}^2$$

$$s = \sqrt{s^2} = \sqrt{8.57} = 2.93 \text{ km}$$

Because we want a 95% confidence interval estimate,

$$1 - \alpha = 0.95$$

Thus,

$$\alpha = 0.05 \text{ and } \alpha/2 = 0.025$$

$$t_{\alpha/2, n-1} = t_{0.025, 40} = 2.021$$

The 95% confidence interval estimate of μ is

$$\bar{X} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} = 7.70 \pm 2.021 \frac{2.93}{\sqrt{41}} = 7.70 \pm 0.92$$

or

$$\text{LCL} = 6.78 \quad \text{and} \quad \text{UCL} = 8.62$$

Interpreting the results

We estimate that the mean distance travelled by taxi lies between 6.78 and 8.62 kilometres. The general manager can use the estimate to determine the effect of different pricing policies on her company. With the interval estimate she could determine upper and lower bounds on revenues generated from the new rates. She may also be able to use the results to judge the performance and honesty of individual drivers. We remind you that the accuracy of the interval estimate is dependent upon the validity of the sampling process and the distribution of the distances (they are required to be normal). If the distribution is extremely non-normal, the inference may be invalid.



Using the computer

Excel workbook

To produce the interval estimate, if you can calculate the sample mean and sample standard deviation and know the sample size, you can use the **Estimators** workbook, which is available in the Workbooks tab on the companion website. In this example, as the population standard deviation is unknown, we use the **t-Estimate_Mean** worksheet.

Excel output for Example 10.2

	A	B	C	D	E
1	t-Estimate of a Mean				
2					
3	Sample mean	27.19	Confidence Interval Estimate		
4	Sample standard deviation	2.93		7.70	± 0.92
5	Sample size	41	Lower confidence limit		6.78
6	Confidence level	0.95	Upper confidence limit		8.62

COMMANDS

- 1 Type the data into one column or open the data file (**XM10-02**). In any empty cells, calculate the sample mean (=AVERAGE(A1:A41)) and sample standard deviation (=STDEV(A1:A42)).
- 2 Open the **Estimators** workbook and click the **t-Estimate_Mean** worksheet. In cells B3 and B4, type or copy the values of the sample mean and sample standard deviation respectively. If you use **Copy**, also use **Paste Special** and **Values**. In cells B5 and B6, type the value of n (41) and the confidence level (0.95) respectively.

XLSTAT

XLSTAT output for Example 10.2

	A	B	C	D
1	95% confidence interval on the mean:			
2	[6.774, 8.621]			

(Note: This is only a partial output).

COMMANDS

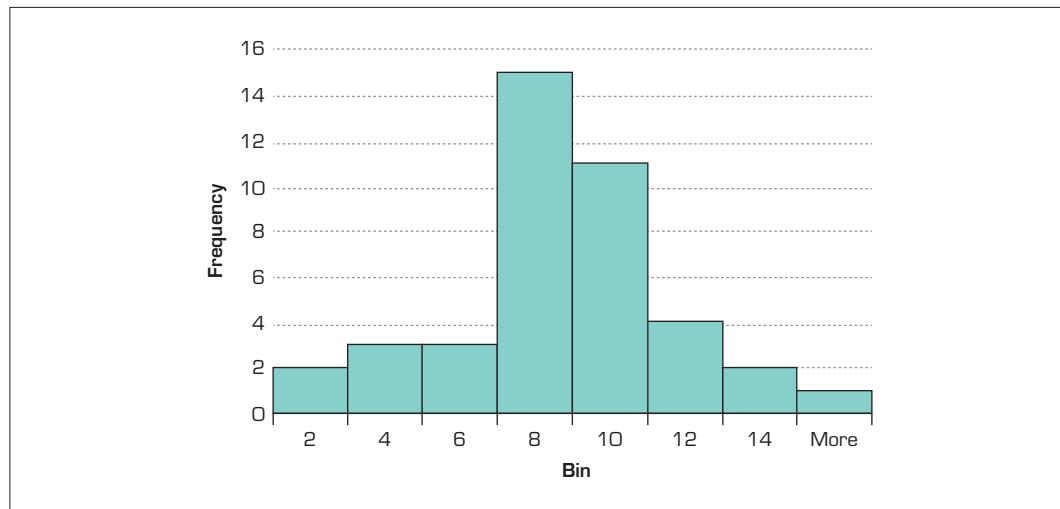
- 1 Type the data into one column or open the data file (**XM10-02**).
- 2 Click **XLSTAT** and **One-sample t-test** and **z-test**.
- 3 In the **Data:** dialog box type the input range (**A1:A100**). Click **Column labels** if the first row contains the name of the variable (as in this example). Check **Student's t-test** and do not check z-test.
- 4 Click the **Options** tab and choose **Mean ≠ Theoretical mean** in the **Alternative hypothesis:** box. Type the value of α (5, as a per cent) in the **Significance:** box. If there are blanks in the column (usually used to represent missing data) click **Missing data, Remove the observations**. For the **Variance for t-test:** check **User defined: Variance:** and type the value of s^2 (**8.5849**). Click **OK** and then **Continue**.

Checking the required conditions

When we introduced the Student t distribution, we pointed out that the t -statistic $(\bar{X} - \mu)/(s/\sqrt{n})$ is Student t distributed only if the population from which we have sampled is normal. We also noted that the techniques introduced in this section are robust, meaning that if the population is non-normal, the techniques are still valid provided that the population is not *extremely* non-normal. Although there are several statistical tests that can determine if data are non-normal, at this point we suggest drawing the histogram to see the shape of the

distribution. Excel can be used to draw the histogram for Example 10.2, which is shown in **Figure 10.9**. The histogram suggests that the variable may be normally distributed or at least not extremely non-normal.

FIGURE 10.9 Histogram for Example 10.2



10.3d Estimating the total number of successes in a large finite population

The inferential techniques introduced thus far were derived by assuming infinitely large populations. In practice, however, most populations are finite. (Infinite populations are usually the result of some endlessly repeatable process, such as flipping a coin or selecting items with replacement.) When the population is small, we must adjust the test statistic and interval estimator using the *finite population correction factor* introduced in Chapter 9 (page 364). However, we can ignore the correction factor in populations that are large relative to the sample size. Large populations are defined as populations that are at least 20 times the sample size. Finite populations allow us to use the confidence interval estimator of a mean to produce a confidence interval estimator of the population total. To estimate the total, we multiply the lower and upper confidence limits of the estimate of the mean by the population size. Thus, the confidence interval estimator of the total is:

$$N \left[\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

In Example 10.1, suppose that we wish to estimate the total number of hours Australian children watch television, given the number of children in Australia is 4 856 253. The 95% confidence interval estimate of the total of size N is:

$$N \left[\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] = 4856253 [25.623, 28.759]$$

and so

$$\text{LCL} = 124431765.5 \text{ hours and UCL} = 139660974.3 \text{ hours.}$$

10.3e Developing an understanding of the statistical concept 1

The concept developed in this section is that, to expand the application to more realistic situations, we must use another sampling distribution when the population variance is unknown. The Student t distribution was derived by W.S. Gosset for this purpose.

Another important development introduced earlier in this section is the use of the term ‘degrees of freedom’. We will encounter this term many times in this book, so a brief discussion of its meaning is warranted.

The Student t distribution is based on using the sample variance to estimate the unknown population variance. The sample variance is defined as

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}$$

To calculate s^2 , we must first determine \bar{X} . Recall that sampling distributions are derived by repeated sampling of size n from the same population. In order to repeatedly take samples to calculate s^2 , we can choose any numbers for the first $n - 1$ observations in the sample. However, we have no choice on the n th value because the sample mean must be calculated first. To illustrate, suppose that $n = 3$ and we find $\bar{x} = 10$. We can have x_1 and x_2 assume any values without restriction. However, x_3 must be such that $\bar{x} = 10$. For example, if $x_1 = 6$ and $x_2 = 8$, then x_3 must equal 16. Therefore, there are only two degrees of freedom in our selection of the sample. We say that we lose one degree of freedom because we had to calculate \bar{X} . Notice that the denominator in the calculation of s^2 is equal to the degrees of freedom. This is not a coincidence and will be repeated throughout this book.

10.3f Developing an understanding of the statistical concept 2

The t -statistic, like the z -statistic, measures the difference between the sample mean \bar{x} and the hypothesised value of μ in terms of the number of standard errors. However, when the population standard deviation σ is unknown, we estimate the standard error by s/\sqrt{n} .

10.3g Developing an understanding of the statistical concept 3

When we introduced the Student t distribution in Section 10.3a, we pointed out that it is more widely spread out than the standard normal. This circumstance is logical. The only variable in the z -statistic is the sample mean \bar{x} , which will vary from sample to sample. The t -statistic has two variables: the sample mean \bar{x} and the sample standard deviation s , both of which will vary from sample to sample. Because of the greater uncertainty, the t -statistic will display greater variability.

10.3h Missing data

In real statistical applications, we occasionally find that the data set is incomplete. In some instances the statistics practitioner may have failed to properly record some observations or some data may have been lost. In other cases, respondents may refuse to answer. For example, in political surveys in which the statistics practitioner asks voters for whom they intend to vote in the next election, some people will answer that they haven't decided or that their vote is confidential and refuse to answer. In surveys in which respondents are asked to

report their income, people often refuse to divulge this information. This is a troublesome issue for statistics practitioners. We can't force people to answer our questions. However, if the number of non-responses is high, the results of our analysis may be invalid because the sample is no longer truly random. To understand why, suppose that people who are in the top quarter of household incomes regularly refuse to answer questions about their incomes. The estimate of the population household income mean will be lower than the actual value.

The issue can be complicated. There are several ways to compensate for non-responses. The simplest method is simply to eliminate them. To illustrate, suppose that in a political survey, respondents are asked for whom they intend to vote in a two-candidate race. Surveyors record the results as 1 = candidate A, 2 = candidate B, 3 = 'Don't know', and 4 = 'Refuse to say'. If we wish to infer something about the proportion of decided voters who will vote for candidate A, we can simply omit codes 3 and 4. If we are doing the work manually, we will count the number of voters who prefer candidate A and the number who prefer candidate B. The sum of these two numbers is the total sample size.

In the language of statistical software, nonresponses that we wish to eliminate are collectively called *missing data*. Software packages deal with missing data in different ways. The Appendix to this chapter presents Excel instructions for *Missing Data*, and *Recoding Data* describes how to address the problem of missing data in Excel as well as how to recode data. In Excel, the nonresponses appear as blanks.

We complete this section with a review of how to identify this technique. To recognise when to use the *t*-confidence interval estimator of a population mean, remember the following factors.

IN SUMMARY

Factors that identify the *t*-interval estimator of a population mean

- 1 *Problem objective*: to describe a single population
- 2 *Data type*: numerical (quantitative)
- 3 *Type of descriptive measurement*: central location
- 4 *Population variance*: unknown

EXERCISES

The following exercises can be solved manually or by using the Excel **Estimators** workbook that is available from the companion website.

Learning the techniques

- 10.46 XR10-38** The following data were drawn from a normal population.

4	8	12	11	14	6	12	8	9	5
---	---	----	----	----	---	----	---	---	---

Estimate the population mean with 90% confidence.

- 10.47** You are given the following statistics:

$$\bar{X} = 156.3; \quad s = 14.9; \quad n = 12.$$

Estimate the population mean with 95% confidence.

- 10.48 a** A random sample of 8 observations was drawn from a normal population. The sample mean and sample standard deviation are $\bar{X} = 40$ and $s = 10$. Estimate the population mean with 95% confidence.

- b** Repeat part (a) assuming that you know that the population standard deviation is $\sigma = 10$.
- c** Explain why the interval estimate produced in part (b) is narrower than that in part (a).

- 10.49 a** Estimate the population mean with 90% confidence given the following statistics:

$$\bar{X} = 175; \quad s = 30; \quad n = 5.$$

- b** Repeat part (a) assuming that you know that the population standard deviation is $\sigma = 30$.
- c** Explain why the interval estimate produced in part (b) is narrower than that in part (a).

- 10.50 a** In a random sample of 500 observations drawn from a normal population, the sample mean and sample standard deviation were calculated as $\bar{X} = 350$ and $s = 100$. Estimate the population mean with 99% confidence.
- b** Repeat part (a) assuming that you know that the population standard deviation is $\sigma = 100$.
- c** Explain why the interval estimates were virtually identical.

- 10.51 XR10-51** A parking officer is conducting an analysis of the amount of time left on parking meters. A quick survey of 15 cars that have just left their metered parking spaces produced the following times (in minutes). Estimate with 95% confidence the mean amount of time left for all the vacant meters.

22	15	1	14	0	9	17	31	
18	26	23	15	33	28	20		

- 10.52 XR10-52** Part of a university lecturer's job is to publish his or her research. This task often entails reading a variety of journal articles to keep up to date. To help determine faculty standards, the dean of a business school surveyed a random sample of 12 lecturers across the country and asked them to count the number of journal articles they read in a typical month. These data are listed here. Estimate with 90% confidence the mean number of journal articles read monthly by lecturers.

9	17	4	23	56	30	
41	45	21	10	44	20	

- 10.53 XR10-53** Most owners of digital cameras store their pictures on the camera. Some will eventually download these to a computer or print them using their own printers or use a commercial printer. A film-processing company wanted to know how many pictures were stored on cameras. A random sample of 10 digital camera owners produced the data given here. Estimate with 95% confidence the mean number of pictures stored on digital cameras.

25	6	22	26	31	18	13	20	14	2
----	---	----	----	----	----	----	----	----	---

- 10.54 a** A random sample of 25 observations was drawn from a population. The sample mean and standard deviation are $\bar{X} = 510$ and $s = 125$. Estimate the population mean with 95% confidence.
- b** Repeat part (a) with $n = 50$.

- c** Repeat part (a) with $n = 100$.
- d** Describe what happens to the confidence interval estimate when the sample size increases.

- 10.55 a** The mean and standard deviation of a sample of 100 is $\bar{X} = 1500$ and $s = 300$. Estimate the population mean with 95% confidence.
- b** Repeat part (a) with $s = 200$.
- c** Repeat part (a) with $s = 100$.
- d** Discuss the effect on the confidence interval estimate of decreasing the standard deviation s .

- 10.56 a** A statistics practitioner drew a random sample of 400 observations and found that $\bar{X} = 700$ and $s = 100$. Estimate the population mean with 90% confidence.
- b** Repeat part (a) with a 95% confidence level.
- c** Repeat part (a) with a 99% confidence level.
- d** What is the effect on the confidence interval estimate of increasing the confidence level?

Applying the techniques

- 10.57 Self-correcting exercise.** A real estate company appraised the market value of 20 homes in a regional suburb in Western Australia and found that the sample mean and the sample standard deviation were \$473000 and \$46000 respectively. Estimate the mean appraisal value of all the homes in this area with 90% confidence. (Assume that the appraised values are normally distributed.)

- 10.58** A NSW Department of Consumer Affairs officer responsible for enforcing laws concerning weights and measures routinely inspects containers to determine if the contents of 10kg bags of potatoes weigh at least 10kg as advertised on the container. A random sample of 25 bags that claim that the net weight is 10kg yielded the following statistics:

$$\bar{X} = 10.52 \quad s^2 = 1.43$$

Estimate with 95% confidence the mean weight of a bag of potatoes. Assume that the weights of 10kg bags of potatoes are normally distributed.

- 10.59** A manufacturer of a brand of designer jeans realises that many retailers charge less than the suggested retail price of \$80. A random sample of 20 retailers reveals that the mean and the standard deviation of the prices of the jeans are \$64 and \$5 respectively. Estimate with 90% confidence the

mean retail price of the jeans, assuming that the prices of jeans are normally distributed.

- 10.60** An advertisement for a major washing machine manufacturer claims that its repair operators are the loneliest in the world because its washing machines require the smallest number of service calls. To examine this claim, researchers drew a random sample of 100 owners of five-year-old washing machines of that manufacturer. The mean and the standard deviation of the number of service calls in the five-year period were 4.3 and 1.8 respectively. Find the 90% confidence interval estimate for the mean number of service calls for all five-year-old washing machines produced by that manufacturer.

Computer/manual applications

The following exercises require the use of a computer and software. The answers may be calculated manually using the sample statistics based on the data provided.

- 10.61 XR10-61** The following observations were drawn from a large population.

22	18	25	28	19	20	24	26	19
26	27	22	23	25	25	18	20	26
18	26	27	24	20	19	18		

- a Estimate the population mean with 95% confidence.
- b What is the required condition of the techniques used in part (a)? Use a graphical technique to check if that required condition is satisfied.

Sample statistics: $\bar{X} = 22.6$; $s = 3.416$; $n = 25$.

- 10.62 XR10-62** A growing concern for educators in Australia is the number of teenagers who have part-time jobs while they attend high school. It is generally believed that the amount of time teenagers spend working is deducted from the amount of time devoted to school work. To investigate this problem, a school guidance counsellor took a random sample of 200 15-year-old high school students and asked how many

hours per week each worked at a part-time job. The results are recorded and some of these data are listed below.

0	6	4	7	...	9	5	0
---	---	---	---	-----	---	---	---

Estimate with 95% confidence the mean amount of time all 15-year-old high school students devote per week to part-time jobs.

Sample statistics: $\bar{X} = 5.125$; $s = 3.310$; $n = 200$.

- 10.63 XR10-63** Schools are often judged on the basis of how well their students perform academically. But what about the quality of school lunches provided? According to Health Department rules, the maximum percentage of daily kilojoules that should come from fat is 30%. To judge how well they are doing, a sample of schools was drawn. For each school the percentage of kilojoules from fat in the lunches was measured and recorded. Estimate with 95% confidence the mean percentage of kilojoules from fat in school lunches.

Sample statistics: $\bar{X} = 29.14$; $s = 4.62$; $n = 49$.

- 10.64 XR10-64** To help estimate the size of the disposable razor market, a random sample of men was asked to count the number of shaves they had with each razor. The responses are recorded. If we assume that each razor is used once per day, estimate with 95% confidence the number of days a pack of 10 razors will last.

Sample statistics: $\bar{X} = 13.94$; $s = 2.16$; $n = 212$.

- 10.65 XR10-65** Companies that sell groceries over the internet are called e-grocers. Customers enter their orders, pay by credit card and receive delivery by truck. A potential e-grocer analysed the market and determined that to be profitable the average order would have to exceed \$85. To determine whether an e-grocery would be profitable in one large city, she offered the service and recorded the size of the order for a random sample of customers. Estimate with 95% confidence the average size of an e-grocery order in this city.

Sample statistics: $\bar{X} = 89.27$; $s = 17.30$; $n = 85$.

10.4 Estimating the population proportion p

In this section we continue to address the problem of describing a single population. However, we shift our attention to populations of nominal data, which means that the population consists of nominal or categorical values. For example, in a brand preference survey in which the statistician practitioner asks consumers of a particular product which brand they purchase, the values of the random variable are the brands. If there are five brands, the values could be represented by their names, by letters (e.g. A, B, C, D and E), or by numbers (e.g. 1, 2, 3, 4 and 5). When numbers are used, it should be understood that the numbers only represent the name of the brand, are completely arbitrarily assigned, and cannot be treated as real numbers; that is, we cannot calculate means and variances.

Therefore, when the data are nominal, we count the number of occurrences of each value and calculate the proportions. Thus, the parameter of interest in describing a population of nominal data is the population proportion p . In Section 9.5, this parameter was used to calculate probabilities based on the binomial experiment. One of the characteristics of the binomial experiment is that there are only two possible outcomes per trial. Most practical applications of inference about p involve more than two outcomes. However, in most cases we are interested in only one outcome, which we label a ‘success’. All other outcomes are labelled as ‘failures’. For example, in brand-preference surveys we are interested in our company’s brand. In political surveys we wish to estimate or test the proportion of voters who will vote for one particular candidate – most likely the one who has paid for the survey.

Our task in this section is to develop the technique of estimating the population proportion. The logical statistic used in making inferences about a population proportion is the sample proportion. Thus, given a sample drawn from the population of interest, we will calculate the number of successes divided by the sample size. As we did in Chapter 7, in which we discussed the binomial distribution, we label the number of successes X ; hence, the sample proportion is X/n . As before, we denote this sample proportion by \hat{p} .

Point estimator of a population proportion

The point estimator of the population proportion p is the sample proportion,

$$\hat{p} = \frac{X}{n}$$

where X is the number of successes in the sample and n is the sample size.

In Section 9.5 we presented the approximate sampling distribution of \hat{p} . (The actual distribution is based on the binomial distribution, which does not lend itself to statistical inference.) The sampling distribution of \hat{p} is approximately normal with mean

$$\mu_{\hat{p}} = E(\hat{p}) = p$$

and standard deviation

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

where $q = 1 - p$. Hence \hat{p} is an unbiased and consistent estimator of p . Put simply, this means that \hat{p} is the best estimator of p .

Sampling distribution of the sample proportion

The sample proportion \hat{p} is approximately normally distributed, with mean p and standard deviation $\sqrt{pq/n}$, provided that n is large such that $np \geq 5$ and $nq \geq 5$.

Since \hat{p} is approximately normal, it follows that the standardised variable

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

is distributed approximately as a standard normal distribution.

10.4a Estimating a population proportion

Using the algebra employed in the previous two sections, we attempt to construct the interval estimator of p from the sampling distribution of \hat{p} . The result is

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{pq}{n}}$$

This formula is useless, however, because p and q are unknown. (If they were known, there would be no need to estimate p .) As a result, we estimate the standard deviation of \hat{p} with $\sqrt{\hat{p}\hat{q}/n}$ to produce the following formula.

Confidence interval estimator of p

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \text{ where } \hat{q} = 1 - \hat{p}$$

The use of this estimator is based on the assumption that \hat{p} is approximately normally distributed. As we explained above, this assumption requires that the sample size be sufficiently large and np and nq must be at least 5. However, as p and q are unknown, we will define n as sufficiently large to use the confidence interval estimator above if $n\hat{p} \geq 5$ and $n\hat{q} \geq 5$.

EXAMPLE 10.3

LO7

Who pays for the medical cost of smoking-related illnesses?

It is a general belief that smokers are a burden to the community, especially when they become ill. However, social workers argue that as smokers are already paying a very high consumption tax to the government when they purchase cigarettes, the government should bear the cost of medical treatment for smoking-related illnesses. A survey was conducted to find out the opinion of the general public on whether the individual smoker or society should bear the medical cost of smoking-related illnesses. In a random sample of 800 Australians, 160 were of the view that society should bear the cost of medical treatment for smoking-related illnesses. Estimate, with 99% confidence, the true proportion of people who were of the opinion that society should bear the medical cost of smoking-related illnesses.





Solution

Identifying the technique

The problem objective is to describe the opinion of the community on the issue of whether society should bear the medical cost of smoking-related illnesses. The data type is nominal, because the values of the variable are *yes* and *no*. It follows that we wish to estimate the population proportion p : the proportion of the general public who were of the opinion that society should bear the medical cost of smoking-related illnesses. The confidence interval estimator of p is

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \text{ where } \hat{q} = 1 - \hat{p}$$

provided that $n\hat{p} \geq 5$ and $n\hat{q} \geq 5$.

Calculating manually

From the data we observe that

$$\hat{p} = \frac{160}{800} = 0.20$$

and

$$\hat{q} = 1 - \hat{p} = 1 - 0.20 = 0.80$$

Therefore, $n\hat{p} = 800(0.20) = 160 > 5$ and $n\hat{q} = 800(0.80) = 640 > 5$.

The confidence level is

$$1 - \alpha = 0.99$$

so

$$z_{\alpha/2} = z_{0.005} = 2.575$$

The 99% confidence interval estimate is

$$\begin{aligned} \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} &= 0.20 \pm 2.575 \sqrt{\frac{(0.20)(0.80)}{800}} \\ &= 0.20 \pm 0.0364 \end{aligned}$$

The confidence limits are

$$\text{LCL} = 0.1636 \text{ and UCL} = 0.2364$$

The proportion of Australians who believe that society should pay the medical costs of smoking-related illnesses is therefore estimated to lie between 16.36% and 23.64%.

Using the computer

EXCEL workbook

We cannot use Data Analysis in Excel for this purpose. However, we can use the worksheet **z-Estimate_Proportion** in the **Estimators** workbook to obtain the values of LCL and UCL. The output is as follows:

Excel output for Example 10.3

	A	B	C	D	E
1	z-Estimate of a Proportion				
2					
3	Sample proportion	0.20	Confidence Interval Estimate		
4	Sample size	800		0.20	\pm 0.0364
5	Confidence level	0.99	Lower confidence limit		0.1636
6			Upper confidence limit		0.2364





COMMANDS

- 1 Open an Excel file. In any empty cell, calculate the sample proportion of 'yes' responses (**=160/800**), which is printed as **0.20**.
- 2 Open the Excel worksheet **z-Estimate_Proportion** in the **Estimators** workbook. In cells B3–B5, type the sample proportion \hat{p} (**0.20**), sample size n (**800**) and confidence level $1 - \alpha$ (**0.99**), respectively, to obtain the values of LCL and UCL.

XLSTAT

XLSTAT output for Example 10.3

	B	C	D	E	F
17	99% confidence interval on the proportion (Wald):				
18	[0.164, 0.236]				

COMMANDS

- 1 Open an Excel file. In any empty cell (B1), calculate the sample proportion of 'yes' by dividing the number of 'yes' (**160**) by the sample size (**800**), which is printed as **0.20**.
- 2 Click **XLSTAT**, **Parametric tests**, and **Tests for one proportion**.
- 3 Type the sample **Proportion:** (**0.20**), the **Sample size:** (**800**), and any **Test proportion:** (This value will not affect the confidence interval estimate.) Under **Data format:** check **Proportion**. Click **z test**.
- 4 Click the **Options** tab and choose **Proportion – Test proportion ≠ D**. Type any **Hypothesized difference (D)** (**0**). (This too will not affect the confidence interval estimate). Type the **Significance level (%)** (**1**). Under the heading **Variance** (confidence interval), click **Sample** and under **Confidence interval**, click **Wald**.

EXAMPLE 10.4

LO7

Assessing flood damage to Australian businesses

XM10-04 In late December 2010 and January 2011, Queensland suffered the worst floods in recent history. The floods affected tens of thousands of homes and businesses, and destroyed many of them. According to the National Australia Bank survey held in February 2011, the fallout was wide-ranging. It caused disruption not only to Queensland businesses but also to business nationally. The survey, which was conducted for 1500 medium or large businesses with more than 50 employees nationwide, found that 10% of businesses had experienced some disruption or had to close. [This proportion was 25% for Queensland businesses.] The data with 1 (affected) and 0 (not affected) were recorded. Estimate with 95% confidence the proportion of all nationwide businesses that experienced disruption or closure due to the floods.

Solution

Identifying the technique

The problem objective is to describe the population of Australian businesses affected by the 2010–11 Queensland floods. Each business is categorised as being either affected or unaffected. We recognise that the data are nominal. To help you differentiate between nominal and numerical data, suppose that the report had analysed the *cost* of repairing the damage. In that case, for each business, analysts would have recorded the cost of repairs, and thus the data would have been numerical.

The parameter of interest is the proportion p of Australian businesses affected by the 2010–11 Queensland floods. The interval estimator is

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

provided that $n\hat{p} \geq 5$ and $n\hat{q} \geq 5$.



Calculating manually

The sample size $n = 1500$, and the sample proportion $\hat{p} = 0.10$ (and $\hat{q} = 1 - \hat{p} = 0.90$). Therefore, $n\hat{p} = 1500(0.10) = 150 > 5$ and $n\hat{q} = 1350 > 5$. We set the confidence level at 95%, so $\alpha = 0.05$, $\alpha/2 = 0.025$ and $z_{\alpha/2} = z_{0.025} = 1.96$. The 95% confidence interval estimate of p is

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.10 \pm 1.96 \sqrt{\frac{(0.10)(0.90)}{1500}} = 0.10 \pm 0.02$$

The lower and upper confidence limits are

$$\text{LCL} = 0.08 \text{ and UCL} = 0.12$$

Interpreting the results

We estimate that between 8% and 12% of all businesses in Australia were affected after the 2010–11 Queensland floods. We can use this estimate in a variety of ways. First, the estimate provides us with a measure of the magnitude of the effect on the Australian business sector. For example, we can say that at least 8% of the businesses were closed or disrupted. Second, we can use the interval estimate to estimate the cost of repairs.

Using the computer

EXCEL Workbook

Excel is not designed to calculate interval estimates for population proportions.

The number of successes from raw data can be obtained using the COUNTIF function in Excel, from which you can calculate \hat{p} and then produce LCL and UCL using the **z-Proportion** worksheet in the **Estimators.xlsx** workbook (which can also be employed for ‘what-if’ analyses).

Excel output for Example 10.4

	A	B	C	D	E
1	z-Estimate of a Proportion				
2					
3	Sample proportion	0.10	Confidence Interval Estimate		
4	Sample size	1500		0.10	± 0.0152
5	Confidence level	0.95		Lower confidence limit	0.0848
6				Upper confidence limit	0.1152

COMMANDS

- 1 Open the data file (**XM10-04**). In any empty cell (C2), calculate the number of ‘successes’ ($=\text{COUNTIF(A2:A1501,1)}$). In another cell (C3), divide that number by the sample size (**1500**) to obtain the sample proportion ($=\text{C2}/1500$). This gives the sample proportion 0.1.
- 2 Open the Excel worksheet **z-Estimate_Proportion** in the **Estimators** workbook. In cells B3–B5, type the sample proportion \hat{p} (**0.10**), sample size n (**1500**) and confidence level $1 - \alpha$ (**0.95**) respectively. The upper and lower confidence limits will be provided as output.

USING XLSTAT

Alternatively, you can use **XLSTAT** to print the lower and upper confidence limits of the interval estimate of p .

XLSTAT output for Example 10.4

	B	C	D	E	F
17	95% confidence interval on the proportion (Wald):				
18	[0.085, 0.115]				



COMMANDS

- 1 Open the data file (**XM10-04**). In any empty cell (C2), calculate the number of 'successes' (**=COUNTIF A2:A1501,0**). In another cell (C3), divide that number by the sample size (**1500**) to obtain the sample proportion (**=C2/1500**). This gives the sample proportion 0.1.
- 2 Click **XLSTAT**, **Parametric tests**, and **Tests** for one proportion.
- 3 Type the sample Proportion: **(0.1)**, the Sample size: **(1500)**, and any Test proportion: (This value will not affect the confidence interval estimate.) Under Data format: check **Proportion**. Click **z test**.
- 4 Click the **Options** tab and choose **Proportion – Test proportion ≠ D**. Type any **Hypothesized difference (D)**. (This too will not affect the confidence interval estimate). Type the Significance level (%) **(5)**. Under the heading **Variance (confidence interval)**, click **Sample** and under **Confidence** interval, click **Wald**.

10.4b Estimating the total number of successes in a large finite population

As was the case with the inference about a mean, the techniques in this section assume infinitely large populations. When the populations are small, it is necessary to include the finite population correction factor. In our definition, a population is small when it is less than 20 times the sample size. When the population is large and finite, we can estimate the total number of successes in the population.

To produce the confidence interval estimator of the total, we multiply the lower and upper confidence limits of the interval estimator of the proportion of successes by the population size. The confidence interval estimator of the total number of successes in a large finite population of size N is

$$N \left(\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \right)$$

In Example 10.3, suppose that we wish to estimate the number of people who were of the opinion that society should bear the medical cost of smoking-related illnesses, given the number of adults in Australia is $N = 20.55$ million. The 99% confidence interval estimate of the number of people who were of the opinion that society should bear the medical cost is

$$N \left(\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \right) = 20.55 [0.1636, 0.2364] = [3.362, 4.858]$$

and so

$$\text{LCL} = 3.362 \text{ million and UCL} = 4.858 \text{ million}$$

We complete this section by reviewing the factors that tell us when to estimate a population proportion.

IN SUMMARY

Factors that identify the interval estimator of p

- 1 *Problem objective:* to describe a single population
- 2 *Data type:* nominal (categorical)

EXERCISES

Learning the techniques

These exercises can be solved manually or by using Excel's **Estimators.xlsx** workbook, which can be obtained from the companion website.

- 10.66** Given that the proportion of successes $\hat{p} = 0.84$ and $n = 600$, estimate p with 90% confidence.
- 10.67** In a random sample of 250, we found 75 successes. Estimate the population proportion of successes with 99% confidence.
- 10.68** Estimate p with 95% confidence, given $X = 27$ and $n = 100$.
- 10.69** Estimate p with 95% confidence, given that a random sample of 100 produced $\hat{p} = 0.2$.
- 10.70** **a** In a random sample of 200 observations, we found the proportion of successes to be $\hat{p} = 0.48$. Estimate with 95% confidence the population proportion of successes.
b Repeat part (a) with $n = 500$.
c Repeat part (a) with $n = 1000$.
d Describe the effect on the confidence interval estimate of increasing the sample size.
- 10.71** **a** The proportion of successes in a random sample of 400 was calculated as $\hat{p} = 0.50$. Estimate the population proportion with 95% confidence.
b Repeat part (a) with $\hat{p} = 0.33$.
c Repeat part (a) with $\hat{p} = 0.10$.
d Discuss the effect on the width of the confidence interval estimate of reducing the sample proportion.

Applying the techniques

- 10.72** **Self-correcting exercise.** In a random sample of 1000 picture tubes produced in a large plant, 80 were found to be defective. Estimate with 95% confidence the true proportion of defective picture tubes produced at this plant.
- 10.73** In a survey of 250 voters prior to an election, 40% indicated that they would vote for the incumbent candidate. Estimate with 90% confidence the population proportion of voters who support the incumbent.
- 10.74** Surveyors asked a random sample of women in a major city what factor was the most important

in deciding where to shop. The results appear in the following table. If the sample size was 1200, estimate with 95% confidence the proportion of women who identified price and value as the most important factor.

Factor	Percentage (%)
Price and value	40
Quality and selection of merchandise	30
Service	15
Shopping environment	15

- 10.75** In a Household Economic Survey about life satisfaction among 5849 New Zealand households, the following results were reported. Estimate with 95% confidence the proportion of all New Zealand households that are satisfied or very satisfied with their lives.

Source	Percentage (%)
Very satisfied	26.0
Satisfied	50.5
Neither satisfied nor dissatisfied	14.9
Dissatisfied	6.5
Very dissatisfied	2.1

Source: Statistics New Zealand. CC Attribution 4.0 International (CC BY 4.0) <https://creativecommons.org/licenses/by/4.0/>.

- 10.76** What type of educational background do CEOs have? In one survey, 344 CEOs of medium and large companies were asked whether they had an MBA degree. There were 97 MBAs. Estimate with 95% confidence the proportion of all CEOs of medium and large companies who have MBAs.

- 10.77** The dean of a business school wanted to know whether the graduates of her school had used a statistical inference technique during their first year of employment after graduation. She surveyed 314 graduates and asked about the use of statistical techniques. After tallying up the responses, she found that 204 had used statistical inference within one year of graduation. Estimate with 90% confidence the proportion of all business school graduates who use their statistical education within a year of graduation.

- 10.78** The GO transportation system of buses and commuter trains operates on the honour system. Train travellers are expected to buy their tickets before boarding the train. Only a small number of people will be checked on the train to see whether they bought a ticket. Suppose that a random sample of 400 train travellers was sampled and 68 of them had failed to buy a ticket. Estimate with 95% confidence the proportion of all train travellers who do not buy a ticket.
- 10.79** Refer to Exercise 10.78. Assuming that there are 1 million travellers per year and the fare is \$8.00, estimate with 95% confidence the amount of revenue lost each year.

Computer/manual applications

- 10.80 XR10-80** An insurance company suspects that many homes are being destroyed during tropical cyclones due to poor building-code rules and enforcement. An investigator for the company decided to analyse the damage done in a

cyclone-affected area to homes built before 1980 and to those built after 1980. One of the findings was that in areas where the sustained winds were more than 200 km/h, 33% of the houses built after 1980 were uninhabitable. Suppose that, after examining a sample of 300 homes, a statistics practitioner recorded whether the house was uninhabitable (1) or habitable (0). With 90% confidence, estimate the proportion of all homes exposed to winds of more than 200 km/h that were uninhabitable after this cyclone.

Sample frequencies: $n(0) = 158$; $n(1) = 142$.

- 10.81 XR10-81** An increasing number of people are giving gift cards as Christmas presents. To measure the extent of this practice, a random sample of people was asked (survey conducted December 26–29) whether they had received a gift card for Christmas. The responses are recorded as 1 = No and 2 = Yes. Estimate with 95% confidence the proportion of people who received a gift card for Christmas.

Sample frequencies: $n(1) = 92$; $n(2) = 28$.

10.5 Determining the required sample size

As you have seen in the previous four sections, interval estimates can often provide useful information about the value of a parameter. If the interval is too wide, however, its use is quite limited.

In Example 10.1, the interval estimate of the average time spent by children watching television per week was 25.6 hours to 28.8 hours. If the program manager is to use this estimate as input for his advertising plan, he needs greater precision. Fortunately, statistics practitioners can control the width of the interval by determining the sample size necessary to produce narrow intervals.

Suppose that before gathering the data, the program manager had decided that he needed to estimate the mean to within 0.5 hours of the true value. The phrase ‘to within 0.5 hours’ means that the interval estimate is to be of the form

$$\bar{X} \pm 0.5$$

That is, the program manager has specified the number of hours following the plus/minus sign to be 0.5.

The formula for the confidence interval estimate of μ is

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

It follows therefore that

$$z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 0.5$$

After some algebraic manipulation the equation becomes

$$n = \left(\frac{z_{\alpha/2} \sigma}{0.5} \right)^2$$

We have specified the confidence level to be 95%, thus $z_{\alpha/2} = 1.96$. The value of σ is 8. Thus,

$$n = \left(\frac{(1.96)(8)}{0.5} \right)^2 = 983.45 = 984 \text{ (rounded up)}$$

Note that when calculating the sample size, the value is always rounded up, as any increase in sample size would result in narrower intervals.

To produce the 95% confidence interval estimate of the mean $\bar{X} \pm 0.5$, we need to sample 984 children. Notice that all that is left to be done is to collect the data on the number of hours the 984 randomly selected children spend watching television per week and calculate the sample mean. If the sample mean is (say) 26, the interval estimate becomes 26 ± 0.5 .

10.5a Error of estimation

In Chapter 5, we pointed out that sampling error is the difference between the sample and the population that exists only because of the observations that happened to be selected for the sample. Now that we have discussed estimation, we can define the sampling error as the difference between an estimator and a parameter. We can also define this difference as the error of estimation. In this chapter, this can be expressed as the difference between \bar{X} and μ .

The **error of estimation** is the absolute difference between the point estimator and the parameter. For example, the point estimator of μ is \bar{X} , so in that case

$$\text{Error of estimation} = |\bar{X} - \mu|$$

The maximum error of estimation is called the error bound and is denoted B . In the above example, $B = 0.5$.

error of estimation

The absolute difference between the statistic and the parameter.

10.5b Determining the sample size to estimate a mean

Now we derive a general formula for the **sample size required to estimate the population mean μ** . Let B represent the sampling error we are willing to tolerate, which is the quantity following the \pm sign. When σ is known:

$$B = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

With the same algebraic manipulation, we derive the formula for the sample size:

$$n = \left(\frac{z_{\alpha/2}\sigma}{B} \right)^2$$

sample size required to estimate the population mean μ

The sample size required to estimate μ given the confidence level of the interval estimator and the size of the error bound.

Sample size required to estimate μ with σ^2 known

$$n = \left(\frac{z_{\alpha/2}\sigma}{B} \right)^2 \quad \text{or} \quad \frac{z_{\alpha/2}^2\sigma^2}{B^2}$$

In order to find n , we had to have some value for α and σ . The value of σ could be approximated using estimates from previous experiments or from knowledge about the population.

A popular method of approximating σ for normal or near-normal populations is to begin by using the range of the random variable. In practice, this task is relatively simple. A conservative estimate of σ is the range divided by 4; that is, $\sigma \approx \text{range}/4$. This method is quite effective because approximating the range is often easy. It should be noted, however, that the estimate is valid only for normal or near-normal populations.

EXAMPLE 10.5

LO5

Determining the sample size to estimate the mean tree diameter

A timber company has just acquired the rights to a large tract of land containing thousands of trees. The company manager needs to be able to estimate the amount of timber that can be harvested in a tract of land to determine whether the effort will be profitable. To do so, the manager needs an estimate of the mean diameter of the trees. He decides to estimate that parameter to within 2.5 cm with 90% confidence. A forester familiar with the territory guesses that the diameters of the trees are normally distributed with a standard deviation of 15 cm. Determine the required sample size (number of trees).

Solution**Identifying the technique**

The problem objective is to describe the population mean of a single population. The data type is numerical. We want to estimate the required sample size to achieve a desired error bound.

Calculating manually

Before the sample is taken, the forester needs to determine the sample size using the formula on page 411, as follows:

The maximum error of estimation or error bound $B = 2.5$ cm.

The confidence level is 90%. That is, $1 - \alpha = 0.90$.

Thus, $\alpha = 0.10$ and $\alpha/2 = 0.05$. Therefore $z_{\alpha/2} = z_{0.05} = 1.645$.

The diameter of trees population is normally distributed with standard deviation $\sigma = 15$ cm.

Thus, the required sample size can be calculated as

$$n = \left(\frac{z_{\alpha/2}\sigma}{B} \right)^2 = \left(\frac{1.645 \times 15}{2.5} \right)^2 = 97.42$$

As the sample size has to be a whole number, we round up to 98.

Using the computer**Using Excel workbook**

To produce the required sample size, if you have the population standard deviation (or approximate) and know the error bound B , you can use the **Estimators** workbook. In this example, as we need the required sample size to estimate the mean, we use the **Sample-size_Mean** worksheet.

Excel output for Example 10.5

	A	B
1	Confidence Intervals (Population mean):	
2	Calculating the required sample size	
3		
4	Sigma (known)=	15
5	Confidence level (1-alpha)=	0.90
6	B=	2.5
7	Alpha=	0.10
8	n=	98

COMMANDS

Open the **Estimators** workbook and click the **Sample-size_Mean** worksheet. In cells B4–B6, type the population standard deviation (15), confidence level (0.90) and error bound B (2.5) respectively. The value of n (98, rounded up) will be displayed in cell B8.

In this chapter, we have assumed that we know the value of the population standard deviation. In practice, this is seldom the case. (The confidence interval estimator of the population mean introduced in Section 10.3 when σ is unknown is more realistic.) It is frequently necessary to ‘guesstimate’ the value of σ to calculate the sample size; that is, we must use our knowledge of the variable with which we are dealing to assign some value to σ .

Unfortunately, we cannot be very precise in this guess. However, in guesstimating the value of σ , we prefer to err on the high side. For Example 10.5, if the forester had determined the sample size using $\sigma = 30$, he would have calculated

$$n = \left(\frac{z_{\alpha/2}\sigma}{B} \right)^2 = \left(\frac{1.645 \times 30}{2.5} \right)^2 = 389.67 \text{ (rounded to 390)}$$

Using $n = 390$ (assuming that the selected sample of size $n = 390$ gives a sample mean value of 62.5), the 90% confidence interval estimate is

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 62.5 \pm 1.645 \times \frac{30}{\sqrt{390}} = 62.5 \pm 2.5$$

This interval is as narrow as the forester wanted ($B = 2.5$).

What happens if the standard deviation is *smaller* than assumed? If we discover that the standard deviation is less than that we assumed when we determined the sample size, the confidence interval estimator will be narrower and therefore more precise. Suppose that after the sample of 98 trees was taken (assuming again that $\sigma = 15$), the forester discovers that $\sigma = 7.5$. The confidence interval estimate is

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 62.5 \pm 1.645 \times \frac{7.5}{\sqrt{98}} = 62.5 \pm 1.25$$

which is narrower than that wanted by the forester ($B = 2.5$). Although this means that he would have sampled more trees than needed, the additional cost is relatively low when compared to the value of the information derived.

10.5c Determining the required sample size for estimating a population proportion

To find the required sample size for estimating p , we solve the equation

$$B = z_{\alpha/2} \sqrt{\frac{pq}{n}}$$

to obtain the following formula:

$$n = \left(\frac{z_{\alpha/2} \sqrt{pq}}{B} \right)^2 \quad \text{where } q = 1 - p$$

Sample size required to estimate p

$$n = \left(\frac{z_{\alpha/2} \sqrt{pq}}{B} \right)^2 \quad \text{or} \quad \frac{z_{\alpha/2}^2 pq}{B^2} \quad \text{where } q = 1 - p$$

Suppose that, working with the data in Example 10.4 discussed above, we want to estimate the proportion of all nationwide businesses that experienced disruption or closure due to the floods to within 0.03, with 95% confidence. This means that, when the sample is taken, we wish the interval estimate to be $\hat{p} \pm 0.03$. Hence, $B = 0.03$. Since $1 - \alpha = 0.95$, we know that $z_{\alpha/2} = 1.96$; therefore,

$$n = \left(\frac{1.96 \sqrt{pq}}{0.03} \right)^2$$

To solve for n we need to know p and q . Unfortunately, we cannot replace them with the sample values \hat{p} and \hat{q} as these values are unknown – the sample has not yet been taken. At this point, we can use one of two methods to solve for n .

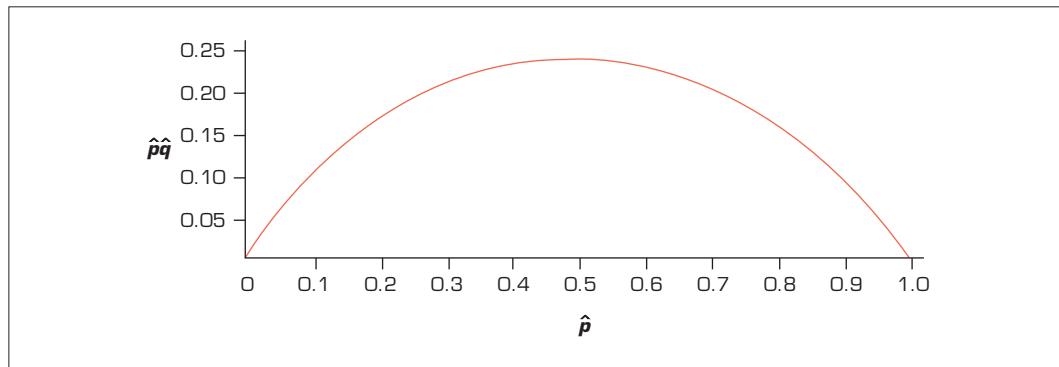
Method 1

If we have no knowledge even of the approximate values of \hat{p} and \hat{q} , we let $\hat{p} = \hat{q} = 0.5$. We choose $\hat{p} = 0.5$ (and thus, $\hat{q} = 0.5$) because the product $\hat{p}\hat{q}$ equals its maximum value for this value of \hat{p} . (Figure 10.10 illustrates this point.)

This, in turn, results in a conservative value of n , and, as a result, the confidence interval will be no wider than the interval $\hat{p} \pm 0.03$. If, when the sample is drawn, \hat{p} does not equal 0.5, the interval estimate will be better (that is, narrower) than planned. Thus,

$$n = \left(\frac{1.96\sqrt{(0.5)(0.5)}}{0.03} \right)^2 = (32.67)^2 = 1068$$

FIGURE 10.10 Graphing $\hat{p}\hat{q}$ versus \hat{p}



If it turns out that $\hat{p} = 0.5$, the interval estimate is 0.5 ± 0.03 . If not, the interval estimate will be narrower. For instance, if it turns out that $\hat{p} = 0.2$, the estimate is 0.2 ± 0.024 , which is better than we had planned.

Incidentally, this is the sample size used in opinion polls of the type frequently referred to in newspapers, magazines and television. These polls usually estimate proportions to within 3%, with 95% confidence. (The media often state the 95% confidence level as ‘19 times out of 20’.)

Method 2

If we have some idea about the value of \hat{p} , we can use it to determine n . For example, if we believe that \hat{p} will turn out to be approximately 0.2, we can solve for n as follows:

$$n = \left(\frac{1.96\sqrt{(0.2)(0.8)}}{0.03} \right)^2 = (26.13)^2 = 683$$

Notice that this produces a smaller value of n (thus reducing sampling costs) than does Method 1. If \hat{p} actually lies between 0.2 and 0.8, however, the estimate will not be as good as we wanted, because the interval will be wider than desired.

If you have ever wondered why opinion polls almost always estimate proportions to within 3%, consider the sample size required to estimate a proportion to within 1%.

$$n = \left(\frac{1.96\sqrt{(0.5)(0.5)}}{0.01} \right)^2 = (98)^2 = 9605$$

The sample size 9605 is 9 times the sample size needed to estimate a proportion to within 3%. Thus, to divide the width of the interval by 3 requires multiplying the sample size by 9. The cost would also increase considerably. For most applications, the increase in accuracy

(created by decreasing the width of the interval estimate from 0.03 to 0.01) does not overcome the increased cost. Interval estimates with 5% or 10% bounds (sample sizes 385 and 97 respectively) are generally considered too wide to be useful. Thus, the 3% bound is the happy compromise between cost and accuracy.

EXAMPLE 10.6

LOG

Proportion of shoppers who will buy a new liquid detergent

A market analyst in Sydney wants to estimate the proportion of shoppers who will buy a new type of liquid detergent. How large a sample should he take in order to estimate that proportion to within 0.04, with 90% confidence?

Solution

Identifying the technique

The problem objective is to describe a single population, the proportion of shoppers who will buy the new type of liquid detergent, p . The data are nominal.

Calculating manually

The parameter to be estimated is the population proportion p , with a confidence level of $1 - \alpha = 0.90$.

$$z_{\alpha/2} = z_{0.05} = 1.645$$

The error bound is

$$B = 0.04$$

Unfortunately, as we haven't taken the sample yet, we do not know the values of \hat{p} and \hat{q} . The best way of proceeding is to select the values of \hat{p} and \hat{q} that produce the largest possible value of n (as discussed in Method 1). Thus, we set $\hat{p} = 0.5$ and $\hat{q} = 0.5$, and hence

$$n = \frac{z_{\alpha/2}^2 (0.5)(0.5)}{B^2} = \frac{(1.645)^2 (0.25)}{(0.04)^2} = 423$$

A sample of 423 shoppers should be taken to estimate the proportion of shoppers who will buy the new detergent to within 0.04, with 90% confidence.

Using the computer

Using Excel workbook

To produce the required sample size, if you have an estimate for the population proportion (or use 0.5) and know the error bound B , you can use the **Estimators** workbook. In this example, as we need the required sample size to estimate the mean, we use the **Sample-size_Proportion** worksheet.

Excel output for Example 10.6

	A	B
1	Confidence Intervals (Sample proportion):	
2	Calculating the sample size for a fixed width (2B)	
3	phat =	0.50
4	Confidence level (1-alpha)=	0.90
5	Width/2 = B =	0.04
6	alpha =	0.10
7	Required sample size n=	423

COMMANDS

Open the **Estimators** workbook and click the **Sample-size_Proportion** worksheet. In cell B3, type the estimated \hat{p} or use 0.5. In cells B4 and B5, type the confidence level (**0.90**) and error bound B (**0.04**), respectively. The value of the required sample size n (**423**, rounded up) will be displayed in cell B7.

EXERCISES

Learning the techniques

- 10.82** Determine the sample size required to estimate μ to within 10 units with 99% confidence. We know that the range of the population observations is 200 units.
- 10.83** Find the required sample size, n , given that we wish to estimate μ to within 10 units with 95% confidence, and assuming that $\sigma = 100$.
- 10.84** **a** Determine the sample size required to estimate a population mean to within 10 units given that the population standard deviation is 50. A confidence level of 90% is judged to be appropriate.
b Repeat part (a) changing the standard deviation to 100.
c Repeat part (a) using a 95% confidence level.
d Repeat part (a) given that we now wish to estimate the population mean to within 20 units.
- 10.85** Review Exercise 10.84. Describe what happens to the sample size when:
a the population standard deviation increases.
b the confidence level increases.
c the bound on the error of estimation increases.
- 10.86** **a** A statistics practitioner would like to estimate a population mean to within 50 units with 99% confidence, given that the population standard deviation is 250. What sample size should be used?
b Repeat part (a) changing the standard deviation to 50.
c Repeat part (a) using a 95% confidence level.
d Repeat part (a) given that we now wish to estimate the population mean to within 10 units.
- 10.87** Review Exercise 10.86. Describe what happens to the sample size when:
a the population standard deviation decreases.
b the confidence level decreases.
c the bound on the error of estimation decreases.
- 10.88** **a** Determine the sample size necessary to estimate a population mean to within 1 with 90% confidence given that the population standard deviation is 10.
b Suppose that the sample mean was calculated as 150. Estimate the population mean with 90% confidence.
- 10.89** **a** Repeat part (b) in Exercise 10.88 after discovering that the population standard deviation is actually 5.

- b** Repeat part (b) in Exercise 10.88 after discovering that the population standard deviation is actually 20.

- 10.90** How large a sample should be taken to estimate a population proportion to within 0.05 with 95% confidence?

Applying the techniques

- 10.91** **Self-correcting exercise.** One of the tasks of the operations managers of production plants is to estimate the average time it takes to assemble a product. The operations manager of a large production plant would like to estimate the average time a worker takes to assemble a new electronic component. After observing a number of workers assembling similar devices, she noted that the shortest time taken was 10 minutes, and the longest time taken was 22 minutes. Determine how large a sample of workers should be taken if the manager wishes to estimate the mean assembly time to within 20 seconds with 99% confidence level.
- 10.92** A medical researcher wants to investigate the amount of time it takes for headaches to be relieved after patients take a new prescription painkiller. She plans to use statistical methods to estimate the mean of the population of relief times. She believes that the population is normally distributed with a standard deviation of 20 minutes. How large a sample should she take to estimate the mean time to within 1 minute with 90% confidence?
- 10.93** The operations manager of a plant making mobile telephones has proposed rearranging the production process to be more efficient. He wants to estimate the time to assemble the telephone using the new arrangement. He believes that the population standard deviation is 15 seconds. How large a sample of workers should he take to estimate the mean assembly time to within 2 seconds with 95% confidence?
- 10.94** A medical statistician wants to estimate the average weight loss of people who are on a new diet plan. In a preliminary study, she found that the smallest weight loss was 3 kg and the largest weight loss was 39 kg. How large a sample should be drawn to estimate the mean weight loss to within 2 kg with 90% confidence?

- 10.95** A marketing manager is in the process of deciding whether to introduce a new product. He has concluded that he needs to perform a market survey in which he asks a random sample of

people whether they will buy the product. How large a sample should he take if he wants to estimate the proportion of people who will buy the product to within 3% with 99% confidence?

10.6 Applications in marketing: Market segmentation

Market segmentation separates consumers of a product into different groups in such a way that members of each group are similar to each other and there are differences between groups. Market segmentation grew out of the realisation that a single product can seldom satisfy the needs and wants of all consumers. For example, the market for new cars must be segmented because there is a wide variety of needs that a car must satisfy. There are often identifiable segments of the market to which specifically designed products can be directed.

There are many ways to segment a market. **Table 10.4** lists several different segmentation variables and their market segments. For example, car manufacturers can use education levels to segment the market. It is likely that high school graduates would be quite similar to others in this group and that members of this group would differ from university graduates. We would expect those differences to include the types and brands of cars each group would choose to buy. However, it is likely that income level would differentiate more clearly between segments. Statistical techniques can be used to help determine the best way to segment the market. These statistical techniques are more advanced than can be studied in this textbook. Consequently, we will focus our attention on other statistical applications.

In Example 10.7 below, we demonstrate how statistical methods can be used to determine the size of a segment, which is used to determine its profitability. This aspect is crucial because not all segments are worth pursuing. In some instances the size of the segment is too small or the costs of satisfying a segment may be too high. The size of a segment can be determined in several ways.

The census can provide useful information. For example, we can determine the number of Australians in various age categories. For other segments we may need to survey members of a population and use the inferential techniques introduced in this chapter.

TABLE 10.4 Market segmentation

Segmentation variable	Segments
Geographic	
Countries	Australia, Canada, China, Japan, Korea, Taiwan, United States
Country regions	Metropolitan, Urban, Rural
Demographic	
Age	Under 5, 5–12, 13–19, 20–29, 30–50, over 50
Education	Some high school, high school graduate, some TAFE college, TAFE college or university graduate
Income	Under \$20 000, \$20 000–\$29 999, \$30 000–\$50 000, over \$50 000
Marital status	Single, married, divorced, widowed
Social	
Religion	Catholic, Protestant, Jewish, Islam, Buddhist, Hindu
Class	Upper class, middle class, working class, lower class
Behaviour	
Media usage	TV, internet, newspaper, magazine
Payment method	Cash, cheque, Visa, MasterCard

EXAMPLE 10.7

LO7

Segmentation of the credit card market

XM10-07 A new credit card company is investigating various market segments to determine whether it is profitable to direct its advertising specifically at each one. One of the market segments is composed of Asian migrants. The latest census indicates that there are 2.01 million Asian migrants in Australia. A survey of 475 Asians asked each how they usually pay for the products they purchase. The responses are:

- 1 Cash
- 2 Cheque
- 3 Visa
- 4 MasterCard
- 5 Other credit card.

The responses are recorded. Estimate with 95% confidence the number of Asian migrants in Australia who usually pay by credit card.

Solution**Identifying the technique**

The problem objective is to describe the method of payment of the Asian migrants in Australia. The data are nominal. Consequently, the parameter we wish to estimate is the proportion p of Asian migrants in Australia who classify themselves as users of credit cards. The confidence interval estimator we need to employ is

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Calculating manually

To solve manually, we count the number of 3s (Visa card), 4s (MasterCard) and 5s (Other credit cards) in the file. We find them to be 167, 146 and 34 respectively. Thus,

$$\begin{aligned}\hat{p} &= \frac{(167 + 146 + 34)}{475} = 0.7305 \\ \hat{q} &= 1 - \hat{p} = 1 - 0.7305 = 0.2695\end{aligned}$$

and

$$\begin{aligned}n\hat{p} &= 475(0.7305) = 347 > 5 \\ n\hat{q} &= 475(0.2695) = 128 > 5\end{aligned}$$

The confidence level is $1 - \alpha = 0.95$. It follows that $\alpha = 0.05$, $\alpha/2 = 0.025$ and $z_{\alpha/2} = z_{0.025} = 1.96$. The 95% confidence interval estimate of p is

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.7305 \pm 1.96 \sqrt{\frac{(0.7305)(0.2695)}{475}} = 0.7305 \pm 0.0399$$

and so

$$\text{LCL} = 0.6906 \text{ and UCL} = 0.7704$$

To estimate the number of Asian migrants in Australia who usually pay by credit card, we multiply the proportion by the number of Asian migrants in Australia.

Interpreting the results

We estimate that the proportion of Asian migrants in Australia who usually pay by credit card lies between 0.6906 and 0.7704. Because there are 2.01 million Asian migrants in the Australian population, we estimate that the number of Asian migrants in Australia who usually pay by credit card falls between



LCL = $2.01(0.6906) = 1.388106$ million

and

UCL = $2.01(0.7704) = 1.548504$ million

Using the computer

The Excel commands are the same as in Example 10.4.

Excel output for Example 10.7

	A	B	C	D	E
1	z-Estimate of a Proportion				
2					
3	Sample proportion	0.7305	Confidence Interval Estimate		
4	Sample size	475		0.7305	\pm 0.0399
5	Confidence level	0.95		Lower confidence limit	0.6906
6				Upper confidence limit	0.7704

Now we will answer the opening problem described in this chapter's introduction.

SPOTLIGHT ON STATISTICS

Segmentation of the breakfast cereal market: Solution

Identifying the technique

The problem objective is to describe the healthy eating habits of the population of Australian adults. The data are nominal. Consequently, the parameter we wish to estimate is the proportion, p , of Australian adults who classify themselves as concerned about healthy eating. The confidence interval estimator we need to employ is

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$



Calculating manually

To solve manually, we count the number of 1s in the file. We find this value to be 269. Thus,

$$\begin{aligned}\hat{p} &= \frac{X}{n} = \frac{269}{1250} = 0.2152 \\ \hat{q} &= 1 - \hat{p} = 1 - 0.2152 = 0.7848\end{aligned}$$

and

$$\begin{aligned}n\hat{p} &= 1250(0.2152) = 269 > 5 \\ n\hat{q} &= 1250(0.7848) = 981 > 5\end{aligned}$$

The confidence level is $1 - \alpha = 0.95$. It follows that $\alpha = 0.05$, $\alpha/2 = 0.025$ and $z_{\alpha/2} = z_{0.025} = 1.96$.

The 95% confidence interval estimate of p is

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.2152 \pm 1.96 \sqrt{\frac{(0.2152)(0.7848)}{1250}} = 0.2152 \pm 0.0228$$

and

LCL = 0.1924 and UCL = 0.2380

Interpreting the results

We estimate that the proportion of Australian adults who are in group 1 lies between 0.1924 and 0.2380. Because there are 18798064 adults in the population, we estimate that the number of adults who belong to group 1 falls between

$$\text{LCL} = 18798064(0.1924) = 3616748$$

and

$$\text{UCL} = 18798064(0.2380) = 4473939$$

Using the computer

The Excel commands are the same as in Example 10.4.

Excel output for Opening example

	A	B	C	D	E
1	z-Estimate of a Proportion				
2					
3	Sample proportion	0.2152	Confidence Interval Estimate		
4	Sample size	1250	0.2152	± 0.0228	
5	Confidence level	0.95	Lower confidence limit	0.1924	
6			Upper confidence limit	0.2380	

We will return to the subject of market segmentation in later chapters in which we demonstrate how statistics can be employed to determine whether differences actually exist between segments.

EXERCISES

The following exercises require the use of a computer and software. The answers may be calculated manually using the sample statistics based on the data provided.

10.96 XR10-96 A new credit card company is investigating various market segments to determine whether it is profitable to direct its advertising specifically at each one. One of the market segments is composed of Asian migrants. The latest census indicates that there are 2.01 million Asian migrants in Australia. A survey of 475 Asians asked each how they usually pay for the products they purchase. The responses are:

- 1 Cash
- 2 Cheque
- 3 Visa
- 4 MasterCard
- 5 Other credit card.

The responses are recorded. Estimate with 95% confidence the number of Asian migrants in Australia who usually pay by credit card.

Sample frequencies: $n(1) = 81$; $n(2) = 47$; $n(3) = 167$; $n(4) = 146$; $n(5) = 34$.

10.97 XR10-97 A university in Victoria is investigating expanding its evening programs. It wants to target people aged between 25 and 35 years, who have completed high school but did not complete a university degree. To help determine the extent and type of offerings, the university needs to know the size of its target market. A survey of 320 adults aged between 25 and 35 years was drawn and each person was asked to identify his or her highest educational attainment. The responses are:

- 1 Did not complete high school
- 2 Completed high school only
- 3 Completed some vocational study
- 4 A university graduate

The responses are recorded. In 2019, there were about 1147315 people between the ages of 25 and 35 in Victoria. Estimate with 95% confidence the number of adults in Victoria between 25 and 35 years of age who were in the market segment the university wishes to target.

Sample frequencies: $n(1) = 63$; $n(2) = 125$; $n(3) = 45$; $n(4) = 87$.

10.98 XR10-98 A major department store chain segments the market for women's apparel by its identification of values. There are three segments:

- 1** Conservative
- 2** Traditional
- 3** Contemporary

Questionnaires about personal and family values are used to identify into which segment a woman falls. Suppose that the questionnaire was sent to a random sample of 1836 women. Each

woman was classified using the codes 1, 2 and 3. The data are recorded using these codes. The 2014 ABS figures indicate that there are about 9225090 adult women in Australia.

- a** Estimate with 95% confidence the proportion of adult Australian women who are classified as traditional.
- b** Estimate the size of the traditional market segment.

Sample frequencies: $n(1) = 418$; $n(2) = 536$; $n(3) = 882$.

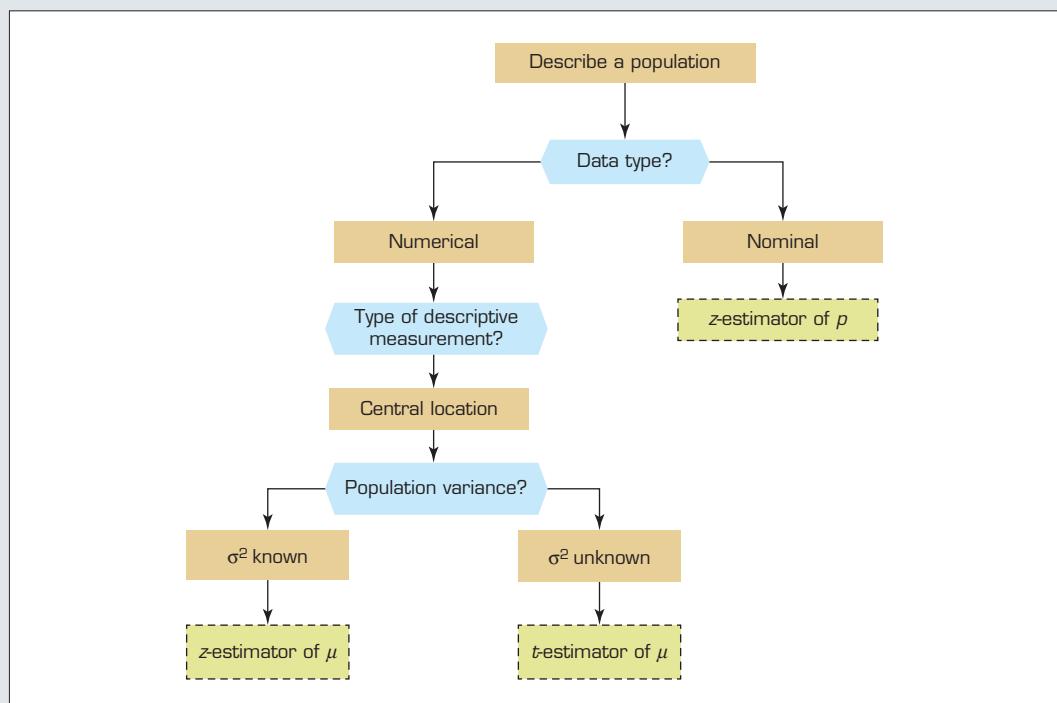
Study Tools

CHAPTER SUMMARY

There are two types of statistical inference: *estimation* and *hypothesis testing*. Both types of inference can be applied to problems in which the objective is to describe a single population and in which the data type is either numerical or nominal. In this chapter, we considered the statistical inference type, estimation. We estimate the *parameters* μ and p using an *interval estimator* with a specified *confidence level*. There are two different estimators of μ : one is based on the *standard normal distribution* (*z-formula*), which is used when σ^2 is known; and the other is based on the *Student t distribution* formula, which is used when σ^2 is unknown and σ^2 is replaced with s^2 . Student *t* distribution requires that the population from which we are sampling is normally distributed. As \hat{p} is approximately normally distributed, the formula used to estimate p involves the standard normal random variable *z*.

Table 10.5 summarises the relevant formulas and the conditions required for their use.

In this chapter we also discussed how to determine the sample sizes required to estimate μ and p . These formulas are shown in **Table 10.6**.



The data files for Examples, Exercises and Cases are provided on the companion website (accessible through <https://login.cengagebrain.com/>).

SYMBOLS

Symbol	Pronounced	Represents
$1 - \alpha$	One-minus-alpha	Confidence level
B		Error bound or maximum allowable error of estimation
$z_{\alpha/2}$	<i>z-alpha-divided-by-2</i>	Value of <i>z</i> such that the area to its right is equal to $\alpha/2$

SUMMARY OF FORMULAS

TABLE 10.5 Summary of interval estimators of μ and p

Parameter	Confidence interval estimator	Required conditions
μ (Numerical data)	$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$	σ^2 is known; X is normally distributed or $n \geq 30$
μ (Numerical data)	$\bar{X} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$	σ^2 is unknown and estimated by s^2 ; X is normally distributed
p (Nominal data)	$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$	$n\hat{p} \geq 5$ and $n\hat{q} \geq 5$, where $\hat{q} = 1 - \hat{p}$

TABLE 10.6 Summary of sample sizes for estimating μ and p

Parameter	Sample size	Require condition
μ (Numerical data)	$n = \left(\frac{z_{\alpha/2} \sigma}{B} \right)^2$ or $\frac{z_{\alpha/2}^2 \sigma^2}{B^2}$	σ can be estimated from previous knowledge or approximated by range/4
p (Nominal data)	$n = \left(\frac{z_{\alpha/2} \sqrt{\hat{p}\hat{q}}}{B} \right)^2$ or $\frac{z_{\alpha/2}^2 \hat{p}\hat{q}}{B^2}$	\hat{p} can be estimated from previous knowledge or $\hat{p} = 0.5$; $\hat{q} = 1 - \hat{p}$

SUPPLEMENTARY EXERCISES

10.99 A health department official is investigating the possibility of allowing doctors to advertise their services. In a survey designed to examine this issue, 91 doctors were asked whether they believed that doctors should be allowed to advertise. A total of 23 respondents supported advertising by doctors. Estimate with 90% confidence the proportion of all doctors who support advertising.

10.100 In a study to determine the size of loan requests at a suburban bank, the mean in a random sample of 25 requests was \$15 000, with a standard deviation of \$4000. Assuming that loan requests are normally distributed, estimate with 95% confidence the mean loan request.

10.101 A firm's management is contemplating modifying one of its products. To help in making a decision, management representatives wish to conduct a market survey that will enable them to estimate the proportion of potential customers who would buy the new product. They wish to estimate this proportion to within 4% with 95% confidence. How large a sample should be drawn?

10.102 A university wants to determine the average income their students earn during the summer. A random sample of 25 second-year business students produced the following statistics (where x is measured in hundreds of dollars):

$$\sum x_i = 826.6 \quad \sum x_i^2 = 27\,935.7$$

- a Estimate the mean summer employment income for all second-year business students, with 99% confidence.
- b Does the estimate in part (a) pertain to all business students? Does it pertain to all university students? Explain.

10.103 A duty-free shop in Perth is considering moving from its present mid-city location to a location in Fremantle. One factor in this decision is the amount of time the company's employees spend getting to work. A random sample of 20 employees reveals that the mean and the standard deviation of the time required to get to work are 36.5 and 11.3 minutes respectively. Estimate with 95% confidence the mean time spent getting to work for all of this shop's employees.

10.104 A department store wants to estimate the average account of all its credit-card customers to within \$20, with 99% confidence. A quick analysis tells us that the smallest account is \$0, while the largest is \$1000.

- a Determine the sample size.
 - b Suppose that a survey was performed, and the sample mean is \$300. Find a 99% confidence interval estimate of μ , assuming that the value of σ used in part (a) is correct.
- (Hint: This question should take you no more than 5 seconds to answer.)

10.105 In a survey, a random sample of 500 recent university graduates, who had studied part-time, were asked how many years it had taken them to complete their bachelor's degrees. The results (in percentages) are as follows:

Five years	49%
Six years	27%
Seven years	9%
Eight years	15%

Estimate with 95% confidence the mean number of years taken to complete the bachelor's degree.

10.106 XR10-106 Researchers asked a random sample of business executives how many days they take for holidays annually. The results appear in the following table (in percentages). If the sample size was 800, estimate with 99% confidence the mean number of days spent on vacation by business executives.

Days*	Frequency (%)
0–5	9
5–10	24
10–15	31
15–20	23
20–25	13

*Each interval includes the lower limit but excludes the upper limit.

Computer/manual applications

The following exercises require a computer and software. Alternatively, they can be solved manually using the sample statistics provided.

10.107 XR10-107 The routes of postal deliveries are carefully planned so that each deliverer works between 7.0 and 7.5 hours per shift. The planned routes assume an average walking speed of 2 km/h and no shortcuts across lawns. In an experiment to examine the amount of time deliverers actually spend completing their shifts, a random sample of 75 postal deliverers were secretly timed. The data from the survey are recorded and some of these are shown below.

6.9	6.9	7.3	...	7.1	7.0	7.0
-----	-----	-----	-----	-----	-----	-----

Sample statistics: $n = 75$; $\bar{X} = 6.91$; $s = 0.226$.

- a Estimate with 99% confidence the mean shift time for all postal deliverers.
- b Check to determine if the required condition for this statistical inference is satisfied.

10.108 XR10-108 The manager of a branch of a major bank wants to improve service. She is thinking about giving \$1 to any customer who waits in line for a period of time that is considered excessive. (The bank ultimately decided that more than 8 minutes is excessive.) However, to get a better idea about the level of current service, she undertakes a survey of customers. A student is hired to measure the time spent waiting in line by a random sample of 50 customers. Using a stopwatch, the student determined the amount of time between the time the customer joined the line and the time he or she reached a teller. The times were recorded and some are listed below.

1.4	6.1	10.4	...	10.8	7.3	7.9
-----	-----	------	-----	------	-----	-----

Sample statistics: $n = 50$; $\bar{X} = 5.79$; $s = 2.86$.

- a Construct a 90% confidence interval estimate of the mean customer waiting time.
- b Check to ensure that the required condition for the estimate is satisfied.

10.109 XR10-109 A manufacturer of designer jeans has pitched her advertising to develop an expensive and classy image. The suggested retail price is \$75. However, she is concerned that retailers are undermining her image by offering the jeans

at discount prices. To better understand what is happening, she randomly samples 30 retailers who sell her product, determines the prices and records them. She would like a 95% confidence interval estimate of the mean selling price of the jeans at all retail stores. (*Caution:* Missing data.)

- a Determine the 95% confidence interval estimate.
- b What assumption must be made to be sure that the estimate produced in part (a) is valid? Use a graphical technique to check the required condition.

Sample statistics: $n = 28$; $\bar{X} = 62.79$; $s = 5.32$.

10.110 XR10-110 In an examination of consumer loyalty in the travel business, 72 first-time visitors to a tourist attraction were asked whether they planned to return. The responses were recorded with coding 2 = Yes and 1 = No. Estimate with 95% confidence the proportion of all first-time visitors who planned to return to the same destination.

Sample data frequencies: $n(1) = 24$; $n(2) = 48$.

Case Studies

CASE 10.1 Estimating the monthly average petrol price in Queensland

C10-01 The Royal Automobile Club Queensland (RACQ) publishes the monthly average unleaded petrol prices (cents/litre) for selected Queensland centres. The data below are for the month of September 2019. Estimate a 90% confidence interval for the mean unleaded petrol price for Queensland. Also, state any assumptions you would need to make, and verify them using the data.

Qld Locality	Average petrol price	Qld Locality	Average petrol price
Atherton	143.9	Lockyer Valley	141.2
Beaudesert	146.0	Longreach	155.1
Biloela	149.7	Mackay	139.2
Blackwater	159.9	Mareeba	145.7
Bowen	141.9	Maryborough	138.6
Brisbane Metro	144.4	Miles	135.1
Bundaberg	136.1	Moranbah	138.9
Cairns	139.5	Mount Isa	148.7
Charters Towers	145.5	Nambour	144.7
Childers	143.7	Noosa	147.8
Dalby	138.9	Rockhampton	141.0
Emerald	149.9	Roma	134.4
Gladstone	135.4	Somerset	141.4
Goondiwindi	134.2	Toowoomba	138.8
Gympie	136.8	Townsville	137.7
Hervey Bay	142.0	Tully	142.9
Ingham	148.2	Warwick	134.6
Innisfail	140.2	Whitsunday	133.4
Kingaroy	141.2	Yeppoon	137.6

Source: Royal Automobile Club Queensland (RACQ) calculations using Oil Price Information Service (OPIS) data, September 2019.

CASE 10.2 Cold men and cold women will live longer!

C10-02 A recent press release lists a number of regions in Australia where people enjoy Australia's highest life expectancies, with Canberra (ACT) in the lead for both women and men. The following table is a sample of locations in various Australian states and territories listing the life expectancy of men and women. Based on this information, the Minister for Ageing would like you to estimate the average life expectancy of male and female Australians. Also give 95% interval estimates for the two parameters: the average life expectancy of male and female Australians.

Life expectancy in years of males and females (years), Australia

City	Men	Women
Adelaide	79.7	84.0
Brisbane	79.5	84.1
Canberra	81.0	84.8
Darwin	74.9	80.5
Hobart	78.3	82.5
Melbourne	80.3	84.4
Perth	80.1	84.6
Sydney	79.8	84.2

Source: Australian Bureau of Statistics, November 2012, Deaths Australia, Australian Bureau of Statistics, cat no 3302.0, ABS Canberra.

CASE 10.3 Super fund managers letting down retirees

C10-03 A recent report from the OECD entitled ‘Pensions at a Glance’ compared Australia with 33 other countries and ranked it as the second lowest on social equity, with 36% of pensioners living below the poverty line. It also reported that the Australian government spends 3.5% of its GDP on pensions, which is below the OECD average of 7.9%.

The following table provides the pension fund’s 5-year average annual real investment rates of return for 2008–13 for a randomly selected number of OECD countries. Construct a 95% interval estimate for the pension fund’s 5-year average real investment return for the developed world for the period 2008–13 and comment on the Australian pension fund.

Pension fund’s 5-year real net investment rates of return in selected OECD countries, 2008–13

Country	Percentage return
Australia	2.1%
Canada	7.4%
Greece	-0.3%
Japan	3.8%
Netherlands	7.4%
New Zealand	2.8%
Norway	5.8%
South Korea	1.1%
Spain	2.7%
US	5.7%

Source: *Pension Markets in Focus*, OECD, 2014.

Appendix 10.A

Excel instructions for missing data and for recoding data

Missing data

Excel (the techniques listed under **Data Analysis...**) addresses the problem inconsistently. Some functions recognise blank cells as missing data; others do not. The safest and easiest way to omit missing data is to simply delete the blank cells. Sort the data (Click **Data** and **Sort...**) in order, highlight the cells you wish to delete, and with one keystroke delete the missing data.

Recoding data

To recode data we employ a logical function. Click **fx**, **Logical**(Function category), and **IF**(Function name). To illustrate suppose that in column A you have stored data consisting of codes 1–6 and you wish to convert all 4s, 5s and 6s to 9s. Activate cell B1 and type

=IF(A1>=4,9,A1)

This logical function determines whether the value in cell A1 is greater than or equal to 4. If so, Excel places a 9 in cell B1. If A1 is less than 4, B1 = A1. Dragging to fill in column B converts all 4s, 5s, and 6s to 9s and stores the results in column B.

If 4s, 5s and 6s represent nonresponses you can replace these codes with a blank. Type the following in cell B1:

=IF(A1>=4,"",A1)

Estimation: Two populations

Learning objectives

This chapter discusses the sampling distribution and estimation of the difference in population means and the difference in population proportions.

At the completion of this chapter, you should be able to:

- L01** recognise when the parameter of interest is the difference between two population means or proportions
- L02** estimate the difference between two population means when the population variances are known
- L03** estimate the difference between two population means when the population variances are unknown
- L04** recognise when the samples are independently drawn from two populations and when they are taken from a matched pairs experiment
- L05** estimate the difference between two population means in a matched pairs experiment
- L06** estimate the difference between two population proportions
- L07** calculate the minimum sample size required to estimate the difference between two population means and between two population proportions.

CHAPTER OUTLINE

Introduction to statistics

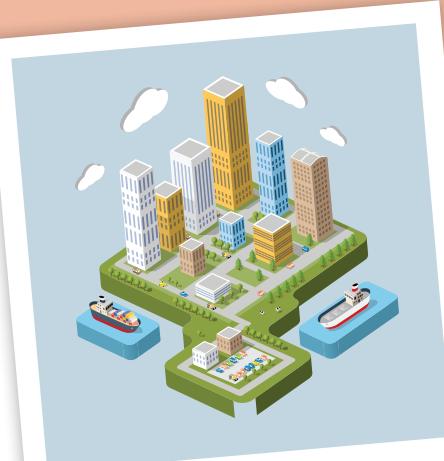
- 11.1** Estimating the difference between two population means ($\mu_1 - \mu_2$) when the population variances are known: Independent samples
- 11.2** Estimating the difference between two population means ($\mu_1 - \mu_2$) when the population variances are unknown: Independent samples
- 11.3** Estimating the difference between two population means with matched pairs experiments: Dependent samples
- 11.4** Estimating the difference between two population proportions, $p_1 - p_2$

SPOTLIGHT ON STATISTICS

Selecting a location for a new department store

Often when an upmarket department store chain such as Myer or David Jones wishes to open a new store in a region, management needs to make a decision regarding the location of the new store based on a number of consumer characteristics. One of the main location-based characteristics used in such decision making is the average income of residents living in the surrounding areas.

The management of a chain of department stores wants to know if there is a difference in the average annual income of potential customers at two possible sites for a new store. In



Source: Shutterstock.com/Alexzel

one location, a random sample of 100 households showed a mean annual income of \$166 000. In the other location, the mean annual income of 75 households was \$134 000. Assume that annual income in both locations are normally distributed with a standard deviation of \$10 000. Estimate with 99% confidence the difference between the average annual incomes in the two locations. The solution is shown on pages 435–6.

Introduction

We can compare learning how to use statistical techniques to learning how to drive a car. We began by describing what you are going to learn in this book (Chapter 1) and then presented the essential background material (Chapters 2–8). Learning the concepts of statistical inference and applying them the way we did in Chapter 9 is comparable to driving a car in an empty parking lot. You are driving, but it's not a realistic experience. Learning the concepts about a single population in Chapter 10 is like driving on a quiet side street with little traffic. The experience represents real driving, but many of the difficulties have been eliminated. In this chapter, you begin to drive for real, with many of the actual problems faced by licensed drivers, and the experience prepares you to tackle the next difficulty. In this chapter and Chapter 13, we present a variety of techniques used to compare two population parameters.

Recall that in Chapter 10 we discussed the basic principles of estimation and presented the confidence interval estimators used to describe a single population with both numerical and nominal data types. In this chapter we extend our presentation to cover a variety of estimation methods when the objective involves comparing two populations.

In Sections 11.1–11.3, we deal with numerical variables; the parameter of interest is the difference between two means, $\mu_1 - \mu_2$. The difference between the data in Sections 11.1 and 11.2 and Section 11.3 introduces yet another factor that determines the correct statistical method – the design of the experiment used to gather the data. In Sections 11.1 and 11.2, the samples are independently drawn, but in Section 11.3 the samples are taken from a matched pairs experiment.

Section 11.4 addresses the problem of comparing two populations of nominal data. The parameter to be tested and estimated is the difference between two proportions, $p_1 - p_2$.

We offer the following examples to illustrate applications of these estimation techniques.

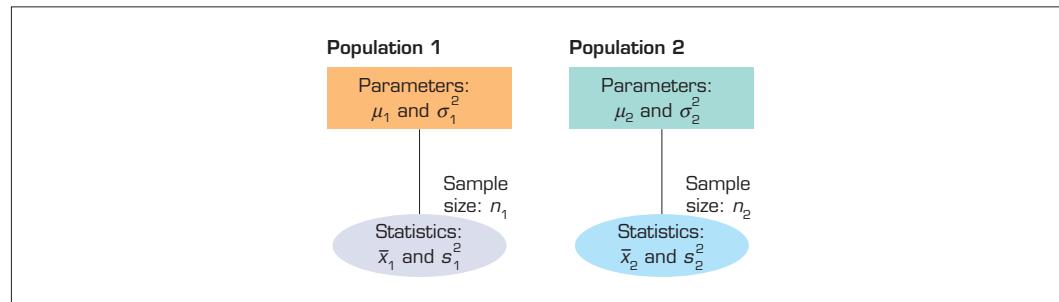
- 1** Firms that use components manufactured by other companies in producing their own finished products are often concerned about the quality, reliability and price of the components. If two competing suppliers of a component are available, the firm's manager may wish to compare the reliability of the two products. For example, a car manufacturer currently equips its product with a certain brand of tyre. If a similarly priced brand of tyre becomes available, the decision about which brand to use should be based on which tyre, on average, lasts longer. In this situation, the data type is numerical (tyre life is usually measured by the number of kilometres until wear-out), and the problem objective is to compare the life of the two populations of tyres. The parameter to be estimated is the difference between the two population means, $\mu_1 - \mu_2$, where μ_1 and μ_2 are average life times of brand 1 and brand 2 tyres respectively.
- 2** Market managers and advertisers are eager to know which segments of the population are buying their products. If they can determine these groups, they can target their advertising messages and tailor their products to these customers. For example, if advertisers determine that the decision to purchase a particular household product is made more frequently by men than by women, the interests and concerns of men will be the focus of most commercial messages. The advertising media used also depend on whether the product is of greater interest to men or to women. The most common way of measuring this factor is to find the difference in the proportions of men (p_1) and women (p_2) buying the product. In these situations, the data type is nominal and the problem objective is to compare the two populations (male and female). The parameter to be estimated is the difference between two proportions, $p_1 - p_2$.

11.1 Estimating the difference between two population means ($\mu_1 - \mu_2$) when the population variances are known: Independent samples

In order to estimate the difference between two population means, the statistics practitioner draws independent random samples from each of the two populations. We define independent samples as samples that are completely unrelated to one another.

Figure 11.1 depicts the sampling process. Observe that we draw a sample of size n_1 from population 1 and a sample of size n_2 from population 2. For each sample, we calculate the sample mean and the sample variance.

FIGURE 11.1 Independent samples from two populations



In Chapter 9, we showed that \bar{X} is the best estimator of μ and \bar{X} is approximately normally distributed when we have a reasonably large sample. By a similar analysis, we can show that $\bar{X}_1 - \bar{X}_2$ is the best estimator of $\mu_1 - \mu_2$, where \bar{X}_1 is the mean of a sample of size n_1 from a large population whose mean and variance are μ_1 and σ_1^2 respectively, and where \bar{X}_2 is the mean of a sample of size n_2 from another large population with mean and variance μ_2 and σ_2^2 respectively.

11.1a Sampling distribution of $\bar{X}_1 - \bar{X}_2$

The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is derived from an extension of the central limit theorem and the basic rules of expectation and variance. It can be shown that \bar{X} is normally distributed when X is normally distributed, or \bar{X} is approximately normally distributed when n is sufficiently large (central limit theorem, see Chapter 9). Similarly, it can also be shown that $\bar{X}_1 - \bar{X}_2$ is normally distributed when X_1 and X_2 are normal. If the populations are non-normal, then the sampling distribution is only approximately normal for large sample sizes. The required sample sizes depend on the extent of non-normality. However, for most populations, sample sizes of 30 or more are sufficient. By the rules of expectation and variance, we derive the expected value and the variance of the sampling distribution of $\bar{X}_1 - \bar{X}_2$:

$$\begin{aligned} \text{Mean} &= \mu_{\bar{x}_1 - \bar{x}_2} = E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2 \\ \text{Variance} &= \sigma_{\bar{x}_1 - \bar{x}_2}^2 = V(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \end{aligned}$$

In the above derivation of variance of $\bar{X}_1 - \bar{X}_2$, it is assumed that the two samples are independent. Thus, it follows that in repeated independent sampling from two populations

with means μ_1 and μ_2 and standard deviations σ_1 and σ_2 , respectively, the sampling distribution of $\bar{X}_1 - \bar{X}_2$ is normal with mean

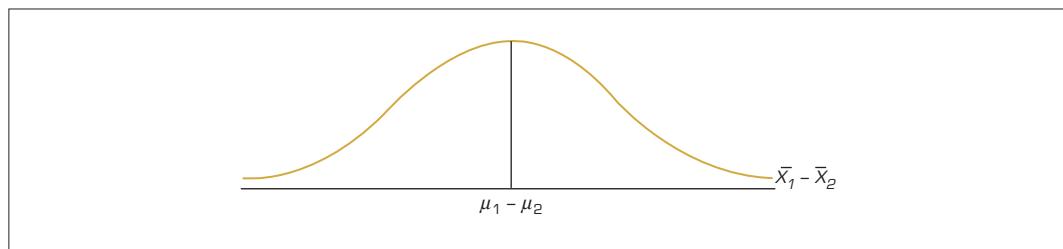
$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

and standard deviation (which is the *standard error of the difference between two means*) is

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is depicted in **Figure 11.2**.

FIGURE 11.2 Sampling distribution of $\bar{X}_1 - \bar{X}_2$



Sampling distribution of $\bar{X}_1 - \bar{X}_2$

- 1 $\bar{X}_1 - \bar{X}_2$ is normally distributed if the populations are normal, and approximately normal if the populations are non-normal and the sample sizes are large.
- 2 The expected value of $\bar{X}_1 - \bar{X}_2$ is

$$E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$$

- 3 The variance of $\bar{X}_1 - \bar{X}_2$ is

$$V(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

- 4 The standard error of $\bar{X}_1 - \bar{X}_2$ is

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

As mentioned at the beginning of this section, we assume that the samples are independent. This means that the selection of sample members drawn from one population is not affected by the selection of sample members drawn from the other population. The issue of dependent samples is discussed in greater detail in Sections 11.3 and 13.2. Consequently, all the examples and exercises deal with independent samples until then. We also assume that the two population variances are known – even though this assumption is quite unrealistic. The more realistic case is dealt with in Section 11.2, in which we assume that σ_1^2 and σ_2^2 are unknown.

We now extend our work in Chapter 10 based on the fact that the sampling distribution of $\bar{X}_1 - \bar{X}_2$ is approximately normal, with mean $\mu_1 - \mu_2$ and variance $(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)$.

11.1b Estimating $\mu_1 - \mu_2$ when σ_1^2 and σ_2^2 are known

Using exactly the same set of arithmetic operations we used in Chapter 10 to develop the confidence interval estimators for a single population mean μ , we determine the confidence interval estimator of the difference between two population means.

Confidence interval estimator of $\mu_1 - \mu_2$, when σ_1^2 and σ_2^2 are known

$$(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Keep in mind that this interval estimator is valid as long as the populations are normal or the sample sizes are large enough, the population variances are known and the samples are independent.

EXAMPLE 11.1

L01 L02

CEO Salaries: Males vs Females

Until recently, most Chief Executive Officers (CEOs) of companies have been men. However, as government encourages companies to increase the number of female CEOs, a number of companies now have female CEOs. A researcher decided to examine the success of female CEOs of medium-size companies by comparing their salaries (excluding bonuses and outliers) with those of their male counterparts. The researcher took a random sample of 100 male and 100 female CEOs and recorded their salaries for the preceding year. The mean income for male CEOs was \$430 000 and for the female CEOs average salary was \$402 500. From past records, we have information that CEO incomes are normally distributed with a standard deviation of \$7000 for both males and females. Estimate with 90% confidence the difference in mean income between all male and female CEOs of medium-size companies.

Solution

Identifying the technique

The problem objective is to compare two populations, and the data type is numerical.

Consider the two populations,

X_1 = Income earned by male CEOs of medium-size companies

X_2 = Income earned by female CEOs of medium-size companies

We wish to compare the mean income of male CEOs with the mean income of female CEOs, from medium-size companies. We define

μ_1 = mean income earned by male CEOs of medium-size companies

μ_2 = mean income earned by female CEOs of medium-size companies

Therefore, the parameter to be estimated is $\mu_1 - \mu_2$.

As both populations are normally distributed, the distribution of $\bar{X}_1 - \bar{X}_2$ is also normal. Furthermore, σ_1^2 and σ_2^2 are both known ($\sigma_1 = \sigma_2 = 7000$). Therefore, we use the z-estimate. The $(1 - \alpha)100\%$ confidence interval estimator for $\mu_1 - \mu_2$ is

$$(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

► Calculating manually

We are given the following values:

$$\bar{X}_1 = 430000; \quad \bar{X}_2 = 402500$$

$$\sigma_1 = \sigma_2 = 7000$$

$$n_1 = n_2 = 100$$

The confidence level is

$$1 - \alpha = 0.90$$

Hence,

$$z_{\alpha/2} = z_{0.05} = 1.645$$

The 90% confidence interval estimate of $\mu_1 - \mu_2$ is

$$\begin{aligned} (\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} &= (430000 - 402500) \pm 1.645 \sqrt{\frac{7000^2}{100} + \frac{7000^2}{100}} \\ &= 27500 \pm 1628 \end{aligned}$$

The confidence limits are

$$\text{LCL} = \$25872 \text{ and UCL} = \$29128$$

Interpreting the results

We estimate that the difference between the mean incomes of male and female CEOs of medium-size companies lies between \$25872 and \$29128. Expressed another way, the mean income of male CEOs is between \$25872 and \$29128 more than mean income of female CEOs. Consequently, company managers need to seriously consider ways of reducing the gender gap between salaries of male and female CEOs.

Using the computer

Excel doesn't provide confidence interval estimates for $\mu_1 - \mu_2$. However, the interval estimates can be calculated using the **z-Estimate_2Means** worksheet in the **Estimators** workbook that is provided on the companion website (accessible through <https://login.cengagebrain.com/>). The output is shown below.

Excel output for Example 11.1

	A	B	C	D	E	F
1	z-Estimate of the Difference Between Two Means (Known Variances)					
2						
3		Sample 1	Sample 2	Confidence Interval Estimate		
4	Sample mean	430000.0	402500.0	27500.00	±	1628.3
5	Population variance	49000000.0	49000000.0	Lower confidence limit		25871.7
6	Sample size	100	100	Upper confidence limit		29128.3
7	Confidence level	0.90				

COMMANDS

To estimate the difference between two means with known variances, first calculate the sample means of the two independent sample observations. Then open the **z-Estimate_2Means** worksheet in the **Estimators** workbook and type in the values of the sample means, the given population variances and the sample sizes, as well as the confidence level.

Now we will answer the opening problem described in this chapter's introduction.

SPOTLIGHT ON STATISTICS

Selecting a location for a new upmarket department store: Solution

Identifying the technique

The problem objective is to compare two population means. The data are numerical. Consequently, the parameter we wish to estimate is $\mu_1 - \mu_2$, where

μ_1 = mean income earned by households in Location 1

μ_2 = mean income earned by households in Location 2

As both populations are normally distributed, the distribution of $\bar{X}_1 - \bar{X}_2$ is also normal. Furthermore, σ_1^2 and σ_2^2 are both known. Therefore, the confidence interval estimator is

$$(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

We are given the following values (in '000):

$$\bar{X}_1 = 166000; \quad \bar{X}_2 = 134000$$

$$\sigma_1 = 10000; \quad \sigma_2 = 10000$$

$$n_1 = 100; \quad n_2 = 75$$

The confidence level is

$$1 - \alpha = 0.90$$

Hence,

$$z_{\alpha/2} = z_{0.05} = 1.645$$

The 90% confidence interval estimate of $\mu_1 - \mu_2$ is

$$\begin{aligned} (\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} &= (166000 - 134000) \pm 1.645 \sqrt{\frac{10000^2}{100} + \frac{10000^2}{75}} \\ &= 32000 \pm 2512.56 \end{aligned}$$

The confidence limits are

$$\text{LCL} = 29487 \text{ and UCL} = 34513$$

Interpreting the results

We estimate that the difference between the average income of households in location 1 and location 2 lies between \$29487 and \$34513. Expressed another way, the mean income of households in location 1 is between \$29487 and \$34513 higher than the mean income of households in location 2. Consequently, management can be more confident about choosing location 1 in which to open the new store.

Using the computer

Excel doesn't provide confidence interval estimates for $\mu_1 - \mu_2$. However, the interval estimates can be calculated using the **z-Estimate_2Means** worksheet in the **Estimators** workbook. The commands are as in Example 11.1. The Excel output is shown below.



Source: TS shutterstock.com/Alexzel



	A	B	C	D	E	F
1	z-Estimate of the Difference Between Two Means (Known Variances)					
2						
3		Sample 1	Sample 2	Confidence Interval Estimate		
4	Sample mean	166000.0	134000.0	32000.00	±	2512.56
5	Population variance	1000000000	1000000000	Lower confidence limit		29487.44
6	Sample size	100	75	Upper confidence limit		34512.56
7	Confidence level	0.90				

Here is a brief summary of the factors that tell us when to use the z -interval estimator of the difference between two population means.

IN SUMMARY

Factors that identify the z -interval estimator of $\mu_1 - \mu_2$

- 1 *Problem objective:* to compare two populations
- 2 *Data type:* numerical (quantitative)
- 3 *Descriptive measurement:* central location
- 4 *Population variances:* known
- 5 *Experimental design:* independent samples
- 6 *Population distributions:* either normal or sample sizes are large enough to apply the central limit theorem ($n > 30$)

11.1c Selecting the sample sizes to estimate $\mu_1 - \mu_2$

The method used to determine the sample sizes required to estimate $\mu_1 - \mu_2$ is a simple extension of the method used to determine the sample size to estimate μ (Chapter 10). Since the confidence interval estimator is

$$(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

the maximum error of estimation or error bound B is set equal to

$$B = z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

In order to solve for n_1 and n_2 , we specify the confidence level $(1 - \alpha)$, and the values of σ_1^2 and σ_2^2 . Finally, we let $n_1 = n_2$, and solve the equation below.

Sample sizes necessary to estimate $\mu_1 - \mu_2$

$$n_1 = n_2 = \left[\frac{z_{\alpha/2} \sqrt{\sigma_1^2 + \sigma_2^2}}{B} \right]^2 = \frac{z_{\alpha/2}^2 (\sigma_1^2 + \sigma_2^2)}{B^2}$$

EXAMPLE 11.2

LO7

Comparing the clothing purchasing habits of men and women

The manager of a major clothing manufacturer wants to compare the annual expenditure on clothing of men and women. She decides to estimate the difference in mean annual expenditure to within \$100 with 95% confidence. How large should the two sample sizes be, if we assume that the range of expenditure is \$800 for males and \$1200 for females and the populations of male and female expenditures on clothing are normal?

Solution**Identifying the technique**

The problem objective is to determine the required sample sizes. The data type is numerical. Both populations of interest are normally distributed and the samples are independent.

Calculating manually

The error bound is \$100 and the confidence level is 0.95. Hence, $B = \$100$ and $z_{\alpha/2} = z_{0.025} = 1.96$.

We approximate σ_1 and σ_2 by using the following formulas:

$$\sigma_1 \approx \frac{\text{Range}}{4} = \frac{800}{4} = \$200 \text{ (for males)}$$

$$\sigma_2 \approx \frac{\text{Range}}{4} = \frac{1200}{4} = \$300 \text{ (for females)}$$

Thus,

$$n_1 = n_2 = \left[\frac{1.96 \sqrt{200^2 + 300^2}}{100} \right]^2 = 49.94 \approx 50$$

Interpreting the results

In order to estimate $\mu_1 - \mu_2$ to within \$100 with 95% confidence, we should take samples of 50 men and 50 women.

Using the computer

The required sample sizes can be calculated using the **Sample size-2Means** worksheet in the **Estimators** workbook. The output is shown below.

Excel output for Example 11.2

	A	B	C
1	Calculating the sample sizes for a fixed width (2B)		
2		Sample 1	Sample 2
3	Variance	40000	90000
4	Width/2 = B	100	
5	Confidence level	0.95	
6	Sample size	50	50

COMMANDS

To estimate required sample sizes, open the **Sample size-2Means** worksheet in the **Estimators** workbook, then type in the values of the given population variances (or as calculated above), B and the confidence level.

EXERCISES

Learning the techniques

- 11.1** Assume you are given the following information:

$n_1 = 100$	$n_2 = 200$
$\bar{X}_1 = 510$	$\bar{X}_2 = 480$
$\sigma_1 = 80$	$\sigma_2 = 90$

Determine the 95% confidence interval estimate of $\mu_1 - \mu_2$.

- 11.2** A random sample of 25 observations from a normal population whose variance is 110 produced a sample mean of 45. A random sample of 40 from another normal population whose variance is 250 had a sample mean of 80. Estimate with 90% confidence the difference between the two population means.

- 11.3** The following information has been received:

$n_1 = 50$	$n_2 = 50$
$\bar{X}_1 = 175$	$\bar{X}_2 = 150$
$\sigma_1 = 40$	$\sigma_2 = 50$

- a Estimate $\mu_1 - \mu_2$ with 99% confidence.
- b Repeat part (a) with 95% confidence.
- c Repeat part (b) with 90% confidence.

Applying the techniques

- 11.4 Self-correcting exercise.** In order to help make a financial decision, an investor observes 25 returns on one type of investment and 35 returns on a second type of investment. The sample means are $\bar{X}_1 = \$12.5$ and $\bar{X}_2 = \$11.3$. Assume that the returns are normally distributed with standard deviations $\sigma_1 = \sigma_2 = \$5$. Estimate the difference in mean returns with 95% confidence.

- 11.5** A survey of 200 second-year university business students revealed that their mean monthly income from summer jobs was \$1150. Another survey of 200 third-year university business students showed that their mean monthly income from summer jobs

was \$1300. If monthly incomes from summer jobs are normally distributed, with population variances σ_1^2 (second year) = 35000 and σ_2^2 (third year) = 42000, estimate the difference in mean monthly income from summer jobs between second- and third-year business students. Use a confidence level of 90%.

- 11.6** To compare the hourly wages paid to workers of two large companies, random samples of 50 wage earners are drawn from each company. The average hourly wages of the two samples of wage earners are $\bar{X}_1 = \$22.50$ and $\bar{X}_2 = \$23.70$. Assuming that the population standard deviations of wages in the two companies are $\sigma_1 = 6.00$ and $\sigma_2 = 5.40$, estimate with 95% confidence the difference in average hourly wages between the two companies.

- 11.7** The management of a chain of department stores wants to know if there is a difference in the average annual income of potential customers at two possible sites for a new store. In one location, a random sample of 100 households showed a mean annual income of \$63000. In the other location, the mean annual income of 75 households was \$67000. Answer the following questions assuming that $\sigma_1 = \sigma_2 = \$5000$.
- a Estimate with 99% confidence the difference between the average annual incomes in the two locations.
 - b How large should the sample sizes be in order to estimate the difference between average incomes in the two areas to within \$1000 with 99% confidence?

- 11.8** An advertising consultant for a major brewer wants to compare the annual beer consumption of men with that of women. She decides to estimate the difference in mean annual consumption to within 10 litres with 99% confidence. How large should the sample sizes be, if we assume that the range of consumption is 400 litres for males and 200 litres for females and the two populations of consumption are normally distributed?

11.2 Estimating the difference between two population means ($\mu_1 - \mu_2$) when the population variances are unknown: Independent samples

In the previous section, our inferences about $\mu_1 - \mu_2$ assumed that the population variances σ_1^2 and σ_2^2 were known. Unfortunately, it is quite unusual for μ_1 and μ_2 to be unknown while σ_1^2 and σ_2^2 are known. In this section, we address the more realistic problem of making inferences about $\mu_1 - \mu_2$ when σ_1^2 and σ_2^2 are unknown.

11.2a Estimating $\mu_1 - \mu_2$ when σ_1^2 and σ_2^2 are unknown and unequal ($\sigma_1^2 \neq \sigma_2^2$)

When the population variances are unknown, we proceed along the same lines as we used in Chapter 10, in which we estimated the population mean for the case when the population variance is unknown. That is, we substitute the sample variances s_1^2 and s_2^2 for the unknown population variances σ_1^2 and σ_2^2 . If the population variances are unknown and estimated by sample variances s_1^2 and s_2^2 , we use the t -interval estimator given below to estimate $\mu_1 - \mu_2$. This estimator requires that the following occur:

- 1 The two population random variables X_1 and X_2 are normally distributed.
- 2 The population variances are unknown and not equal.
- 3 The two samples are independent.

Confidence interval estimator of $\mu_1 - \mu_2$, when σ_1^2 and σ_2^2 are unknown and $\sigma_1^2 \neq \sigma_2^2$

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2, \text{d.f.}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \text{where d.f.} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left(s_1^2/n_1 \right)^2}{n_1-1} + \frac{\left(s_2^2/n_2 \right)^2}{n_2-1}}$$

EXAMPLE 11.3

L01 L03

Dietary effects of high-fibre breakfast cereals: Part I

XM11-03 Despite some controversy, scientists generally agree that high-fibre cereals reduce the likelihood of various forms of cancer. However, one scientist claims that people who eat high-fibre cereal for breakfast will consume, on average, fewer kilojoules for lunch than people who don't eat high-fibre cereal for breakfast. If this is true, high-fibre cereal manufacturers will be able to claim another advantage of eating their product – potential weight reduction for dieters.

As a preliminary test of the claim, 150 people were randomly selected and asked what they regularly ate for breakfast and lunch. Each person was identified as either a consumer or a non-consumer of high-fibre breakfast cereal, and the number of kilojoules consumed at lunch was measured and recorded. These data are listed below. Estimate with 95% confidence the difference between the mean kilojoule intake of consumers and non-consumers, assuming that the two consumption populations are normally distributed.





Calories consumed at lunch by consumers of high-fibre cereal								
568	646	607	555	530	714	593	647	650
498	636	529	565	566	639	551	580	629
589	739	637	568	687	693	683	532	651
681	539	617	584	694	556	667	467	
540	596	633	607	566	473	649	622	

Calories consumed at lunch by non-consumers of high-fibre cereal									
705	754	740	569	593	637	563	421	514	536
819	741	688	547	723	553	733	812	580	833
706	628	539	710	730	620	664	547	624	644
509	537	725	679	701	679	625	643	566	594
613	748	711	674	672	599	655	693	709	596
582	663	607	505	685	566	466	624	518	750
601	526	816	527	800	484	462	549	554	582
608	541	426	679	663	739	603	726	623	788
787	462	773	830	369	717	646	645	747	
573	719	480	602	596	642	588	794	583	
428	754	632	765	758	663	476	490	573	

Solution

Identifying the technique

The problem objective is to compare the population of consumers of high-fibre cereal with the population of non-consumers. The data are numerical (obviously, as we have recorded real numbers). This problem objective–data type combination tells us that the parameter to be estimated is the difference between two means $\mu_1 - \mu_2$, the difference between the mean kilojoule intake of consumers (μ_1) and that of non-consumers (μ_2).

To identify the correct interval estimator, the scientist instructs the computer to output the sample variances. They are

$$s_1^2 = 4102.98 \text{ and } s_2^2 = 10669.77$$

Considering the large difference between the two sample variances, there is reason to believe that the population variances can be considered as unequal.

The two populations are normal. Thus, we use the unequal-variances confidence interval estimator:

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2, \text{d.f.}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \text{where d.f.} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\left(\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1} \right)}$$

Calculating manually

From the data we calculate the following statistics:

$$\bar{X}_1 = 604.02$$

$$\bar{X}_2 = 633.23$$

$$s_1 = 64.05$$

$$s_2 = 103.29$$





The number of degrees of freedom of the t distribution is

$$\begin{aligned} \text{d.f.} &= \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}\right)} \\ &= \frac{\left[\frac{(64.05)^2}{43} + \frac{(103.29)^2}{107}\right]^2}{\left[\frac{(64.05)^2}{43}\right] + \left[\frac{(103.29)^2}{107}\right]} = 122.6 \end{aligned}$$

which we round to 123.

Because we wish to estimate with 95% confidence, we determine

$$\alpha = 0.05$$

$$t_{\alpha/2, \text{d.f.}} = t_{0.025, 123} \approx t_{0.025, 120} = 1.980$$

Thus, the 95% confidence interval estimate of the difference in mean kilojoule intake between consumers and non-consumers is

$$\begin{aligned} (\bar{X}_1 - \bar{X}_2) &\pm t_{\alpha/2, \text{d.f.}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\ &= (604.02 - 633.23) \pm 1.980 \sqrt{\frac{(64.05)^2}{43} + \frac{(103.29)^2}{107}} \\ &= -29.21 \pm 27.65 \end{aligned}$$

The confidence limits are

$$\text{LCL} = -56.86 \text{ and UCL} = -1.56$$

Interpreting the results

The difference in mean kilojoule intake between consumers and non-consumers is estimated to fall between -56.86 and -1.56 . That is, we estimate that non-consumers of high-fibre cereal eat an average of between 2 and 57 kilojoules more than do consumers.

Using the computer

As in Example 11.1, Excel doesn't provide confidence interval estimators for $\mu_1 - \mu_2$. However, the interval estimates can be calculated using the **Estimators** workbook. The output is shown below.

Using the Estimators workbook

	A	B	C	D	E	F
1	t-Estimate of the Difference Between Two Means (Unequal-Variances)					
2						
3		Sample 1	Sample 2	Confidence Interval Estimate		
4	Sample mean	604.02	633.23	-29.21	±	27.65
5	Sample variance	4102.976	10669.77	Lower confidence limit		-56.86
6	Sample size	43	107	Upper confidence limit		-1.56
7	Confidence level	0.95				
8	Degrees of freedom	122.60				

COMMANDS

To estimate the difference between two means with unknown and unequal population variances, first calculate the sample means and sample variances of the two independent sample observations. Open the **t-Estimate_2Means (Uneq-Var)** worksheet in the **Estimators** workbook and enter the values of the sample means, sample variances and sample sizes, as well as the confidence level.

Before proceeding to another example of a confidence interval estimator of the difference between two population means, let us review how to recognise when to use the unequal-variances t -interval estimator of $\mu_1 - \mu_2$ we introduced in this section.

IN SUMMARY

Factors that identify the unequal-variances t -interval estimator of $\mu_1 - \mu_2$

- 1 *Problem objective:* to compare two populations
- 2 *Data type:* numerical (quantitative)
- 3 *Descriptive measurement:* central location
- 4 *Population variances:* unknown but not equal
- 5 *Experimental design:* independent samples
- 6 *Population distributions:* normal

11.2b Estimating $\mu_1 - \mu_2$ when σ_1^2 and σ_2^2 are unknown but equal ($\sigma_1^2 = \sigma_2^2$)

This confidence interval estimator is only valid if the population variances are not equal. If the population variances are equal, another interval estimator must be used. This estimator requires that the following occur:

- 1 The two population random variables X_1 and X_2 are normally distributed.
- 2 The two population variances are equal – that is, $\sigma_1^2 = \sigma_2^2$.
- 3 The two samples are independent.

Confidence interval estimator of $\mu_1 - \mu_2$ with σ_1^2 and σ_2^2 unknown and $\sigma_1^2 = \sigma_2^2$

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2, \text{d.f.}} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

and

$$\text{d.f.} = n_1 + n_2 - 2$$

pooled variance estimate

The weighted average of two sample variances.

The quantity s_p^2 , called the **pooled variance estimate**, is the weighted average of two sample variances. It combines both samples to produce a single estimate of the population variance. This is made possible by the requirement that $\sigma_1^2 = \sigma_2^2$ so that the common variance can be estimated by pooling the two sample variances. It makes sense to use the pooled variance estimate because, by combining both samples, we produce a better estimate. The sampling distribution is Student t , with $n_1 + n_2 - 2$ degrees of freedom. Recall from Chapter 10 that a Student t distribution can only be used if the populations are normal.

EXAMPLE 11.4

LO1 LO3

Comparing the durability of original and replacement shock absorbers: Part I

XM11-04 A nationally known manufacturer of replacement shock absorbers claims that its product lasts longer than the type of shock absorber that the car manufacturer installs. To examine this claim, researchers equipped eight cars with the original shock absorbers and another eight cars with the replacement shock absorbers. The cars were driven until the shock absorbers were no longer effective and the number of kilometres travelled before this occurred were recorded. The results are shown in the following table. Estimate with 90% confidence the difference in mean kilometres travelled until the shock absorbers fail between the two types of products. Assume that the two populations of lifetimes of shock absorbers are normally distributed.

Number of kilometres ('000)	
Original shock absorber	Replacement shock absorber
39.6	35.7
34.2	52.0
47.0	46.8
40.9	58.5
50.6	45.7
27.5	52.4
43.5	41.3
36.3	43.8

Solution**Identifying the technique**

The problem objective is to compare two populations whose data type is numerical. (We compare the kilometres driven with the two kinds of shock absorbers.) The parameter to be estimated is $\mu_1 - \mu_2$.

Calculating manually

The population variances are unknown; therefore, in order to identify the correct confidence interval estimator, we need to calculate the sample variances. From the data we calculate the following statistics:

Original shock absorber	Replacement shock absorber
$\bar{X}_1 = 39.95$	$\bar{X}_2 = 47.03$
$s_1^2 = 54.02$	$s_2^2 = 51.22$
$n_1 = 8$	$n_2 = 8$

We have $s_1^2 = 54.02$ and $s_2^2 = 51.22$. Because s_1^2 is close to s_2^2 , we can infer that the population variances may be equal. Furthermore, the two populations are normally distributed. Thus, we employ the confidence interval estimator:

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2, n_1+n_2-2} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

The pooled variance estimate is

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2} = \frac{7(54.02) + 7(51.22)}{14} = 52.62$$

The specified confidence level is

$$1 - \alpha = 0.90$$



Hence, we need to find $t_{\alpha/2, \text{d.f.}} = t_{0.05, \text{d.f.}}$ with d.f. = $n_1 + n_2 - 2 = 14$. From Table 4 in Appendix B, we find

$$t_{0.05, 14} = 1.761$$

The 90% confidence interval estimate of the difference in mean longevity between the two kinds of shock absorbers is

$$\begin{aligned} (\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2, n_1+n_2-2} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} &= (39.95 - 47.03) \pm 1.761 \sqrt{52.62 \left(\frac{1}{8} + \frac{1}{8} \right)} \\ &= -7.08 \pm 6.39 \end{aligned}$$

Thus, we find confidence limits of

$$\text{LCL} = -13.47 \text{ and UCL} = -0.69$$

Interpreting the results

The difference in mean kilometres until shock absorber failure between the two types of shock absorbers is estimated to lie between -13.47 and -0.69 thousand kilometres. That is, the replacement shock absorber is estimated to last on average between 690 and 13470 kilometres longer than the original shock absorber. The manufacturer's claim appears to be believable.

Using the computer

Excel doesn't provide confidence interval estimates for $\mu_1 - \mu_2$. However, we can calculate the interval estimates using the **Estimators** workbook. The output is shown below.

Using the Estimators workbook

	A	B	C	D	E	F
1	t-Estimate of the Difference Between Two Means (Equal-Variances)					
2						
3		Sample 1	Sample 2	Confidence Interval Estimate		
4	Sample mean	39.95	47.03	-7.08	±	6.39
5	Sample variance	4102.976	51.22	Lower confidence limit		-13.47
6	Sample size	8	8	Upper confidence limit		-0.69
7	Confidence level	0.9				
8	Pooled Variance	52.62				

COMMANDS

To estimate the difference between two means with unknown but equal population variances, open the **t-estimate_2Means (Eq-Var)** worksheet in the **Estimators** workbook. Then enter in the values of the sample means, the sample variances and the sample sizes, as well as the confidence level.

Here is a summary of how we recognise when to use the equal-variances t -interval estimator.

IN SUMMARY

Factors that identify the equal-variances t -interval estimator of $\mu_1 - \mu_2$

- 1 *Problem objective:* to compare two populations
- 2 *Data type:* numerical (quantitative)
- 3 *Descriptive measurement:* central location
- 4 *Population variance:* unknown but equal
- 5 *Experimental design:* independent samples
- 6 *Population distributions:* normally distributed

11.2c Violations of the required conditions

We have already pointed out that to ensure the validity of a statistical technique, the statistics practitioner must be certain that the technique is appropriate and that all requirements for its use are satisfied. For example, all the methods presented in this section require that the populations sampled are normally distributed. As before, we can check to see if the normality requirement is satisfied by drawing histograms of the data for the two variables X_1 and X_2 . In the inference about $\mu_1 - \mu_2$ with σ_1^2 and σ_2^2 unknown, we can only use the pooled variance estimator when $\sigma_1^2 = \sigma_2^2$. When they are equal, we calculate and use the pooled variance estimate s_p^2 .

An important principle is being applied here and will be applied again in this chapter (Section 11.4) and in later chapters. The principle can be loosely stated as: where possible, it is advantageous to pool sample data to estimate the sampling distribution standard deviation.

In the application above, we are able to pool data because we assume that the two samples were drawn from populations with a common variance. Combining the samples increases the accuracy of the estimate. Thus, s_p^2 is a better estimator of the common variance than either s_1^2 or s_2^2 .

When the two population variances are unequal, we cannot pool the data and produce a common estimator. We must calculate s_1^2 or s_2^2 and use them to estimate σ_1^2 and σ_2^2 respectively.

We will encounter applications in which at least one of the sample sizes is small (less than or equal to 200) but the number of degrees of freedom is 200 or greater. The confidence interval estimator presented here is the appropriate one to use. However, it should be noted that the value of $t_{\alpha/2}$ in such situations approximately equals $z_{\alpha/2}$. (Recall that $t \approx z$ for d.f. ≥ 200 .) The quantity $t_{\alpha/2}$ can be read from either the Normal table (Table 3 in Appendix B) or the Student t table (Table 4 in Appendix B).

It is important for you to realise that if a requirement is not satisfied, a specific technique should not be used, because the results may be invalid. Some students believe that the techniques should be used even when the required conditions are not satisfied, as there is no alternative. This is quite untrue. When the data are not normally distributed – a required condition for the use of the z - or t -statistics – we can use another technique. To draw inferences about $\mu_1 - \mu_2$ from non-normal populations, we use the nonparametric technique – the Wilcoxon rank sum test for independent samples (described in Section 20.2).

EXERCISES

Learning the techniques

*These problems can be solved manually or by using Excel's **Estimators** workbook.*

- 11.9** The following statistics were calculated from the samples of two independent normal populations:

Sample 1	$\bar{X}_1 = 115.6$	$s_1 = 11.8$	$n_1 = 42$
Sample 2	$\bar{X}_2 = 133.0$	$s_2 = 28.6$	$n_2 = 34$

Estimate with 99% confidence the difference between the two population means.

- 11.10** You are given the following summary information calculated from the samples of two independent normal populations:

Sample 1	$\bar{X}_1 = 7.63$	$s_1 = 0.79$	$n_1 = 18$
Sample 2	$\bar{X}_2 = 6.19$	$s_2 = 0.85$	$n_2 = 24$

- a** Estimate $\mu_1 - \mu_2$ with 90% confidence.
b What assumptions must you make in order to answer part (a)?

- 11.11** **XR11-11** The following random samples were drawn from two normal populations:

Sample 1	14	29	32	18	24		
Sample 2	41	36	40	27	23	32	37

Estimate the difference between their population means with 90% confidence.

- 11.12 a** Random samples of 15 observations from each of two normal populations were drawn, with the following results:

Sample 1	$\bar{X}_1 = 1.48$	$s_1 = 0.18$	$n_1 = 15$
Sample 2	$\bar{X}_2 = 1.23$	$s_2 = 0.14$	$n_2 = 15$

Estimate the difference between the two population means with 90% confidence.

- b** Repeat part (a) increasing the sample sizes to 200 ($n_1 = n_2 = 200$).

- 11.13 a** In random samples of 25 from each of two normal populations, we found the following statistics:

Sample 1	$\bar{X}_1 = 524$	$s_1 = 129$	$n_1 = 25$
Sample 2	$\bar{X}_2 = 469$	$s_2 = 131$	$n_2 = 25$

Estimate the difference between the two population means ($\mu_1 - \mu_2$) with 95% confidence.

- b** Repeat part (a) increasing the standard deviations to $s_1 = 255$ and $s_2 = 260$.
c Describe what happens when the sample standard deviations get larger.
d Repeat part (a) with samples of size 100 ($n_1 = n_2 = 100$).
e Discuss the effects of increasing the sample size.

- 11.14** Random sampling from two normal populations produced the following results:

Sample 1	$\bar{X}_1 = 63$	$s_1 = 18$	$n_1 = 50$
Sample 2	$\bar{X}_2 = 60$	$s_2 = 7$	$n_2 = 45$

- a** Estimate with 90% confidence the difference between the two population means.
b Repeat part (a) changing the sample standard deviations to 41 and 15, respectively.
c What happens when the sample standard deviations increase?
d Repeat part (a), doubling the sample sizes.
e Describe the effects of increasing the sample sizes.

Applying the techniques

- 11.15 Self-correcting exercise.** Three years ago, a 100-hectare site was planted with 250 000 pine seedlings. Half the site was scarified (soil turned up) and the seedlings spot-fertilised; the other half was not fertilised. Random samples of 50 trees

from each half of the site were taken and the foliage of each tree was weighed. Assume that the weights of the pine tree foliage from the site are normally distributed. The results are shown in the table. Estimate with 95% confidence the difference in mean foliage weight between fertilised and unfertilised trees.

Fertilised trees	Unfertilised trees
$\bar{X}_1 = 5.98$	$\bar{X}_2 = 4.79$
$s_1^2 = 0.61$	$s_2^2 = 0.77$

- 11.16** Coupons for purchasing products at discount prices periodically appear in local newspapers and advertisements. A supermarket chain has two different types of coupon for its own brand of bread. Coupon 1 offers two loaves for the price of one and coupon 2 offers a 50-cent discount on the purchase of each loaf. In order to determine the relative selling power of the two plans, the supermarket chain performs the following experiment. The coupons appear in the local newspaper and supermarket chain website on four consecutive weeks (coupon 1 in weeks 1 and 2, and coupon 2 in weeks 3 and 4). The company wants to estimate the difference in average daily sales under the two coupon plans. The average number of loaves sold per day during the first 14 days (the supermarkets are open seven days per week) was 153 with a standard deviation of 10. The average number per day during the second 14 days was 142 with a standard deviation of 10. Assuming that the population of the number of loaves sold per day is normally distributed, estimate with 99% confidence the difference in mean daily sales under the two coupon plans.

- 11.17** A sporting goods manufacturer has developed a new type of golf ball which he believes will travel further than any other type of ball currently in use. In an experiment to verify his claim, he took 100 of his golf balls and 100 golf balls of a leading competitor to a local driving range, where he asked a variety of people to hit the balls with a driver. The distances were measured (in metres), and the statistics shown in the following table were determined. Estimate with 90% confidence the difference between the mean distances of the two brands of golf balls.

New type of golf ball	Competitor's golf ball
$\bar{X}_1 = 193$	$\bar{X}_2 = 184$
$s_1 = 27$	$s_2 = 22$

Computer/manual applications

The following exercises require the use of a computer and software. The answers may be calculated manually using the sample summary information provided.

- 11.18 XR11-18** The managing director of Tastee Inc., a baby-food producer, claims that her company's product is superior to that of the leading competitor, because babies gain weight faster with her product. To test the claim a survey was undertaken. Mothers were asked which baby food they intended to feed their babies. Those who responded Tastee or the leading competitor were asked to keep track of their babies' weight gains over the next two months. There were 15 mothers who indicated that they would feed their babies Tastee and 25 who responded that they would feed their babies the leading competitor's product. Each baby's weight gain (in grams) was recorded (column 1 lists data for Tastee and column 2 for leading competitor results). The data are also shown in the accompanying table.

Weight gain (grams)

Tastee baby food							Competitor's baby food						
30	37	36	36	39	29	31	38	28	32	29	33	38	33
37	39	42	44	41	37	38		30	29	26	27	32	31
							31	29	27	26	33	31	32

- a Estimate with 95% confidence the difference between the mean weight gains of babies fed the two products.
- b Check to ensure that the required condition(s) is satisfied.

Sample statistics: $n_1 = 15$, $\bar{X}_1 = 36.93$, $s_1 = 4.23$
 $n_2 = 25$, $\bar{X}_2 = 31.36$, $s_2 = 3.35$

- 11.19 XR11-19** The data obtained from sampling two populations are recorded. (Column 1 of the data file contains the data, and column 2 specifies the sample.) Some data are shown below.

Observations	25	15	38	...	39	-3	26
Sample	1	1	1	...	2	2	2

- a Estimate the difference in population means with 95% confidence.
- b What is the required condition(s) of the techniques employed in part (a)?
- c Check to ensure that the required condition(s) is satisfied.

Sample statistics: $n_1 = 100$, $\bar{X}_1 = 19.07$, $s_1 = 9.57$
 $n_2 = 140$, $\bar{X}_2 = 16.38$, $s_2 = 25.16$

- 11.20 XR11-20** In assessing the value of radio advertisements, sponsors not only measure the total number of listeners, but also record their ages. The 18–34 age group is considered to spend the most money. To examine the issue, the manager of an FM station commissioned a survey. One objective was to measure the difference between listening habits of the 18–34 and 35–50 age groups. The survey asked 250 people in each age category how much time they spent listening to FM radio per day. The results (in minutes) were recorded. (Column 1 lists the listening times, and column 2 identifies the age group: 1 = 18–34 and 2 = 35–50.) Some data are shown below.

Listening times	75	30	50	...	50	0
Age group	1	1	1	...	2	2

- a Estimate with 95% confidence the difference in mean time spent listening to FM radio between the two age groups.
- b Are the required conditions satisfied for the techniques you used in part (a)?

Sample statistics: $n_1 = 250$, $\bar{X}_1 = 58.99$, $s_1 = 30.77$
 $n_2 = 250$, $\bar{X}_2 = 52.96$, $s_2 = 43.32$

- 11.21 XR11-21** Automobile insurance companies take many factors into consideration when setting insurance premiums. These factors include age, marital status and the number of previous accidents. An insurance company is considering introducing another factor, kilometres driven by the applicant per year. In order to determine the effect of gender on kilometres driven, 100 male and 100 female drivers were surveyed. Each driver was asked how many kilometres he or she drove in the previous year. The distances (in thousands of kilometres) are stored in stacked format (code 1 = male and code 2 = female). (A partial listing of the data is shown below.)

Kms ('000)	11.2	9.2	6.4	...	15.1	7.1
Male/female	1	1	1	...	2	2

- a Estimate with 95% confidence the difference in mean distance driven by male and female drivers.
- b Check to ensure that the required condition(s) of the techniques used in part (a) is satisfied.

Sample statistics: $n_1 = 100$, $\bar{X}_1 = 10.23$, $s_1 = 2.87$
 $n_2 = 100$, $\bar{X}_2 = 9.66$, $s_2 = 2.90$

11.22 XR11-22 A statistics lecturer needs to select a statistical software package for her course. One of the most important features, according to the lecturer, is the ease with which students learn to use the software. She has narrowed the selection to two possibilities: software A, a menu-driven statistical package with some high-powered techniques; and software B, a spreadsheet that has the capability of performing most techniques. To help make her decision, she asks 40 statistics students selected at random to choose one of the two packages. She gives each student a statistics problem to solve by computer with the appropriate manual. The amount of time (in minutes) each student needs to complete the problem is recorded and stored in unstacked format (column 1 = package A, $n_1 = 24$; and column 2 = package B, $n_2 = 16$). A partial listing of the data is provided below.

Package A	88	83	70	81	...	105	82	75
Package B	55	57	67	...	49	67		

- a Estimate with 95% confidence the difference in the mean amount of time needed to learn to use the two packages.
- b What are the required conditions for the techniques used in part (a)?
- c Check to see if the required conditions are satisfied.

Sample statistics: $n_1 = 24$, $\bar{X}_1 = 74.71$, $s_1 = 24.02$
 $n_2 = 16$, $\bar{X}_2 = 52.5$, $s_2 = 9.04$

11.23 XR11-23 One factor in low productivity is the amount of time wasted by workers. Wasted time includes time spent cleaning up mistakes, waiting for more material and equipment, and performing any other activity not related to production. In a project designed to examine the problem, an operations management consultant took a survey of 200 workers in companies that were classified (on the basis of their latest annual profits) as successful and another 200 workers from unsuccessful companies. The amount of time (in hours) wasted during a standard 40-hour work week was recorded for each worker. (Row 1 lists data for successful

companies and row 2 for unsuccessful companies.) Some data appear below.

Successful company	5.8	2.0	6.5	5.3	...	4.1	2.0	5.3
Unsuccessful company	7.6	2.7	10.1	4.1	...	5.8	8.3	0.8

Estimate with 95% confidence how much more time is wasted in unsuccessful firms than in successful ones.

Sample statistics: $n_1 = 200$, $\bar{X}_1 = 5.02$, $s_1 = 1.39$
 $n_2 = 200$, $\bar{X}_2 = 7.80$, $s_2 = 3.09$

11.24 XR11-24 The cruise ship business is rapidly increasing. Although cruises have long been associated with seniors, it now appears that younger people are choosing a cruise for their vacations. To determine whether this is true, an executive for a cruise line sampled passengers 2 years ago and this year and determined their ages. Estimate with 99% confidence the difference in ages between this year and 2 years ago.

Sample statistics: $n_1 = 125$, $\bar{X}_1 = 59.808$, $s_1 = 7.018$
 $n_2 = 159$, $\bar{X}_2 = 57.396$, $s_2 = 6.994$

11.25 XR11-25 High blood pressure (hypertension) is a leading cause of strokes. Medical researchers are constantly seeking ways to treat patients suffering from this condition. A specialist in hypertension claims that regular aerobic exercise can reduce high blood pressure just as successfully as drugs, with none of the adverse side effects. To test the claim, 50 patients who suffer from high blood pressure were chosen to participate in an experiment. For 60 days, half the sample exercised three times per week for 1 hour and did not take medication; the other half took the standard medication. The percentage reduction in blood pressure was recorded for each individual.

- a Estimate with 95% confidence the difference in mean percentage reduction in blood pressure between drugs and exercise programs.
- b Check to ensure that the required condition(s) of the technique used in part (a) is satisfied.

Sample statistics: $n_1 = 25$, $\bar{X}_1 = 13.52$, $s_1^2 = 5.76$
 $n_2 = 25$, $\bar{X}_2 = 9.92$, $s_2^2 = 13.16$

11.3 Estimating the difference between two population means with matched pairs experiments: Dependent samples

In the previous sections we dealt with statistical techniques for estimating the difference between two population means when the samples were independent. Now consider the following experiment in which the samples are not independent.

The managing director of a pharmaceutical company that has recently developed a new non-prescription sleeping pill wants to measure its effectiveness. Her assistant recruits five individuals selected at random to participate in a preliminary experiment. The experiment is performed over two nights. On one night the subject takes the sleeping pill, and on the other a placebo (a pill that looks like a sleeping pill but contains no medication). The order in which the pills are taken by each individual is random. The hours of sleep in each case are shown in the second and third columns of **Table 11.1**.

TABLE 11.1 Number of hours slept by subjects in preliminary experiment and their differences

Subject	Number of hours of sleep		
	Sleeping pill (X_1)	Placebo (X_2)	Difference ($X_D = X_1 - X_2$)
1	7.3	6.8	0.5
2	8.5	7.9	0.6
3	6.4	6.0	0.4
4	9.0	8.4	0.6
5	6.9	6.5	0.4

This type of experiment is called a **matched pairs experiment**, because the number of hours of sleep of each subject was measured twice – once with the sleeping pill and once with the placebo. As a result, there is a natural pairing between the two samples. This means that the samples are *not independent*; once we selected the five subjects to take the sleeping pill, the experiment dictated that we measure the amount of sleep of the *same five subjects* with the placebo.

If we had measured the amount of sleep of five individuals with the sleeping pill and the amount of sleep of *another five* individuals with the placebo, the samples would have been independent. If the samples were independent, we would measure the effectiveness of the sleeping pill by constructing the confidence interval for the difference of the means ($\mu_1 - \mu_2$), where μ_1 is the average amount of sleep with the sleeping pill, and μ_2 is the average amount of sleep with the placebo, using $\bar{X}_1 - \bar{X}_2$ as an estimate for $\mu_1 - \mu_2$.

In this experiment, the two samples are *not* independent, so in order to eliminate the effect of the variations in the subjects' sleeping times, we estimate the **mean of the population of differences** (which we label μ_D), as opposed to the difference between the population means. Note that $\mu_1 - \mu_2 = \mu_D$, but we estimate μ_D because of the way the experiment was performed. This estimation is done by calculating the difference X_D between the amounts of sleep with the sleeping pill and with the placebo for each of the five subjects, as shown in the last column of **Table 11.1**.

matched pairs experiment

One in which each observation from one sample can be matched with an observation in another sample.

mean of the population of differences

The mean of the paired differences in a matched pairs experiment.

11.3a Estimating the mean difference

We estimate the difference between population means by estimating the mean difference μ_D , when the data are produced by a matched pairs experiment. As in the one population case (Section 10.3), we use \bar{X}_D to estimate μ_D .

Confidence interval estimator of μ_D

$$\bar{X}_D \pm t_{\alpha/2, d.f.} \frac{s_D}{\sqrt{n_D}}, \text{ where d.f.} = n_D - 1; n_D = n_1 = n_2$$

assuming X_D is normally distributed.

EXAMPLE 11.5

L04 L05

Comparing the durability of the original replacement shock absorbers: Part II

XM11-05 A nationally known manufacturer of replacement shock absorbers would like to measure the effectiveness of its product in relation to the type of shock absorber that a car manufacturer installs. To make this comparison, eight cars each had one new original and one new replacement shock absorber installed on the rear end and were driven until the shock absorbers were no longer effective. In each case, the number of kilometres until this happened was recorded. The results are shown in the following table.

Car	Number of kilometres driven ('000)	
	Original shock absorber	Replacement shock absorber
1	42.6	43.8
2	37.2	41.3
3	50.0	49.7
4	43.9	45.7
5	53.6	52.5
6	32.5	36.8
7	46.5	47.0
8	39.3	40.7

Estimate the mean difference between kilometres driven with the original and with the replacement shock absorbers.

Solution

Identifying the technique

The data type is numerical (number of kilometres driven), and the problem objective is to compare two populations (kilometres driven with the original and with the replacement shock absorbers). The observations are paired, because each car was equipped with both kinds of shock absorbers. Hence, the parameter of interest is μ_D . We arbitrarily define X_D as the number of kilometres driven on the original shock absorber minus the number of kilometres driven on the replacement shock absorber.

Calculating manually

The values of X_D are as follows:

Car	1	2	3	4	5	6	7	8
X_D	-1.2	-4.1	0.3	-1.8	1.1	-4.3	-0.5	-1.4

From the data, we calculate

$$\bar{X}_D = -1.49 \text{ and } s_D = 1.92; \text{ d.f.} = n_D - 1 = 8 - 1 = 7$$



Therefore, the estimate for mean difference m_D with 95% confidence is

$$\bar{X}_D \pm t_{\alpha/2, n_D-1} \frac{s_D}{\sqrt{n_D}} = -1.49 \pm 2.365 \left(\frac{1.92}{\sqrt{8}} \right) \\ = -1.49 \pm 1.61$$

which simplifies to

$$(-3.10, 0.12)$$

Interpreting the results

The 95% confidence interval estimate of the mean difference between kilometres driven with the original and with the replacement shock absorbers is the interval -3.10 to 0.12 . It is worth noting that we have assumed that the values of X_D are normally distributed.

Using the computer

Excel does not provide confidence interval estimates for μ_D . However, LCL and UCL can be calculated using the **Estimators** workbook (see Example 10.2).

Using the Estimators workbook

	A	B	C	D	E
1	t-Estimate of a Mean				
2					
3	Sample mean	-1.49	Confidence Interval Estimate		
4	Sample standard deviation	1.92		-1.49	± 1.61
5	Sample size	8	Lower confidence limit		-3.10
6	Confidence level	0.95	Upper confidence limit		0.12

COMMANDS

To estimate the mean difference, open the data file (**XM11-05**) and first calculate the difference (X_D) for each observation and then calculate their sample mean and sample variance. As in Example 10.2, open the **t-Estimate_Mean** worksheet in the **Estimators** workbook and then enter in the values of the sample mean (\bar{X}_D), sample standard deviation of (s_{x_D}) and sample size (n_D), as well as the confidence level.

11.3b Recognising the matched pairs experiment

Many students of statistics experience some degree of difficulty in determining whether a particular experiment is independent or matched pairs. The key to recognising a matched pairs experiment is to watch for a natural pairing between one observation in the first sample and one observation in the second sample. If a natural pairing exists, the experiment involves matched pairs.

Here is a summary of how we recognise when to use the *t*-interval estimator for the matched pairs experiment.

IN SUMMARY

Factors that identify the *t*-interval estimator of μ_D

- 1 *Problem objective:* to compare two populations
- 2 *Data type:* numerical (quantitative)
- 3 *Descriptive measurement:* central location
- 4 *Experimental design:* dependent samples (matched pairs)
- 5 *Population distributions:* normally distributed

11.3c Violation of the required condition

The estimator of μ_D requires that the differences X_D are normally distributed. This condition will be satisfied when both X_1 and X_2 are normal or the difference $X_D = X_1 - X_2$ is normal. If the differences are very non-normal, we cannot use the *t*-test of μ_D . Instead, we could employ a nonparametric technique.

EXERCISES

Learning the techniques

- 11.26 XR11-26** You are given the following data generated from a matched pairs experiment. Assume that both populations are normally distributed. Estimate the difference between the means with 95% confidence.

Observation	I	II
1	20	17
2	23	16
3	15	9
4	18	19
5	19	15

- 11.27** For each problem, estimate μ_D with 95% confidence, assuming that both populations are normally distributed.

a $\bar{X}_D = 2 \quad s_D = 4 \quad n_D = 15$
b $\bar{X}_D = -8 \quad s_D = 20 \quad n_D = 50$

- 11.28 XR11-28** You are given the following data generated from a matched pairs experiment. Assuming that both populations are normally distributed, estimate μ_D with 90% confidence.

Observation	Sample 1	Sample 2
1	25	32
2	11	14
3	17	16
4	7	14
5	29	36
6	21	22

Applying the techniques

- 11.29 XR11-29 Self-correcting exercise.** In an attempt to estimate the difference between Queensland petrol prices in 2017 and 2019, average petrol prices for December 2017 and December 2019 in 38 localities in the state of Queensland were recorded. Assuming that petrol prices are normally

distributed, estimate the mean difference in petrol prices for the two periods, 2017 and 2019, with 99% confidence.

(Source: <https://www.racq.com.au/cars-and-driving/cars-owning-and-maintaining-a-car/fuel-prices>)

- 11.30 XR11-30** In a study to determine whether gender affects salary offers for graduating BA students, 10 pairs of students were selected. Each pair consisted of a male student and a female student who had almost identical average marks, courses taken, ages and previous work experience. The highest salary offered to each student upon graduation is shown in the following table. Assume that these salaries follow a normal distribution.

- a Estimate the average difference in salary offers with 90% confidence.
b How should the experiment be redesigned in order to have independent samples? Which design is superior? Explain.

Student pair	Annual salary offer (\$'000)	
	Female student	Male student
1	52	55
2	47	48
3	51	57
4	49	47
5	56	59
6	53	55
7	51	49
8	61	57
9	55	66
10	48	53

- 11.31 XR11-31** In an effort to determine whether a new type of fertiliser is more effective than the type currently in use, researchers took 12 two-acre plots of land scattered throughout the shire. Each plot was divided into two equal-size subplots, one of which was treated with the current fertiliser and the other with the new fertiliser. Wheat was

planted and the crop yields were measured.

The data are provided below.

Plot	Crop yield	
	Current fertiliser	New fertiliser
1	56	60
2	45	49
3	68	66
4	72	73
5	61	59
6	69	67
7	57	61
8	55	60
9	60	58
10	72	75
11	75	72
12	66	68

- a Estimate with 95% confidence the difference in mean crop yields between the two fertilisers.
- b What is the required condition(s) for the validity of the estimates obtained in part (a)?
- c Is the required condition(s) satisfied?
- d Are these data experimental or observational? Explain.
- e How should the experiment be conducted if the researchers believed that the land throughout the shire was essentially the same?

Computer/manual applications

The following exercises require the use of a computer and software. The answers may be calculated manually using the sample statistics provided.

- 11.32 XR11-32** In order to determine the effect of advertising in the local newspaper, a researcher took a sample of 40 businesses that did not advertise in the local newspaper last year but did so this year. The annual sales (in thousands of dollars)

for each business in both years were recorded. Some of these observations appear below.

Store	Sales	
	This year	Last year
1	284	227
2	338	336
3	159	45
4	219	198
:	:	:
38	206	123
39	192	222
40	239	269

- a Estimate with 90% confidence the improvement in sales between the two years.
- b Check to ensure that the required condition(s) of the techniques above is satisfied.
- c Would it be advantageous to perform this experiment with independent samples? Explain why or why not.

Sample statistics: $n_D = 40$, $\bar{X}_D = 29.625$, $s_D = 45.95$.

- 11.33 XR11-33** A restaurant consultant undertook a preliminary study to estimate the difference in tips earned by waiters and waitresses. The study involved measuring and recording the percentage of the total bill left as a tip for one randomly selected waiter and one randomly selected waitress in each of 50 restaurants during a one-week period.

- a Estimate with 95% confidence the difference in the mean tips earned by the waiters and waitresses.
- b Verify the required conditions.

Sample statistics: $\bar{X}_D = -1.16$, $s_D = 2.22$, $n_D = 50$, where $X_D = X[\text{waiter}] - X[\text{waitress}]$.

11.4 Estimating the difference between two population proportions, $p_1 - p_2$

In this section, we present the procedures for estimating the difference between population parameters whose data are nominal. When data are nominal, the only meaningful calculation is to count the number of occurrences of each type of outcome and calculate proportions. Consequently, the parameter to be estimated in this section is the difference between two population proportions, $p_1 - p_2$.

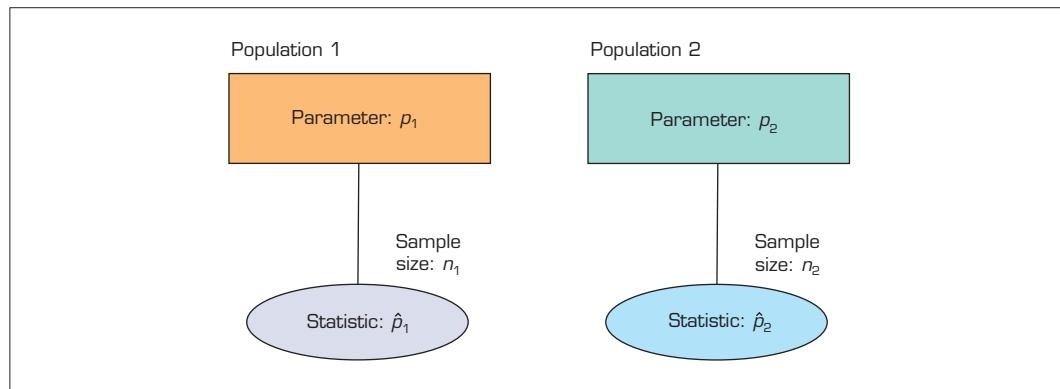
In order to draw inferences about $p_1 - p_2$, we take two independent samples – a sample of size n_1 from population 1 and a sample of size n_2 from population 2. (Figure 11.3 depicts the sampling process.) For each sample, we count the number of successes (recall that we call

anything we are looking for a success) in each sample, which we label X_1 and X_2 , respectively. The sample proportions are then calculated, as follows:

$$\hat{p}_1 = \frac{X_1}{n_1} \quad \text{and} \quad \hat{p}_2 = \frac{X_2}{n_2}$$

Statistics practitioners have proved that the statistic $\hat{p}_1 - \hat{p}_2$ is an unbiased consistent estimator of the parameter $p_1 - p_2$.

FIGURE 11.3 Sampling from two populations of nominal data



11.4a Sampling distribution of $\hat{p}_1 - \hat{p}_2$

Using the same mathematics as we used in Chapter 10 to derive the sampling distribution of the sample proportion \hat{p} , we determine the sampling distribution of the difference between two sample proportions.

Sampling distribution of $\hat{p}_1 - \hat{p}_2$

- 1 The statistic $\hat{p}_1 - \hat{p}_2$ is approximately normally distributed provided the sample sizes are large enough so that $n_1 p_1$, $n_1 q_1$, $n_2 p_2$ and $n_2 q_2$ are all greater than or equal to 5. (Since p_1 , q_1 , p_2 and q_2 are unknown, we express the sample size requirement as $n_1 \hat{p}_1, n_1 \hat{q}_1, n_2 \hat{p}_2$ and $n_2 \hat{q}_2 \geq 5$.)
- 2 The mean of $\hat{p}_1 - \hat{p}_2$ is

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

- 3 The variance of $\hat{p}_1 - \hat{p}_2$ is

$$V(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$$

where we have assumed that the samples are independent.

- 4 The standard deviation is

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

Thus, the variable

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

is approximately standard normally distributed.

11.4b Estimating the difference between two population proportions

The interval estimator of $p_1 - p_2$ can very easily be derived from the sampling distribution of $\hat{p}_1 - \hat{p}_2$.

Confidence interval estimator of $p_1 - p_2$

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

This formula is valid when $n_1 \hat{p}_1, n_1 \hat{q}_1, n_2 \hat{p}_2$ and $n_2 \hat{q}_2 \geq 5$. Notice that the standard deviation of $\hat{p}_1 - \hat{p}_2$ is estimated using $\hat{p}_1, \hat{q}_1, \hat{p}_2$ and \hat{q}_2 .

EXAMPLE 11.6

L01 L06

Popularity of Prime Minister Morrison surges

Surveys have been widely used by politicians as a way of monitoring the opinions of the electorate. The *Newspoll* published its survey results on Prime Minister Morrison's performance before and after the May 2019 Australian Federal election. The question asked was 'Are you satisfied or dissatisfied with the way Prime Minister Scott Morrison is doing his job as the Prime Minister?' Prior to the election (Survey, 14–17 May 2019, Number of respondents 1600), 46% of the surveyed voters reported that they are 'satisfied'. After the election (Survey, 25–28 July 2019, Number of respondents 1600), 51% of the surveyed voters reported that they are 'satisfied'. Estimate with 95% confidence the increase in percentage satisfaction after the election compared to before the election.

Solution

Identifying the technique

The problem objective is to compare two populations. The first is the population of voters satisfied with the Prime Minister's performance after the election and the second is the population of voters satisfied with the Prime Minister's performance before the election. The data are nominal because there are only two possible observations: 'satisfied' and 'dissatisfied'. These two factors tell us that the parameter of interest is the difference between two population proportions, $p_1 - p_2$ (where p_1 = proportion of all voters satisfied with the Prime Minister's performance after the election, and p_2 = proportion of all voters satisfied with the Prime Minister's performance before the election).

Calculating manually

The confidence interval estimator is

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

From the data, we have, $n_1 = 1600$, $n_2 = 1600$, and

$$\hat{p}_1 = 0.51, \hat{p}_2 = 0.46$$

Therefore

$$\hat{q}_1 = 1 - 0.51 = 0.49, \hat{q}_2 = 1 - 0.46 = 0.54$$





Using these values, we can verify that $n_1\hat{p}_1$, $n_1\hat{q}_1$, $n_2\hat{p}_2$ and $n_2\hat{q}_2 \geq 5$. The 95% confidence interval estimate of $p_1 - p_2$ is

$$\begin{aligned} & (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \\ & = (0.51 - 0.46) \pm 1.96 \sqrt{\frac{(0.51)(0.49)}{1600} + \frac{(0.46)(0.54)}{1600}} \\ & = 0.05 \pm 0.035 \end{aligned}$$

The confidence interval estimate is (0.015, 0.085).

Interpreting the results

We estimate that the proportion of voters who were satisfied with the Prime Minister's performance after the election is between 1.5% and 8.5% *more* than the proportion of voters who were satisfied before the election.

Using the computer

Most software packages do not conduct inferential techniques involving two proportions for p_1 and p_2 . However, the **Estimators** workbook can be used to calculate interval estimators for $p_1 - p_2$.

Excel output for Example 11.6

	A	B	C	D	E	F
1	z-Estimate of the Difference Between Two Proportions					
2						
3		Sample 1	Sample 2	Confidence Interval Estimate		
4	Sample proportion	0.5100	0.4600	0.0500	±	0.0346
5	Sample size	1600	1600	Lower confidence limit		0.0154
6	Confidence level	0.95		Upper confidence limit		0.0846

COMMANDS

To estimate the mean difference, open the **z-Estimate_2Proportions** worksheet in the **Estimators** workbook. Insert the sample proportions in cells B4 and C4. If using raw data, the two sample proportions should be calculated separately and inserted in cells B4 and C4 of the workbook. Then enter in the values of the sample sizes (n_1 and n_2), and the confidence level (0.95).

IN SUMMARY

Factors that identify the estimator of $p_1 - p_2$

- 1 *Problem objective:* to compare two populations
- 2 *Data type:* nominal (categorical)
- 3 *Experimental design:* independent samples

11.4c Selecting the sample sizes to estimate $p_1 - p_2$

The sample size required to estimate $p_1 - p_2$ is calculated in essentially the same way as the sample size needed to estimate p . First, we specify the confidence level and the error bound B . Second, we set \hat{p}_1 and \hat{p}_2 equal to 0.5 or some specific values that we believe \hat{p}_1 and \hat{p}_2 are likely to assume. Finally, we solve for the sample sizes by letting $n_1 = n_2$.

Sample sizes necessary to estimate $p_1 - p_2$

$$n_1 = n_2 = \left[\frac{z_{\alpha/2} \sqrt{\hat{p}_1 \hat{q}_1 + \hat{p}_2 \hat{q}_2}}{B} \right]^2$$

EXAMPLE 11.7

LO7

Comparing the percentage of male and female users of no-wait service centres: Part I

A market surveyor wants to estimate the difference in the proportion of male and female car owners who have their oil changed by a national chain of no-wait service centres. The surveyor wishes to estimate the difference in proportions to within 0.04, with 90% confidence. If she believes that the proportion of men who regularly use the service centre is no more than 20% and that the proportion of women who regularly use it is no more than 30%, how large should the samples be?

Solution

Identifying the technique

The data type is nominal. We want to estimate the required sample sizes when \hat{p}_1 and \hat{p}_2 are given.

Calculating manually

Because we want to estimate $p_1 - p_2$ to within 0.04, with 90% confidence,

$$B = 0.04$$

and

$$z_{\alpha/2} = 1.645$$

As p_1 is believed to be no more than 20% and p_2 no more than 30%, we use $\hat{p}_1 = 0.20$ and $\hat{p}_2 = 0.30$. Thus,

$$\begin{aligned} n_1 = n_2 &= \left[\frac{z_{\alpha/2} \sqrt{\hat{p}_1 \hat{q}_1 + \hat{p}_2 \hat{q}_2}}{B} \right]^2 \\ &= \left[\frac{1.645 \sqrt{(0.2)(0.8) + (0.3)(0.7)}}{0.04} \right]^2 \\ &= (25.02)^2 \\ &= 626 \end{aligned}$$

Interpreting the results

The surveyor must draw samples of 626 men and 626 women in order to estimate the difference in proportions to within 0.04 with 90% confidence when $\hat{p}_1 = 0.20$ and $\hat{p}_2 = 0.30$.

Using the computer

Most software packages do not conduct inferential techniques involving two proportions for p_1 and p_2 . However, the **Estimators** workbook can be used to calculate the required sample sizes, n_1 and n_2 .



Excel output for Example 11.7

	A	B	C
1	Calculating the sample sizes for a fixed width (2B) - Known phat		
2		Sample 1	Sample 2
3	phat	0.20	0.30
4	Width/2 = B	0.04	
5	Confidence level	0.90	
6	Sample size	626	626

COMMANDS

To estimate the required sample size, open the **Sample-size_2Populations** worksheet in the **Estimators** workbook. Insert the prior sample proportions in cells B4 and C4. Then enter the values of the width and confidence level (**0.90**). This will produce the required sample sizes ($n_1 = n_2$).

EXAMPLE 11.8

LO7

Comparing the percentage of male and female users of no-wait service centres: Part II

Repeat Example 11.7, but this time assume that the market surveyor has no idea about the values of \hat{p}_1 and \hat{p}_2 .

Solution

Identifying the technique

The data type is nominal. We want to estimate the required sample sizes, when we have no information given on \hat{p}_1 and \hat{p}_2 .

Calculating manually

As the surveyor has no idea about the values of \hat{p}_1 and \hat{p}_2 , she should use the values that produce the largest sample sizes, namely, $\hat{p}_1 = 0.5$ and $\hat{p}_2 = 0.5$. The result is

$$\begin{aligned} n_1 = n_2 &= \left[\frac{z_{\alpha/2} \sqrt{\hat{p}_1 \hat{q}_1 + \hat{p}_2 \hat{q}_2}}{B} \right]^2 \\ &= \left[\frac{1.645 \sqrt{(0.5)(0.5) + (0.5)(0.5)}}{0.04} \right]^2 \\ &= (29.08)^2 \\ &= 846 \end{aligned}$$

Interpreting the results

The surveyor must draw samples of 846 men and 846 women in order to estimate the difference in proportions to within 0.04 with 90% confidence. Because she has to use $\hat{p}_1 = 0.5$ and $\hat{p}_2 = 0.5$ in her preliminary calculation, she must increase each sample size by 220 (when compared to the sample sizes calculated in Example 11.7).

Using the computer

The commands are as in Example 11.6, except that the sample proportion \hat{p} is assumed to be 0.5. The output is presented below.

	A	B	C
9	Calculating the sample sizes for a fixed width (2B) - Unknown phat		
10		Sample 1	Sample 2
11	phat	0.50	0.50
12	Width/2 = B	0.04	
13	Confidence level	0.90	
14	Sample size	846	846

EXERCISES

Learning the techniques

- 11.34** Estimate $p_1 - p_2$ with 90% confidence, given the following:

$n_1 = 500$	$n_2 = 500$
$\hat{p}_1 = 0.56$	$\hat{p}_2 = 0.51$

- 11.35** A random sample of $n_1 = 200$ from population 1 produced $X_1 = 50$ successes, and a random sample of $n_2 = 100$ from population 2 produced $X_2 = 35$ successes. Estimate with 95% confidence the difference between the population proportions.
- 11.36** Random samples of 1000 from each of two populations yielded 300 successes from the first population and 200 successes from the second. Estimate the difference in population success rates between the two populations. Use a confidence level of 99%.
- 11.37** After sampling from two binomial populations, we found the following:

$n_1 = 100$	$n_2 = 100$
$\hat{p}_1 = 0.18$	$\hat{p}_2 = 0.22$

- a Estimate with 90% confidence the difference in population proportions.
- b Repeat part (a), increasing the sample proportions to 0.48 and 0.52 respectively.
- c Review the results in parts (a) and (b) and describe the effects of increasing the sample proportions.

Applying the techniques

- 11.38 Self-correcting exercise.** A market researcher employed by a chain of service centres offering no-wait oil and filter changes wants to know the difference in the fraction of male and female car owners who regularly use the service. Such information will be useful in designing advertising. In a random sample of 500 men, 42 indicated that they frequently have their cars serviced by this chain. A random sample of 300 women showed that 38 use the service. Estimate with 99% confidence the difference in the proportions of men and women who use the oil-change service.

- 11.39** An inspector for the Ministry of Gaming and Sports suspects that a particular blackjack dealer may be cheating when dealing at expensive tables. To test her belief, she observed the dealer at the \$100-limit table and noted that for 400 hands the dealer won 212 times. At the \$3000-limit table, the same dealer won 295 out of 500 deals. Estimate the difference in winning percentage between the two tables. Use a confidence level of 90%.

- 11.40** An author of statistics textbooks lives in Perth, while his publisher is located in Melbourne. Because of the amount of material sent back and forth, the speed of delivery is critical. Two couriers are regularly used. Of 53 deliveries sent through courier 1, 12 were late (delivered past the guaranteed delivery time) while of 41 deliveries by courier 2, five were late. Estimate with 90% confidence the difference in the fraction of late deliveries between the two couriers.

- 11.41** Surveys have been widely used by politicians as a way of monitoring the opinions of the electorate. Six months ago, a survey was undertaken to determine the degree of support for one of the party leaders. Of a sample of 1100, 56% indicated that they would vote for this politician. This month, another survey of 800 voters revealed that 46% now support this leader. Estimate with 95% confidence the decrease in percentage support between now and six months ago.

Computer/manual applications

The following exercises require the use of a computer and software. The answers may be calculated manually using the sample summary information provided.

- 11.42 XR11-42** A random sample of $n_1 = 1000$ from population 1 and a random sample of $n_2 = 600$ from population 2 produced the data recorded in columns 1 and 2 respectively. The results are either success (1) or failure (0). Estimate $p_1 - p_2$ with 99% confidence, where p_1 and p_2 are the proportions of successes.

Sample statistics: $n_1(0) = 301$; $n_1(1) = 699$; $n_2(0) = 156$; $n_2(1) = 444$.

11.43 XR11-43 The data stored in columns 1 and 2 respectively were drawn from random samples from two populations of nominal data for which 1 = success and 0 = failure. Estimate $p_1 - p_2$ with 95% confidence, where p_1 and p_2 are the proportions of successes.

Sample statistics: $n_1(0) = 268$; $n_1(1) = 232$; $n_2(0) = 311$; $n_2(1) = 189$.

11.44 XR11-44 An insurance company manager is thinking about offering discounts on life insurance policies to non-smokers. As part of the analysis, the manager randomly selects 200 men who are 60 years old and asks them if they smoke at least one packet of cigarettes per day and if they have ever suffered from heart disease. The results are recorded using the following format:

Column 1: Smokers (1 = suffer from heart disease; 0 = do not suffer from heart disease)
 Column 2: Non-smokers (1 = suffer from heart disease; 0 = do not suffer from heart disease)

Estimate with 90% confidence the difference between smokers and non-smokers in the fraction of men suffering from heart disease.

Sample statistics: $n_1(0) = 37$; $n_1(1) = 19$; $n_2(0) = 119$; $n_2(1) = 25$.

11.45 XR11-45 A market researcher employed by a chain of service centres offering no-wait oil and filter changes wants to know whether men and women differ in their use of the company's services. Such information would be useful in designing advertising. A random sample of 500 people was selected, and each was asked whether they have their oil and filters changed at the no-wait service centre. The responses and the gender of the respondents were recorded in the following way.

Column 1: Female (1 = use no-wait service; 0 = do not use no-wait service)

Column 2: Male (1 = use no-wait service; 0 = do not use no-wait service)

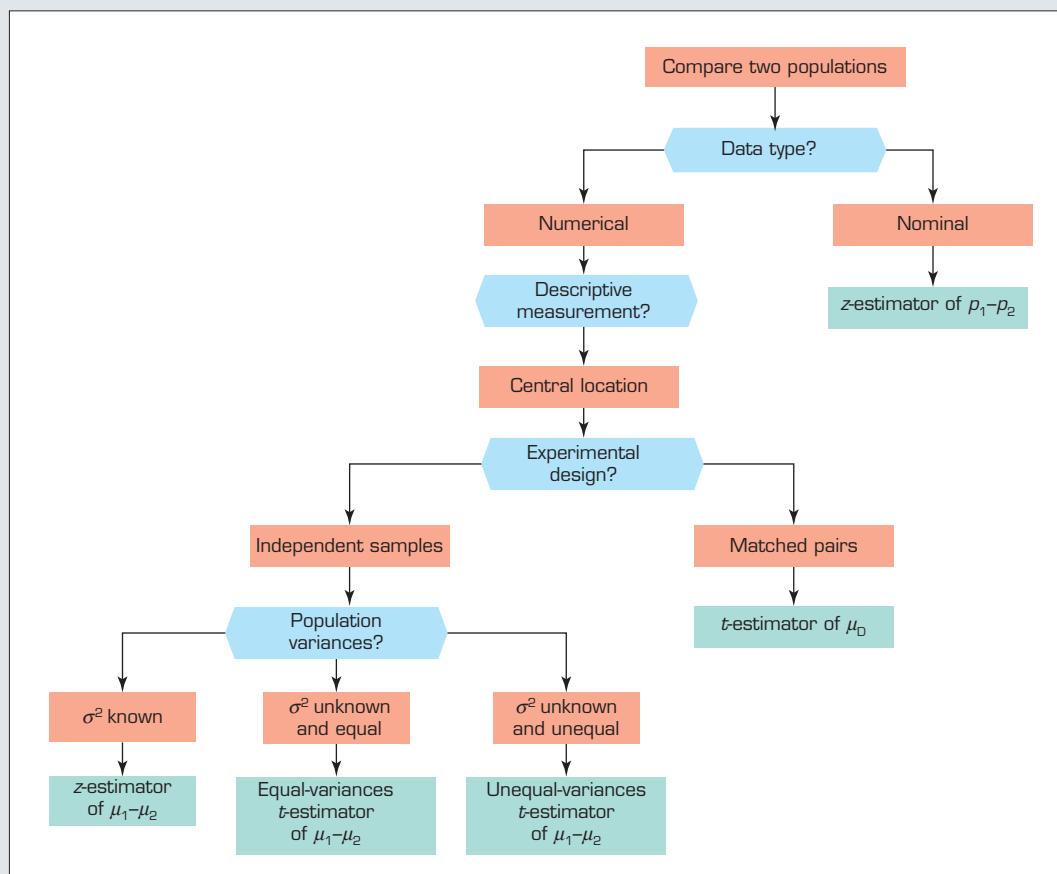
Estimate with 95% confidence the difference between the responses of men and women in their use of this oil-change service.

Sample statistics: $n_1(0) = 171$; $n_1(1) = 67$; $n_2(0) = 176$; $n_2(1) = 86$.

Study Tools

CHAPTER SUMMARY

The statistical techniques used to estimate the *difference between two population means* and the *difference between two population proportions* were described in this chapter. For the numerical data type, when the two population variances are known and when the two populations are normal or the two sample sizes are each larger than 30 and the samples are independent, the *z* distribution is used; when the two populations are normal and their variances are unknown and estimated by the sample variances, and two independent or dependent and matched pairs samples are drawn, the *t* distribution is used. For the nominal data type, when the two sample sizes are large, the *z* distribution is used. The confidence interval estimator formulas are summarised in **Table 11.2**.



The data files for Examples, Exercises and Cases are provided on the companion website (accessible through <https://login.cengagebrain.com/>).

SUMMARY OF FORMULAS

TABLE 11.2 Summary of interval estimators of $\mu_1 - \mu_2$ and $p_1 - p_2$

Parameter	Confidence interval estimator	Required conditions
$\mu_1 - \mu_2$ (numerical data)	$(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	σ_1^2 and σ_2^2 are known; X_1 and X_2 are normally distributed or n_1 and n_2 are large; samples are independent

TABLE 11.2 Continued

Parameter	Confidence interval estimator	Required conditions
$\mu_1 - \mu_2$ (numerical data)	$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2, d.f.} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ d.f. = $\frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\left(\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}\right)}$	σ_1^2 and σ_2^2 are unknown and $\sigma_1^2 \neq \sigma_2^2$; X_1 and X_2 are normally distributed; samples are independent
$\mu_1 - \mu_2$ (numerical data)	$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2, n_1+n_2-2} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$	σ_1^2 and σ_2^2 are unknown and $\sigma_1^2 = \sigma_2^2$; X_1 and X_2 are normally distributed; samples are independent
$\mu_1 - \mu_2$ (numerical data)	$\bar{X}_D \pm t_{\alpha/2, d.f.} \frac{s_D}{\sqrt{n_D}}$ d.f. = $n_D - 1$; $n_D = n_1 = n_2$	$X_D = X_1 - X_2$ is normally distributed; samples are not independent; samples are matched pairs
$p_1 - p_2$ (numerical data)	$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$	$n_1 \hat{p}_1, n_1 \hat{q}_1, n_2 \hat{p}_2$ and $n_2 \hat{q}_2 \geq 5$; samples are independent

SUPPLEMENTARY EXERCISES

- 11.46** A major car manufacturer has plants in a number of countries. To examine productivity in South Korea and Taiwan, a statistics practitioner counted the daily output for one randomly selected plant in South Korea and a similar-sized plant in Taiwan. The mean and the standard deviation of the daily number of cars produced in 15 days were calculated. These statistics are shown in the following table.

Daily car production

South Korea	Taiwan
$\bar{X}_1 = 325$	$\bar{X}_2 = 345$
$s_1 = 12$	$s_2 = 15$

Estimate the difference in the mean number of cars produced by the plants in South Korea and Taiwan. Use a confidence level of 99%.

- 11.47** Use the results from Exercise 11.46 to estimate a 99% confidence interval estimate for the difference in annual production. Assume 250 working days per year.

- 11.48** Doctors have been encouraging their patients to stop smoking for many years. In a study to determine who smokes regularly, a random sample of 1000 people were interviewed. Each person was asked whether or not they smoke regularly, as well as a variety of questions relating to demographic

characteristics (such as age and education). It was found that 248 of the respondents were university graduates. Of these, 52 smoked regularly. Of the remaining 752 respondents, 226 smoked regularly. Estimate with 90% confidence the difference in the fraction of smokers between university graduates and non-graduates.

- 11.49 XR11-49** The impact of the accumulation of carbon dioxide in the atmosphere caused by burning of fossil fuels such as oil, coal and natural gas has been hotly debated for more than a decade. Some environmentalists and scientists have predicted that the excess carbon dioxide will increase the Earth's temperature over the next 50–100 years with disastrous consequences.

To gauge the public's opinion on the subject, a random sample of 400 people had been asked two years ago whether they believed in the greenhouse effect. This year, 500 people were asked the same question. The results are recorded using the following codes: 2 = believe in greenhouse effect; 1 = do not believe in greenhouse effect. Estimate the real change in the public's opinion about the subject, using a 90% confidence level.

Sample statistics: $n_1(1) = 152$; $n_1(2) = 248$; $n_2(1) = 240$; $n_2(2) = 260$.

Case Studies

CASE 11.1 Has demand for print newspapers declined in Australia?

C11-01 Despite the steady growth in online news services, an article in *The Australian* claims that the Australian newspaper market is holding its ground. The table below presents the circulation figures for a number of leading Australian newspapers for 2010 and 2018. Estimate at the 95% confidence level the average difference in Australian newspaper circulation between 2010 and 2018.

Newspaper	Circulation figures ('000)	
	2010	2018
<i>Advertiser</i> (Mon–Fri)	80.8	112.1
<i>Australian Financial Review</i>	75.3	39.8
<i>Canberra Times</i>	32.1	13.8
<i>Courier Mail</i> (Mon–Fri)	206.1	135.0
<i>Daily Telegraph</i> (Mon–Fri)	363.4	221.6
<i>Herald Sun (VIC)</i> (Mon–Fri)	500.8	303.1
<i>NT News</i> (Mon–Fri)	21.1	11.3
<i>Sydney Morning Herald</i> (Mon–Fri)	204.4	78.8
<i>The Age</i> (Mon–Fri)	190.1	74.4
<i>The Australian</i> (Mon–Fri)	136.3	88.6
<i>The Mercury</i>	44.2	28.3
<i>West Australian</i> (Mon–Fri)	192.2	128.4

Source: en.wikipedia.org/wiki/List_of_newspapers_in_Australia_by_circulation. CC BY-SA 3.0 <https://creativecommons.org/licenses/by-sa/3.0/>.

CASE 11.2 Hotel room prices in Australia: Are they becoming cheaper?

C11-02 Australian news media claim that the credit crisis pain being felt by the hotel industry has triggered major reductions in hotel room rates, with discounts of 30% being offered on rooms in Australian cities, where guests are paying as little as \$112 per night. The following table gives average hotel room rates at various locations across Australia in 2010, 2015 and 2019.

Construct 95% confidence interval estimates for the difference in the average room rates in Australian hotels between 2010 and 2019, and between 2015 and 2019.

Average price per room per night in 2010, 2015 and 2019, Australia

Location	Average room rates (\$)		
	2010	2015	2019
Adelaide	140	143	154
Brisbane	154	168	158
Cairns	116	129	173
Canberra	149	248	176
Darwin	139	165	150
Gold Coast	132	151	197

Location	Average room rates (\$)		
	2010	2015	2019
Hobart	124	150	184
Melbourne	165	188	200
Perth	159	178	273
Sydney	170	204	256

Source: Hotel Futures 2019 – Tourism Australia Accommodation, www.dransfield.com.au, www.tourismaccommodation.com.au.

CASE 11.3 Comparing hotel room prices in New Zealand

C11-03 Consider Case 11.2 for Australia. Data for hotel room rates (in New Zealand dollars) at various locations across New Zealand in 2012 and 2020 are recorded. Present a 99% confidence interval estimate for the difference in the average room rates in New Zealand hotels between 2012 and 2020.

Average price (NZ\$) per room per night in 2012 and 2020, New Zealand

Location	2012	2020	Location	2012	2020
Auckland	137	325	Nelson	142	160
Blenheim	173	255	New Plymouth	136	155
Christchurch	154	315	Paihia	159	183
Dunedin	134	173	Palmerston North	109	165
Fox Glacier	166	165	Queenstown	177	350
Franz Josef Glacier	167	176	Rotorua	124	171
Gisborne	139	155	Taupo	158	170
Hamilton	127	185	Tauranga	164	189
Invercargill	115	185	Te Anau	138	198
Lake Tekapo	189	250	Wellington	138	175
Napier	154	166			

Source: www.hotels.com

CASE 11.4 Comparing salary offers for finance and marketing major graduates

C11-04 A number of web-based companies offer job placement services. The manager of one such company wanted to investigate the job offers recent business graduates were obtaining. In particular, she wanted to know whether finance majors were being offered higher salaries than marketing majors. In a preliminary study, she randomly sampled 50 recent graduates, half of whom majored in finance and half in marketing. From each she obtained the initial salary offer (including benefits). Estimate with 95% confidence the difference in average salaries between finance and marketing graduates.

CASE 11.5 Estimating the cost of a life saved

Two prescription medications are commonly used to treat a heart attack. Streptokinase, which has been available since 1959, costs about \$500 per dose. The second medication is tPA, a genetically engineered product that sells for about \$3000 per dose. Both streptokinase and tPA work by opening the arteries and dissolving blood clots, which are the cause of heart attacks. Several previous studies have failed to reveal any differences between the effects of the two medications. Consequently, in many countries where health care is funded by governments, doctors are required to use the less expensive streptokinase. However, the maker of tPA, Genentech Inc., contended that in the earlier studies showing no difference between the two medications, tPA was not used in the right way. Genentech decided to sponsor a more thorough experiment. The experiment was organised in 15 countries and involved a total of 41 000 patients. In this study, tPA was given to patients within 90 minutes, instead of within three hours as in previous trials. Half of the sample of 41 000 patients was treated by a rapid injection of tPA (with intravenous heparin), while the other half received streptokinase (with intravenous heparin). The number of deaths in each sample was recorded. A total of 1497 patients treated with streptokinase died, while 1292 patients who received tPA died. Estimate the cost per life saved by using tPA instead of streptokinase.

Hypothesis testing: Single population

Learning objectives

This chapter presents a systematic, step-by-step approach to testing hypotheses regarding population parameters in decision making for single populations.

At the completion of this chapter, you should be able to:

- L01** understand the fundamental concepts of hypothesis testing
- L02** set up the null and alternative hypotheses, and be familiar with the steps involved in hypothesis testing
- L03** test hypotheses regarding the population mean when the population variance is known
- L04** test hypotheses regarding the population mean when the population variance is unknown
- L05** understand the *p*-value approach to testing hypotheses and calculate the *p*-value of a test
- L06** interpret the results of a test of hypothesis
- L07** calculate the probability of a Type II error and interpret the results
- L08** test hypotheses regarding the population proportion
- L09** understand the consequences of the violation of the required conditions of each test.

CHAPTER OUTLINE

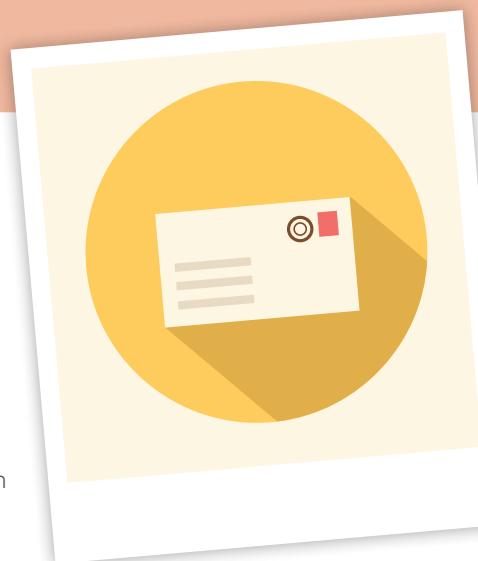
- Introduction
- 12.1** Concepts of hypothesis testing
- 12.2** Testing the population mean when the variance σ^2 is known
- 12.3** The *p*-value of a test of hypothesis
- 12.4** Testing the population mean when the variance σ^2 is unknown
- 12.5** Calculating the probability of a Type II error
- 12.6** Testing the population proportion

SPOTLIGHT ON STATISTICS

SSA envelope plan

Although many customers pay their invoices on time, many companies face the problem that a significant proportion of their customers do not. To encourage customers to pay their invoices on time, companies offer various incentives to their customers.

An express courier service normally sends invoices to customers requesting payment within 30 days. The bill lists an address, and customers are expected to use their own envelopes to return their payments. Currently the mean and standard deviation of the amount of time taken to pay bills are 24 days and 6 days



Source: Shutterstock.com/edel

respectively. The chief financial officer (CFO) believes that including a stamped self-addressed (SSA) envelope would decrease the number of bills paid late. She calculates that the improved cash flow from a two-day decrease in the payment period would pay for the costs of the envelopes and stamps. Any further decrease in the payment period would generate a profit. To test her belief, she randomly selects 220 customers and includes an SSA envelope with their invoices. The numbers of days until payment is received are recorded and stored in file Ch12:

XM12-00. A partial listing of the data appears below. Can the CFO conclude that the plan will be profitable?

Number of days until payment

27 24 14 39 ... 9 14 21 28

After we have introduced the required tools, we will answer this question (see pages 496–9).

Introduction

In Chapters 10 and 11, we introduced estimation and showed how it is used. Now we present the second general procedure of making inferences about a population – hypothesis testing. The purpose of this type of inference is to determine whether enough statistical evidence exists to enable us to conclude that a belief or **hypothesis** about a parameter is supported by the data. You will discover that hypothesis testing has a wide variety of applications in business and economics, as well as many other fields. This chapter will lay the foundation upon which the rest of the book is based. As such, it represents a critical contribution to your development as a statistics practitioner. In Section 12.1 we will introduce the concepts of hypothesis testing, and in Section 12.2 we will develop the method used to test a hypothesis about a population mean when the data type is numerical and the population standard deviation is known. We extend the analysis to real situations for which the population standard deviation is usually unknown in Section 12.4, and in Section 12.6, we deal with hypothesis testing when the data type is nominal. The rest of the chapter deals with related topics.

Examples of hypothesis testing include the following:

- 1 Suppose that a firm that produces agricultural products has developed a new fertiliser. To determine if it improves crop yields, researchers will use it to fertilise a random sample of farms. The resultant crop yields can be measured. The data type is numerical. The parameter of interest here is the mean crop yield μ . The hypothesis to test is that there is any change in the average crop yield based on the sample information.
- 2 Suppose that a company has developed a new product that it hopes will be very successful. After a complete financial analysis, the company's directors have determined that if more than 10% of potential customers buy the product, the company will make a profit. A random sample of potential customers is asked whether they would buy the product. (The sampling procedure and data-collection methods used are as described in Chapter 2.) Statistical techniques can then convert this raw data into information that would permit the company's directors to decide whether to proceed. The data type here is nominal (to buy or not to buy). The parameter is the proportion of customers p who would buy the product. The hypothesis to test is that the proportion is greater than 10% based on the sample information.

hypothesis

A proposition or conjecture that the statistician will test by a means called hypothesis testing.

12.1 Concepts of hypothesis testing

The term 'hypothesis testing' is likely to be new to most readers, but the concepts underlying hypothesis testing are probably familiar. There are a variety of non-statistical applications of hypothesis testing, the best known of which is a criminal trial.

null hypothesis

The proposition about which a decision is to be made in testing a hypothesis, denoted H_0 .

alternative (or research) hypothesis

The proposition, denoted H_A , that will be accepted if the null hypothesis H_0 is rejected.

When a person is accused of a crime, he or she faces a trial. The prosecution presents its case and a jury must make a decision on the basis of the evidence presented. In fact, the jury conducts a test of hypothesis. There are actually two hypotheses that are tested. The first is called the **null hypothesis** and is represented by H_0 (pronounced *H-nought*). The second is called the **alternative or research hypothesis** and is denoted H_A .

In a criminal trial these are

H_0 : The defendant is innocent.

H_A : The defendant is guilty.

12.1a Components of the tests

The tests of hypothesis that we present in this chapter (and in all others) are called parametric tests because they test the value of a population parameter. These tests consist of five components:

- 1 Null hypothesis
- 2 Alternative hypothesis
- 3 Test statistic
- 4 Rejection region
- 5 Decision rule

Null hypothesis

The null hypothesis, H_0 , always specifies one single value for the population parameter being studied. For example, if we wish to test whether the mean weight loss of people who participate in a new weight-reduction program is 10 kg, we would test

$$H_0: \mu = 10$$

To test whether the proportion of defective shoes coming off a production line is equal to 3%, we would test

$$H_0: p = 0.03$$

Alternative hypothesis

The alternative hypothesis, H_A , is really the more important one, because it is the hypothesis that answers our question. The alternative hypothesis can assume three possible forms: ‘greater than’, ‘less than’ or ‘not equal to’ the value shown in the null hypothesis. For example:

A Numerical data

- 1 If a tyre company wanted to know whether the average life of its new radial tyre exceeds its advertised value of 50 000 kilometres, the company would specify the alternative hypothesis as

$$H_A: \mu > 50\,000$$

- 2 If the company wanted to know whether the average life of the tyre is less than 50 000 kilometres, it would test

$$H_A: \mu < 50\,000$$

- 3 If the company wanted to determine whether the average life of the tyre differs from the advertised value, its alternative hypothesis would be

$$H_A: \mu \neq 50\,000$$

In all three cases, the null hypothesis would be

$$H_0: \mu = 50\,000$$

B Nominal data

- 1 If a company wants to know whether its 10% market share has increased as a result of a new advertising campaign, it would specify the alternative hypothesis as

$$H_A: p > 0.10$$

- 2 If it wanted to know whether the campaign has decreased its market share, it would test

$$H_A: p < 0.10$$

- 3 If it wished to determine whether its market share had changed at all, the alternative hypothesis would be

$$H_A: p \neq 0.10$$

In all three cases, the null hypothesis would be

$$H_0: p = 0.10$$

The crucial points to remember about the two hypotheses are summarised in the following box.

Null hypothesis

The null hypothesis H_0 must specify that the population parameter is equal to a single value.

Alternative hypothesis

The alternative hypothesis H_A answers the question by specifying that the parameter is one of the following:

- 1 Greater than the value shown in the null hypothesis
- 2 Less than the value shown in the null hypothesis
- 3 Different from the value shown in the null hypothesis

Test statistic

The purpose of the test is to determine whether it is appropriate to reject or not reject the null hypothesis. (As explained above, we use the term *not reject* instead of *accept*, because the latter can lead to an erroneous conclusion.)

The **test statistic** is based on the point estimator of the parameter to be tested. For example, as sample mean \bar{X} is a point estimator of the population mean μ , \bar{X} will be used as the test statistic to test the hypothesis about the population mean μ . When the population variance σ^2 is known, the test statistic is either the sample mean \bar{X} or its standardised value

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

The variable Z has a standard normal distribution if either the population is normal or n is large (see Section 10.2).

When the population variance σ^2 is unknown, we replace σ by the sample standard deviation s and the standardised test statistic is

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

The variable t has a Student t distribution with $n - 1$ degrees of freedom when the population is normally distributed (see Section 10.3).

To test a population proportion, the test statistic is either the sample proportion \hat{p} or its standardised value

$$Z = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

test statistic

The statistic used to decide whether or not to reject the null hypothesis.

This Z variable has a standard normal distribution when $np \geq 5$ and $nq \geq 5$ (see Section 10.4).

As we present the test for each parameter, we will state the test statistic.

Rejection region

rejection region

The range of values of the test statistic that would cause us to reject the null hypothesis.

critical region

Another name for the rejection region.

acceptance region

Values of the test statistic for which the null hypothesis H_0 is not rejected.

critical value

Value that separates the acceptance and rejection regions.

decision rule

A statement specifying the condition for the rejection of H_0 .

Type I error

The act of rejecting a null hypothesis when it is true.

Type II error

The act of not rejecting a null hypothesis when it is false.

The **rejection region** of a test, also called the **critical region**, consists of all values of the test statistic for which H_0 is rejected. The **acceptance region** consists of all values of the test statistic for which H_0 is not rejected. The **critical value** is the value that separates the critical region from the acceptance region.

Decision rule

The **decision rule** defines the range of values of the test statistic for which the null hypothesis H_0 is rejected in favour of H_A .

To illustrate, suppose that we wish to test

$$H_0: \mu = 1000$$

If we find that the sample mean \bar{X} (which is the test statistic) is quite different from 1000, we say that \bar{X} falls into the rejection region, and we reject the null hypothesis. On the other hand, if \bar{X} is close to 1000, we cannot reject the null hypothesis. The key question answered by the rejection region is: When is the value of the test statistic sufficiently different from the hypothesised value of the parameter to enable us to reject the null hypothesis? The process we use in answering this question depends on the probability of our making a mistake when testing the hypothesis.

As indicated in **Figure 12.1**, the null hypothesis is either true or false, and we must decide either to reject or not to reject the null hypothesis. Therefore, two correct decisions are possible: not rejecting the null hypothesis when it is true, and rejecting the null hypothesis when it is false. Conversely, two incorrect decisions are possible: rejecting H_0 when it is true (this is called a **Type I error**), and not rejecting H_0 when it is false (this is called a **Type II error**).

Because the conclusion we draw is based on sample data, the chance of making one of the two possible errors will always exist. We define

$$P(\text{Making a Type I error}) = \alpha \text{ (Greek letter alpha)}$$

and

$$P(\text{Making a Type II error}) = \beta \text{ (Greek letter beta)}$$

FIGURE 12.1 Results of a test of hypothesis

		H_0 is true	H_0 is false
	Do not reject H_0	Correct decision	Type II error $P(\text{Type II error}) = \beta$
	Reject H_0	Type I error $P(\text{Type I error}) = \alpha$	Correct decision

12.1b Level of significance

The decision rule is based on specifying the value of α , which is also called the **significance level**. We would like for both α and β to be as small as possible, but, unfortunately, there is an inverse relationship between α and β . Thus, for a given sample size, any attempt to reduce one will increase the other (see Section 12.5). The value of α is selected by the decision maker and is usually between 1% and 10%.

The critical concepts in hypothesis testing are as follows:

- 1 There are two hypotheses. One is called the null hypothesis, and the other the alternative or research hypothesis.

- 2 The testing procedure begins with the assumption that the null hypothesis is true.
- 3 The goal of the process is to determine whether there is enough evidence to infer that the alternative hypothesis is true.
- 4 There are two possible decisions:
 - a Conclude that there is enough evidence to support the alternative hypothesis, or
 - b Conclude that there is not enough evidence to support the alternative hypothesis.
- 5 Two possible errors can be made in any test. A Type I error occurs when we reject a true null hypothesis, and a Type II error occurs when we don't reject a false null hypothesis. The probabilities of Type I and Type II errors are

$$P(\text{Type I error}) = \alpha$$

$$P(\text{Type II error}) = \beta$$

If you understand the concepts of hypothesis testing, you will likely find numerous applications in your own life. Just think about the kinds of decisions you commonly make. Most involve a judgement based on a limited amount of information. A misjudgement may lead to one of the two types of error.

12.1c Applications of hypothesis testing in other disciplines

You will find as you progress through this book that a great many statistical techniques involve hypothesis testing. We believe that hypothesis testing is central to the entire subject of statistics, and that it is a vital concept in fields outside statistics. In this subsection, we demonstrate the application of hypothesis testing in two such fields.

Criminal trials

Returning to the criminal trial example, the null and alternative hypotheses:

H_0 : The defendant is innocent.

H_A : The defendant is guilty.

The jury does not know the truth about the hypotheses. They must, however, make a decision based on the evidence presented to them. After the prosecution and defence present their arguments, the jury must decide whether there is enough evidence to support the alternative hypothesis. If so, the defendant is found guilty and sentenced. If the jury finds that there is not enough evidence to support the alternative hypothesis, they render a verdict of not guilty. Notice that they do not conclude that the defendant is innocent – merely not guilty. Our justice system does not allow for a conclusion that the defendant is innocent. Just as in statistical hypothesis testing, juries can make mistakes. A Type I error occurs when an innocent person is found guilty, and a Type II error occurs when a guilty person is acquitted. In the Australian justice system, we place the burden of proof on the prosecution. In theory, this decreases the probability of Type I errors and increases the probability of Type II errors. Our society regards the conviction of innocent people to be a greater sin than the acquittal of guilty people. A Supreme Court judge once said that it is better to acquit 100 guilty persons than to convict one innocent person. In statistical terms, the judge said that the probability of a Type II error should be 100 times the probability of a Type I error.

Drug testing

Every year, dozens of new drugs are developed. It is the responsibility of the Therapeutic Goods Administration (TGA) in Australia to judge the safety and effectiveness of the drugs before allowing them to be sold to the public. This, too, is an example of hypothesis testing. The null and alternative hypotheses are as follows:

H_0 : The drug is not safe and effective.

H_A : The drug is safe and effective.

In most cases, the regulators do not know with certainty what will happen when and if the drug is approved for sale. They must make a decision based on ‘sample’ information. They will usually examine the experiments performed on laboratory animals as well as the human testing before making a decision.

Two types of error are possible. When the TGA approves a drug that turns out to be either unsafe or ineffective, a Type I error is committed. If the TGA disapproves a drug that is actually safe and effective, a Type II error is committed. Unfortunately, it is difficult to decide which error is more serious. Type I errors may lead to deaths through the use of unsafe and/or unnecessary drugs. But so too can Type II errors. By denying the public the use of a drug that works, regulators may be responsible for preventable deaths.

If you understand the concepts of hypothesis testing, you will likely find numerous non-statistical applications in your own life. Just think about the kinds of decisions you commonly make. Most involve a judgement based on a limited amount of information. A misjudgement may lead to one of the two types of error.

12.1d When do we conduct one- and two-tail tests?

As discussed earlier about developing an alternative hypothesis, there are three types of alternatives: the first is when the parameter is not equal to a specified value, the second is when the parameter is tested to be less than a specified value and the third is when the parameter is tested to be greater than a specified value.

Two-tail test

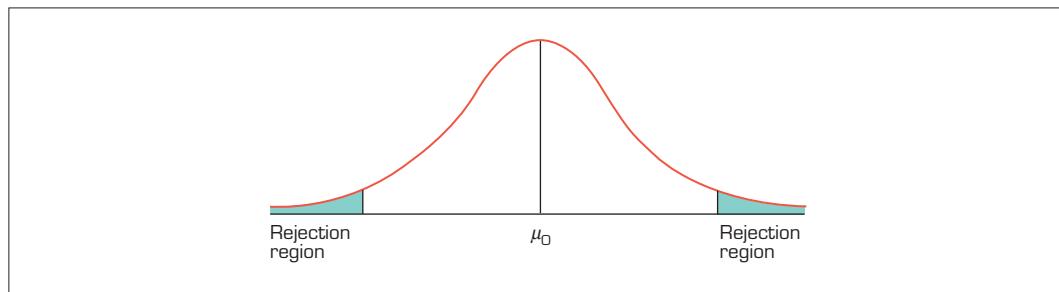
two-tail test

A test with the rejection region in both tails of the distribution, typically split evenly.

In general, a **two-tail test** is conducted whenever the alternative hypothesis specifies that the mean is not *equal* to the value stated in the null hypothesis; that is, when the hypotheses assume the following form:

$$H_0: \mu = \mu_0$$

$$H_A: \mu \neq \mu_0$$



When the sample mean \bar{X} is much smaller than μ_0 or much larger than μ_0 , we have support for H_A . Hence we have two rejection regions, one in each tail.

One-tail test

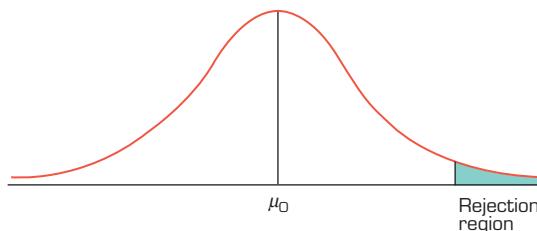
one-tail test

A test with the rejection region in only one tail of the distribution.

There are two one-tail tests. We conduct a **one-tail test** that focuses on the right tail of the sampling distribution whenever we want to know whether there is enough evidence to infer that the mean is greater than the quantity specified by the null hypothesis; that is, when the hypotheses are

$$H_0: \mu = \mu_0$$

$$H_A: \mu > \mu_0$$

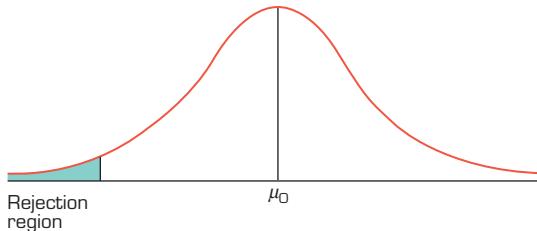


Only when the sample mean \bar{X} is much larger than μ_0 do we have support for H_A . Hence we have the rejection region on the right tail.

The second one-tail test involves the left tail of the sampling distribution. It is used when we want to determine whether there is enough evidence to infer that the mean is less than the value of the mean stated in the null hypothesis. The resulting hypotheses appear in this form:

$$H_0: \mu = \mu_0$$

$$H_A: \mu < \mu_0$$



Only when sample mean \bar{X} is much smaller than μ_0 do we have support for H_A . Hence we have the rejection region on the left tail.

The techniques introduced in this chapter and in Chapters 13–16 require you to decide which of the three forms of the test to employ. Your decision should be made following a six-step process.

12.1e Six-step process for testing hypotheses

In the same way that we illustrated how confidence interval estimates are produced and interpreted, we will illustrate how tests of hypotheses are conducted by testing the population mean when the population variance is known. As you will discover, almost all tests are conducted in the same way. We begin by identifying the technique, which usually involves recognising the parameter to be tested. This is followed by specifying the null and alternative hypotheses. Next comes the test statistic and the rejection region. Finally, we calculate (or let the computer calculate) the value of the test statistic, make a decision and answer the question posed in the problem.

To illustrate, suppose that we want to test

$$H_0: \mu = 50$$

$$H_A: \mu \neq 50$$

If we assume that the population variance is known, the test statistic is

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

We want to reject the null hypothesis whenever the test statistic is a large positive or a large negative number. The value of the level of significance α selected determines what is considered ‘large’. **Figure 12.2** depicts the sampling distribution of the test statistic. If we set $\alpha = 0.05$, we want the total area of the rejection region to equal 0.05. Since we will reject the null hypothesis when Z is either too *large positive* or too *large negative*, our decision rule is

$$\text{Reject } H_0 \text{ if } Z > 1.96 \text{ or if } Z < -1.96$$

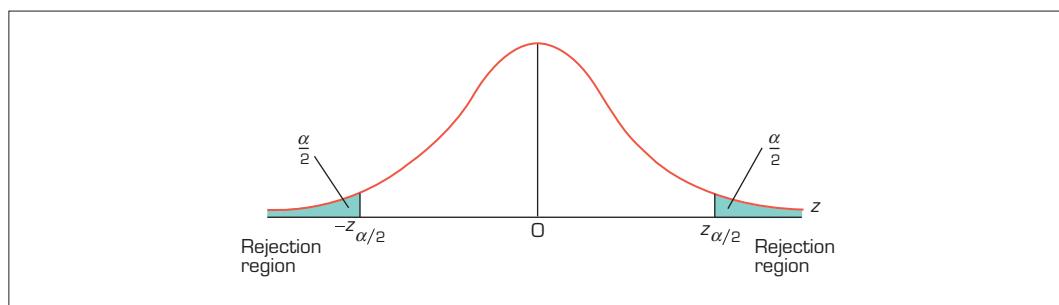
The values -1.96 and $+1.96$ are the *critical values* for this test, and the test is called a *two-tail test* because we will reject the null hypothesis if the test statistic lies in either of the two tails of the sampling distribution. For any value of α , the decision rule is

$$\text{Reject } H_0 \text{ if } Z > z_{\alpha/2} \text{ or if } Z < -z_{\alpha/2}$$

Another way of expressing this is to state

$$\text{Reject } H_0 \text{ if } |Z| > z_{\alpha/2}$$

FIGURE 12.2 Sampling distribution of the test statistic $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$



We use a *one-tail test* if the alternative hypothesis states that the parameter is either greater than or less than the value shown in the null hypothesis. For instance, if we test

$$H_0: \mu = 1000$$

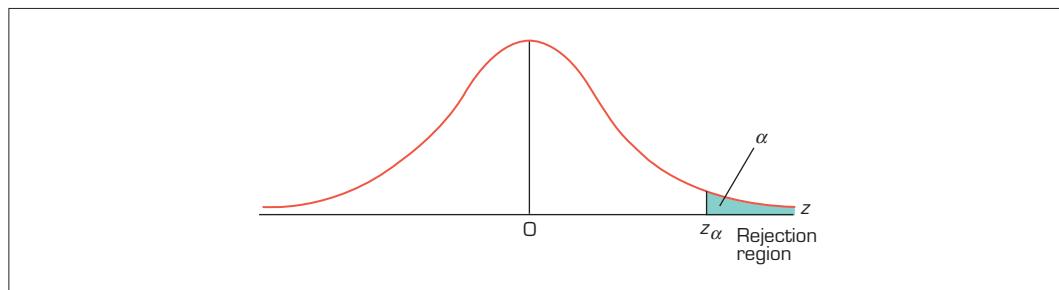
$$H_A: \mu > 1000$$

we reject the null hypothesis only if the value of the test statistic is *too large*. In such a case the decision rule is

$$\text{Reject } H_0 \text{ if } Z > z_\alpha$$

Notice that we use z_α rather than $z_{\alpha/2}$. That is because the entire area of the rejection region is located in one tail of the sampling distribution (see **Figure 12.3**).

FIGURE 12.3 Rejection region for a (right) one-tail test



If we test

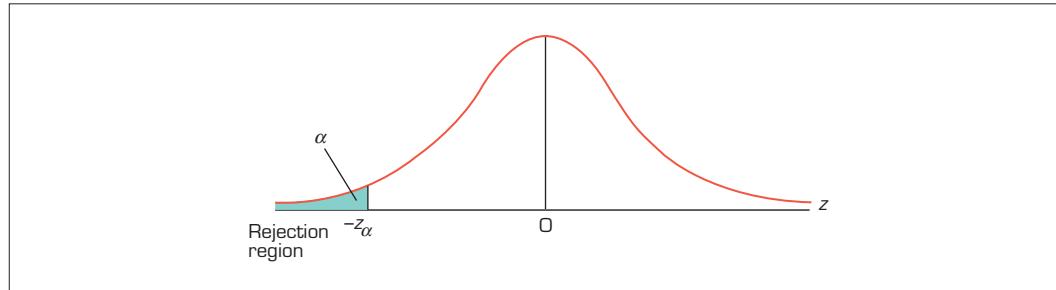
$$H_0: \mu = 1000$$

$$H_A: \mu < 1000$$

the decision rule is (see **Figure 12.4**)

Reject H_0 if $Z < -z_\alpha$

FIGURE 12.4 Rejection region for a (left) one-tail test



In the above two examples, we use a one-tail test because the alternative hypothesis lies entirely on one side of the null hypothesis. Many students have difficulty in determining when they should use a one- or two-tail test and, if it is a one-tail test, which tail to use. Remember that the alternative hypothesis is set up to answer the question posed. Thus, if we have been asked if we can conclude that the mean is different from (say) 500, then

$$H_A: \mu \neq 500$$

and we use a two-tail test. If we have been asked if there is enough evidence to show that the mean is greater than 500, then

$$H_A: \mu > 500$$

and we have a one-tail test (right tail of the sampling distribution). Finally, if we want to know if there is enough evidence to imply that the mean is less than 500, then

$$H_A: \mu < 500$$

and again we use a one-tail test (left tail of the sampling distribution).

The steps involved in hypothesis testing can be summarised in the six steps below.

IN SUMMARY

Six-step process for testing hypotheses

Step 1: Set up the null and alternative hypotheses.

Note: Since the alternative hypothesis answers the question, set this one up first. The null hypothesis will automatically follow.

Step 2: Determine the test statistic and the sampling distribution of the standardised test statistic.

Step 3: Specify the significance level.

Note: We usually set $\alpha = 0.01, 0.05$ or 0.10 , but other values are possible.

Step 4: Define the decision rule in terms of the standardised test statistic.

Note: This involves using the appropriate statistical table from Appendix B to determine the critical value(s) and the rejection region.

Step 5: Calculate the value of the standardised test statistic under the null hypothesis H_0 .

Note: Non-mathematicians need not fear. Only simple arithmetic is needed.

Step 6: Make a decision and answer the question.

Note: This involves comparing the calculated value of the standardised test statistic (step 5) with the decision rule (step 4) and making a decision. Remember to answer the original question. Making a decision about the null hypothesis is not enough.

12.2 Testing the population mean μ when the population variance σ^2 is known

In the previous section we discussed the main components of hypothesis testing. We are now going to use a method to test hypotheses about a population mean when the population variance is known. As you are about to see, the technique of hypothesis testing requires us to fill in the blanks in the following steps:

- 1 H_0 : _____
- H_A : _____
- 2 Test statistic: _____
- 3 Significance level: _____
- 4 Decision rule: _____
- 5 Value of the test statistic: _____
- 6 Conclusion: _____

We will now demonstrate these six steps with an example.

EXAMPLE 12.1

LO2 LO3

Mean diameter of lids of honey jars

XM12-01 A farm that sells honey uses a well-known supplier to supply the lids for its honey containers. The farm has received complaints from its customers that honey is leaking from the container. The farm suspects that the lids are not produced with the exact diameter of 4 cm and hence the leak. The diameters of a random sample of 10 container lids were measured and the results are shown below. Assuming that the population of diameters of container lids is normally distributed with a standard deviation of 0.4 cm, can we conclude at the 5% significance level that the mean diameter is not 4 cm?

3.84	4.00	3.92	4.16	4.24	3.84	3.92	3.76	3.68	3.04
------	------	------	------	------	------	------	------	------	------

Solution

Identifying the technique

The objective of this problem is that we want to know if the average diameter of the honey container lids is different from 4 cm. Thus, the parameter of interest is the population mean μ and the data type is numerical. Therefore, the hypotheses to be tested are

$$H_0: \mu = 4$$

$$H_A: \mu \neq 4$$

As discussed in Section 12.1, the test statistic is the best estimator of the parameter. In Section 10.1 (Chapter 10) we pointed out that the best estimator of a population mean μ is the sample mean \bar{X} . In this example, a sample of 10 container lids produced a sample mean of $\bar{X} = 3.84$ cm. To answer the question posed in this example, we need to answer the question: Is a sample mean of $\bar{X} = 3.84$ sufficiently different from 4 to allow us to infer that the population mean μ is not equal to 4? To answer this question, we need to specify the fourth component of the test – the rejection region.

It seems reasonable to reject the null hypothesis if the value of the sample mean is either large or small relative to 4. If we had calculated the sample mean to be, say, 2.0 or 6.0, it would be quite apparent that the null hypothesis is false and we would reject it. On the other hand, values of \bar{X} close to 4.0 (such as 3.9 or 4.1) do not allow us to reject the null hypothesis because it is entirely possible to observe a sample mean of 3.9 or 4.1 from a population whose mean is 4.0. Unfortunately, the decision is not always so obvious. In this example, the sample mean was calculated to be 3.84, a value neither very far away from nor close to 4.0. In order to make a decision about this sample mean, we need to set up the rejection region. Suppose we define the value of the sample mean that is just small enough to reject the null hypothesis as \bar{X}_L and the value of the sample mean that is just large enough to reject the null hypothesis as \bar{X}_S . Note that \bar{X}_L and \bar{X}_S are the critical values. We can now specify the rejection region as





Reject H_0 if $\bar{X} < \bar{X}_S$ or $\bar{X} > \bar{X}_L$

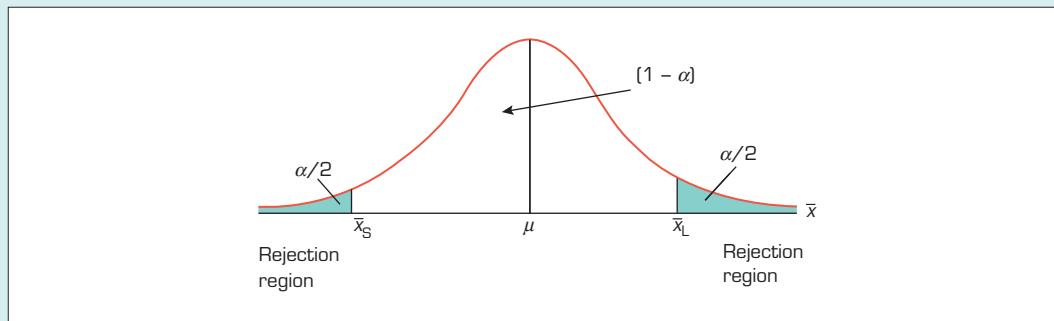
Since a Type I error is defined as rejecting a true null hypothesis, and the probability of committing a Type I error is α , it follows that

$$\alpha = P(\text{Rejecting } H_0 \text{ given that } H_0 \text{ is true})$$

$$= P(\bar{X} < \bar{X}_S \text{ or } \bar{X} > \bar{X}_L | H_0 \text{ is true})$$

Figure 12.5a depicts the sampling distribution and the rejection region. (The central limit theorem tells us that the sampling distribution of the sample mean is either normal or approximately normal for sufficiently large sample sizes.)

FIGURE 12.5A Sampling distribution of \bar{X}



If α is the probability that \bar{X} falls into the rejection region, then $1 - \alpha$ is the probability that it doesn't. Thus,

$$P(\bar{X}_S < \bar{X} < \bar{X}_L | H_0 \text{ is true}) = 1 - \alpha$$

Given that the population is normally distributed, from Section 9.4, we know that the sampling distribution of \bar{X} is also normally distributed, with mean μ and standard deviation σ/\sqrt{n} . As a result, we can standardise \bar{X} and obtain the following conditional probability:

$$P\left(\frac{\bar{X}_S - \mu}{\sigma/\sqrt{n}} < Z < \frac{\bar{X}_L - \mu}{\sigma/\sqrt{n}} | H_0 \text{ is true}\right) = 1 - \alpha$$

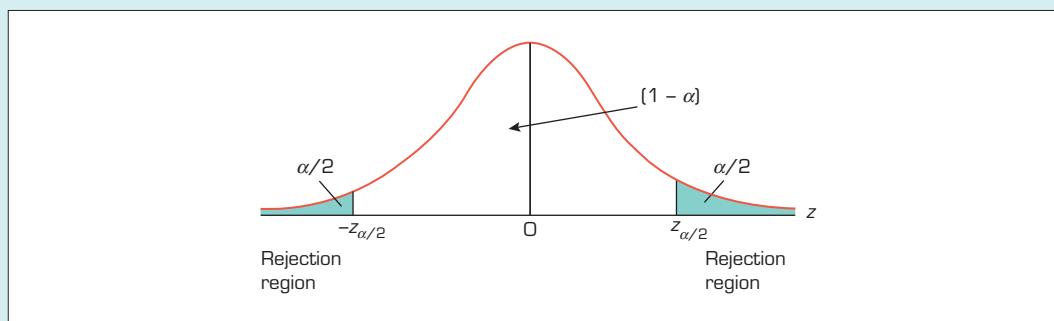
From Section 10.2, for a standard normal random variable, we have

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

As the last two probability statements involve the same distribution (standard normal) and the same probability $(1 - \alpha)$, it follows that the limits are identical. This is depicted in **Figure 12.5b**. Therefore,

$$\frac{\bar{X}_S - \mu}{\sigma/\sqrt{n}} = -z_{\alpha/2} \quad \text{and} \quad \frac{\bar{X}_L - \mu}{\sigma/\sqrt{n}} = z_{\alpha/2}$$

FIGURE 12.5B Sampling distribution of $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$





Calculating manually using \bar{X}

In this example, we know that $\sigma = 0.4$ and $n = 10$, and because the probabilities defined above are conditional upon the null hypothesis being true, we have $\mu = 4$. And finally, with $\alpha = 0.05$, $z_{\alpha/2} = z_{0.025} = 1.96$. We can now solve for \bar{X}_S and \bar{X}_L . First consider

$$\frac{\bar{X}_S - \mu}{\sigma/\sqrt{n}} = -z_{\alpha/2}$$

That is,

$$\bar{X}_S = \mu - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$\bar{X}_S = 4.0 - 1.96 \times \frac{0.4}{\sqrt{10}} = 4.0 - 0.248 = 3.752$$

Similarly, we find

$$\bar{X}_L = \mu + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Therefore,

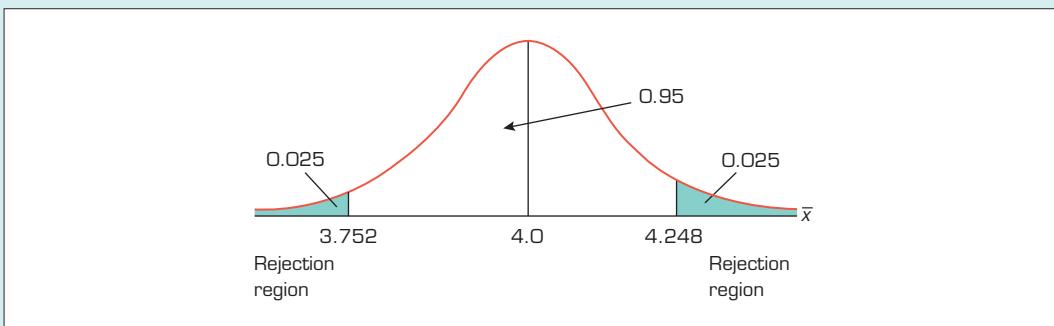
$$\bar{X}_L = 4.0 + 1.96 \times \frac{0.4}{\sqrt{10}} = 4.0 + 0.248 = 4.248$$

Therefore, the rejection region is

$$\bar{X} < 3.752 \text{ or } \bar{X} > 4.248$$

Figure 12.5c depicts the sampling distribution and rejection region in terms of \bar{X} for this example.

FIGURE 12.5C Sampling distribution of \bar{X} for Example 12.1



The sample mean was found to be $\bar{X} = 3.84$. Since the sample mean is not in the rejection region, we do not reject the null hypothesis. Therefore, there is sufficient evidence to infer that the mean diameter is equal to 4 cm.



The preceding test used the test statistic \bar{X} ; as a result, the rejection region had to be set up in terms of \bar{X} . An easier method to use specifies that the test statistic be the standardised value of \bar{X} . That is, we use the standardised test statistic:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

and as shown in **Figure 12.5b**, the rejection region consists of all values of Z that are less than $-z_{\alpha/2}$ or greater than $z_{\alpha/2}$. Algebraically, the rejection region is

$$Z < -z_{\alpha/2} \text{ or } Z > z_{\alpha/2}$$

This is more easily represented as

$$|Z| > z_{\alpha/2}$$

We can redo Example 12.1 using the standardised test statistic following the six-step process introduced in the previous section.

EXAMPLE 12.1 (CONTINUED)

LO2 LO3

Using the standardised test statistic*Step 1: Null and alternative hypotheses:*

$$H_0: \mu = 4$$

$$H_A: \mu \neq 4 \quad (\text{Two-tail test})$$

Step 2: Test statistic:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

 Z has a standard normal distribution as \bar{X} is normal or approximately normal.*Step 3: Level of significance:*

$$\alpha = 0.05$$

*Step 4: Decision rule:*Reject H_0 if $|Z| > z_{\alpha/2} = z_{0.025} = 1.96$ or Reject H_0 if $Z < -1.96, Z > 1.96$

For the next step, we will perform the calculations by hand as well as by computer. (The Excel output and commands are shown at the end of this example.)

*Step 5: Value of the test statistic:***Calculating manually**From the data information, $\bar{X} = 3.84$, $\sigma = 0.4$, $n = 10$ and $\mu_0 = 4.0$.

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{3.84 - 4.0}{0.4/\sqrt{10}} = -1.26$$

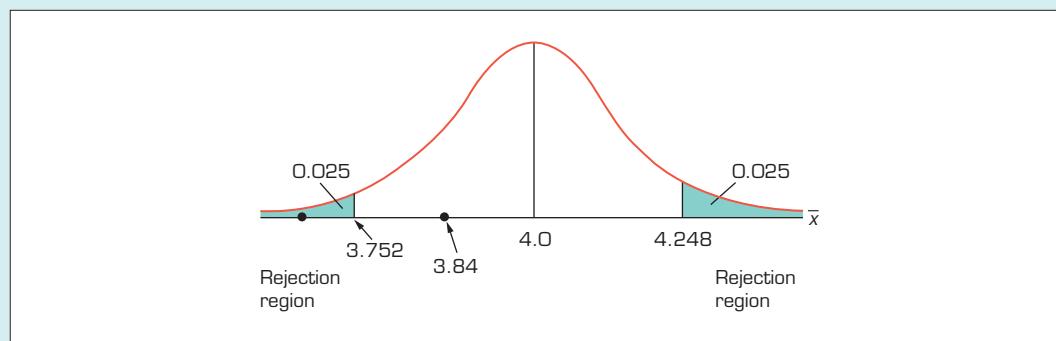
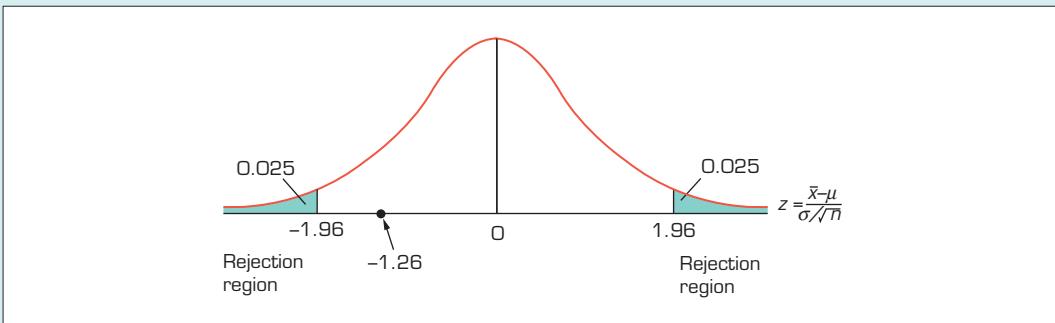
*Step 6: Conclusion:*The value of the test statistic is -1.26 , which is greater than -1.96 , so we do not reject the null hypothesis at the 5% level of significance.**Interpreting the results**As we do not reject the null hypothesis at the 5% level of significance, we conclude that there is not enough evidence to infer that the mean diameter is not equal to 4 cm. As you can see, the conclusions we draw from using the test statistic \bar{X} and the standardised test statistic Z are identical. **Figures 12.6** and **12.7** depict the two sampling distributions, highlighting the equivalence of the two tests.**FIGURE 12.6** Sampling distribution of \bar{X} for Example 12.1

FIGURE 12.7 Sampling distribution of $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ for Example 12.1



Because of its convenience and because statistical software packages use them, the standardised test statistic is used throughout this book. Deferring to popular usage we will refer to the *standardised test statistic* simply as the *test statistic*.

Using the computer

Using Excel workbook

We can use the **z-Test_Mean** worksheet in the **Test Statistics** workbook that is provided on the companion website (accessible through <https://login.cengagebrain.com/>). The output and commands are shown below.

	A	B	C	D
1	z-Test of a Mean			
2				
3	Sample mean	3.84	z Stat	-1.26
4	Population standard deviation	0.4	P(Z<=z) one-tail	0.1030
5	Sample size	10	z Critical one-tail	1.6449
6	Hypothesised mean	4	P(Z<=z) two-tail	0.2059
7	Alpha	0.05	z Critical two-tail	1.9600

COMMANDS

Open the worksheet **z-Test_Mean** available in the **Test Statistics** workbook. Type the values of the sample mean \bar{X} (3.84), population standard deviation σ (0.4), sample size n (10), hypothesised mean μ (4) and alpha α (0.05).

Using XLSTAT

Alternatively, when raw data are available, we can use XLSTAT, an EXCEL Add-in, to perform this task.

	B	C	D	E	F	G
1	Theoretical mean: 4					
2	Significance level (%): 5					
3						
4	Summary statistics:					
5	Variable	Observations	Minimum	Maximum	Mean	Std. deviation
6	Diameter (cms)	10	3.040	4.240	3.840	0.329
7						
8	One-sample z-test / Two-tailed test:					
9	Difference	-0.160				
10	z (Observed value)	-1.265				
11	z (Critical value)	1.960				
12	p-value (Two-tailed)	0.206				
13	alpha	0.05				
14						
15	Test interpretation:					
16	HO: The mean is equal to 4.					
17	HA: The mean is different from 4.					
18	As the computed p-value is greater than the significance level alpha=0.05, one cannot reject the null hypothesis HO.					

(Note: This is only a partial output).

COMMANDS

- 1 Type the data into one column or open the data file (**XM12-01**).
- 2 Click **XLSTAT**, **Parametric Tests** and **One-sample t-test and z-test**.
- 3 In the **Data**: dialog box type the input range (**A1:A11**). Click **Column labels** if the first row contains the name of the variable (as in this example). Check **z-test** and do not check Student's t-test.
- 4 Click the **Options** tab and choose **Mean ≠ Theoretical mean** in the **Alternative hypothesis**: box. Type the value of α (in per cent) in the **Significance**: box (5). If there are blanks in the column (usually used to represent missing data) click **Missing data**, remove the observations. For the **Variance for z-test**: check **User defined: Variance**: and type the value of σ^2 (**0.16**). Click **OK** and then **Continue**.

12.2a Further discussions on the results of the test

In Example 12.1, we did not reject the null hypothesis. Does this *prove* that the alternative hypothesis is false? The answer is no. Because our conclusion is based on sample data (and not on the entire population), we can never prove anything by using statistical inference. Consequently, we summarise the test by stating that *there is enough statistical evidence to infer that the null hypothesis is true and that the alternative hypothesis is false*.

Now suppose that \bar{X} had equalled 3.68 instead of 3.84. We would then have calculated $z = -2.53$, which is in the rejection region. Could we conclude on this basis that there is enough statistical evidence to infer that the null hypothesis is false and hence that $\mu \neq 4$? Again the answer is no because it is absurd to suggest that a sample mean of 3.68 provides *any* evidence to infer that the population mean is 4. Because we are testing a single value of the parameter under the null hypothesis, we can never have enough statistical evidence to establish that the null hypothesis is true (unless we sample the entire population).

Consequently, if the value of the test statistic does not fall into the rejection region, rather than say we *accept the null hypothesis* (which implies that we are stating that the null hypothesis is true) we state that we *do not reject the null hypothesis*, and we conclude that *not enough evidence exists to show that the alternative hypothesis is true*. Although it may appear that we are being overly technical, such is not the case. Your ability to set up tests

of hypotheses properly and to interpret their results correctly very much depends on your understanding of this point. Notice that no matter what the result of the test, the conclusion is based on the alternative hypothesis. In the final analysis, there are only two possible conclusions of a hypothesis test.

IN SUMMARY

Conclusions of a test of hypothesis

If we *reject the null hypothesis*, we conclude that there is enough statistical evidence to infer that the alternative hypothesis is true.

If we *do not reject the null hypothesis*, we conclude that there is not enough statistical evidence to infer that the alternative hypothesis is true.

Observe that, in the end, the alternative hypothesis is the more important one. It is the alternative hypothesis that answers the question, not the null hypothesis. This point is crucial. Whatever you are trying to show statistically must be represented by the alternative hypothesis (bearing in mind that you have only three choices for the alternative hypothesis – greater than, less than, and not equal to).

When we introduced statistical inference in Chapter 9, we pointed out that the first step in the solution is to identify the technique. Part of this process when the problem involves hypothesis testing is the specification of the hypotheses. Because the alternative hypothesis answers the question, we will identify it first. The null hypothesis automatically follows because the null hypothesis must specify equality. However, by tradition, when we list the two hypotheses, the null hypothesis comes first, followed by the alternative hypothesis. All the examples in this book follow this format.

The test illustrated by Example 12.1 is called a *two-tail test* because the rejection region is equally partitioned into the two tails of the sampling distribution. We now present examples to illustrate the *one-tail test*, which is characterised by a rejection region located in only one tail of the sampling distribution.

EXAMPLE 12.2

LO3

Are packs of garlic correctly labelled?

XM12-02 A number of government agencies are devoted to ensuring that food producers package their products in such a way that the weight or volume of the contents listed on the label is correct. A quality control inspector wanted to use statistical inference and techniques to investigate customer complaints regarding the weight of 500g garlic packs imported from overseas by a particular wholesale distributor, Ausvege Ltd. For example, garlic packs whose labels state that the contents have a net weight of 500g must have a net weight of at least 500g. However, it is impossible to check all garlic packs sold in the country. If the mean weight of a sample of the product provides sufficient evidence to infer that the mean weight of all 500g Ausvege garlic packs is less than 500g, then the product label is deemed to be unacceptable. Suppose that the inspector weighs the contents of a random sample of 25 garlic packs labelled ‘Net weight: 500 grams’ distributed by Ausvege Ltd and records the measurements below. Using a 5% significance level, can the inspector conclude that the product label is unacceptable? (Assume that the inspector knows from previous experiments that the weight of all 500g garlic packs distributed by Ausvege Ltd is normally distributed with a standard deviation of 10g.)

Net weight of 25 random ‘500-gram’ garlic packs

500	502	505	498	501	496	504	505	500	499	498	503	510
503	505	499	497	502	504	507	510	495	470	501	480	



Solution

Identifying the technique

The objective of the study is to draw a conclusion about the mean weight of all 500g garlic packs distributed by Ausvege Ltd. Thus, the parameter to be tested is the population mean μ . The data type is numerical. We want to know if there is enough statistical evidence to show that the population mean is less than 500g. We use the six-step process.

Step 1. Null and alternative hypotheses:

$$H_0: \mu = 500$$

$$H_A: \mu < 500 \quad (\text{Left one-tail test})$$

Step 2. Test statistic:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

As the population of weights is normally distributed, \bar{X} is normally distributed. Therefore, as the population standard deviation σ is also known, the standardised test statistic Z has a standard normal distribution.

Step 3. Level of significance:

$$\alpha = 0.05$$

Step 4. Decision rule:

In this example, we locate the rejection region in the left tail of the sampling distribution. To understand why, remember that we are trying to decide if there is enough statistical evidence to infer that the mean is less than 500g (which is the alternative hypothesis). If we observe a large sample mean (and hence a large value of Z), do we want to reject the null hypothesis in favour of the alternative? The answer is an emphatic *no*. It is absurd to conclude that if the sample mean is, say, 510, there is enough evidence to conclude that the mean of all garlic packs is *less* than 500. Consequently, we want to reject the null hypothesis only if the sample mean (and hence the value of Z) is small. How small is small enough? The answer is determined by the significance level and the rejection region. Thus, we set up the rejection region as

$$Z < -z_\alpha = -z_{0.05} = -1.645$$

Therefore, the *decision rule* for the test is:

Reject H_0 if $Z < -1.645$.

Why do we use z_α and not $z_{\alpha/2}$? Because we want the probability of incorrectly rejecting the null hypothesis (Type I error) to be α . Since we reject the null hypothesis only when Z is too small, it follows that the rejection region is $Z < -z_\alpha$.

Note that the direction of the inequality in the rejection region ($Z < -z_\alpha$) matches the direction of the inequality in the alternative hypothesis ($\mu < 500$). Also note the negative sign, since the rejection region is in the left tail (containing values of z less than zero) of the sampling distribution.

For the next step, we will perform the calculations by hand as well as by computer. (The Excel output and commands are shown at the end of this example.)

Step 5. Value of the test statistic:

Calculating manually

From the data, we calculate the sample mean:

$$\bar{X} = 499.76$$

As the population standard deviation is known to be $\sigma = 10$ g, the sample size is $n = 25$, and the value of μ is hypothesised to be 500 under H_0 ($\mu_0 = 500$), we calculate the value of the test statistic as

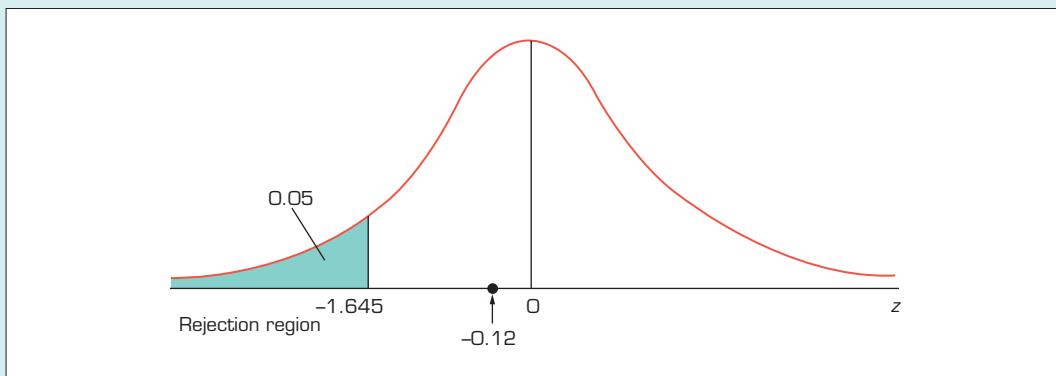
$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{499.76 - 500}{10/\sqrt{25}} = -0.12$$

Step 6: Conclusion:

Because the value of the test statistic $Z = -0.12$ is greater than the critical value -1.645 , we do not reject the null hypothesis. There is insufficient evidence to infer that the mean is less than 500g at the 5% level of significance.

Figure 12.8 depicts the sampling distribution of the standardised test statistic Z .

FIGURE 12.8 Sampling distribution for Example 12.2



Interpreting the results

Because we were not able to reject the null hypothesis, we say that there is not enough evidence to infer that the mean weight of all 500g garlic packs distributed by Ausvege Ltd is less than 500g. Note that there was some evidence to indicate that the population mean is less than 500g. (We did find the sample mean to be 499.76g.) However, to reject the null hypothesis we need enough statistical evidence, and in this case we simply did not have enough reason to reject the null hypothesis in favour of the alternative. In the absence of evidence to show that the mean weight of all garlic packs is less than 500g, the inspector would not find the labels to be unacceptable.

Using the computer

Using Excel workbook

Testing hypotheses for single population with numerical data when the population standard deviation is known can be performed using the **z-Test_Mean** worksheet in the **Test Statistics** workbook that is provided on the companion website (accessible through <https://login.cengagebrain.com/>). The output is shown below.

Excel output for Example 12.2

	A	B	C	D
1	z-Test of a Mean			
2				
3	Sample mean	499.76	z Stat	-0.12
4	Population standard deviation	10	P(Z<=z) one-tail	0.4522
5	Sample size	25	z Critical one-tail	1.6449
6	Hypothesised mean	500	P(Z<=z) two-tail	0.9045
7	Alpha	0.05	z Critical two-tail	1.9600

This Excel workbook output macro summarises the test we are performing. The value of the test statistic is $Z = -0.12$. The p -value of the test will be discussed in the next section. The instructions for the execution of this macro are provided below. See Appendix 12.A on page 529 for an alternative method.



COMMANDS

If the sample mean is already known or computed, activate the **z-test_Mean** worksheet in the **Test Statistics** workbook. Type the population standard deviation (10), sample mean (499.76), sample size (25), hypothesised value of μ (500) and level of significance α (0.05).

Using XLSTAT

Alternatively, when raw data are available, we can use XLSTAT, an EXCEL Add-in, to perform this task.

	B	C	D	E	F	G
1	Theoretical mean: 4					
2	Significance level (%): 5					
3						
4	Summary statistics:					
5	Variable	Observations	Minimum	Maximum	Mean	Std. deviation
6	Weight (gram)	25	470.000	510.000	499.760	8.521
7						
8	One-sample z-test / Two-tailed test:					
9	Difference	-0.240				
10	z (Observed value)	-0.120				
11	z (Critical value)	-1.645				
12	p-value (Two-tailed)	0.452				
13	alpha	0.05				
14						
15	Test interpretation:					
16	HO: The mean is equal to 500.					
17	HA: The mean is lower than 500.					
18	As the computed p-value is greater than the significance level alpha=0.05, one cannot reject the null hypothesis HO.					

(Note: This is only a partial output).

COMMANDS

- Type the data into one column or open the data file (**XM12-02**).
- Click **XLSTAT**, **Parametric Tests** and **One-sample t-test** and **z-test**.
- In the **Data**: dialog box type the input range (**A1:A26**). Click **Column labels** if the first row contains the name of the variable (as in this example). Check **z-test** and do not check Student's t-test.
- Click the **Options** tab and choose **Mean < Theoretical mean** in the **Alternative hypothesis**: box. Type the value of α (in per cent) in the **Significance**: box (5). If there are blanks in the column (usually used to represent missing data) click **Missing data**, remove the observations. For the **Variance for z-test**: check **User defined: Variance**: and type the value of σ^2 (100.0). Click **OK** and then **Continue**.

12.2b Setting up the hypotheses in one-tail tests

In our experience, we have observed that students often have difficulty in setting up one-tail tests of hypotheses. It must be understood that the whole point of the test-of-hypothesis procedure is to determine if there is enough statistical evidence to *support the alternative hypothesis*. Thus, to set up the alternative hypothesis simply decide what the point of the exercise is. In Example 12.1, the objective was to determine if the mean diameter of the population of honey container lids was *not equal* to 4cm. Consequently, we set up the alternative hypothesis as $H_A: \mu \neq 4$. Because the null hypothesis must specify a single value (if it didn't, what value would we use to calculate the test statistic?), the null hypothesis automatically becomes $H_0: \mu = 4$.

In Example 12.2, the purpose of the test was to determine if there was enough evidence to infer that the mean weight of the garlic packs was *less* than 500g (and hence that the labels were unacceptable). We set up the alternative hypothesis as $H_A: \mu < 500$ and the null hypothesis as $H_0: \mu = 500$.¹

You will note that even though we summarise tests by listing the null hypothesis first (by tradition), we actually determine the alternative hypothesis first; the null hypothesis is automatically defined. Let's examine one more example to solidify your understanding of hypothesis testing.

EXAMPLE 12.3

LO3

The impact of the advertising campaign

XM12-03 Past experience indicates that a monthly long-distance telephone bill is normally distributed with a mean of \$17.85 and a standard deviation of \$3.39. After an advertising campaign aimed at increasing long-distance telephone usage, the manager of the telephone company took a random sample of 25 household bills and recorded their monthly usage. The results are listed below. Do the data allow us to infer at the 5% significance level that the campaign was successful?

Monthly long-distance telephone bills

19.61	20.14	19.57	19.26	14.03	19.24	15.98	24.85	26.00
19.46	18.29	16.91	26.15	19.64	16.75	20.52	25.47	18.19
12.56	28.47	14.13	19.72	17.05	13.92	12.38		

Solution

Identifying the technique

This example deals with a normally distributed population of the monthly long-distance telephone bills. Thus, the parameter to be tested is the population mean μ . The data type is numerical. To conclude that the advertising campaign has been successful we are required to show that the mean monthly long-distance telephone bill is greater than \$17.85.

Step 1. Null and alternative hypotheses:

$$H_0: \mu = 17.85$$

$$H_A: \mu > 17.85 \quad (\text{Right one-tail test})$$

Step 2. Test statistic:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

As the population of long-distance telephone bills is normally distributed, \bar{X} is normally distributed.

Therefore, as the standard deviation σ is also known, the standardised test statistic Z has a *standard normal distribution*.

Step 3. Level of significance:

$$\alpha = 0.05$$



¹ Some statistics practitioners prefer to specify the null hypothesis as the opposite of the alternative hypothesis. For Example 12.2, such statistics practitioners would state the hypotheses as

$$H_0: \mu \geq 500$$

$$H_A: \mu < 500$$

However, they would use $\mu = 500$ in calculating the value of the test statistic (just as we did) and therefore actually test the null hypothesis $H_0: \mu = 500$. Their interpretation of the test would be identical to ours.



Step 4. Decision rule:

Reject H_0 if $Z > z_\alpha = z_{0.05} = 1.645$

Step 5. Value of the test statistic:

Calculating manually

From the data, the sample mean $\bar{X} = \$19.13$ from a sample of $n = 25$. The population standard deviation is known to be $\sigma = \$3.39$. The null hypothesis specifies $\mu_0 = 17.85$. Substituting all the parts into the test statistic, we calculate the value of the test statistic as

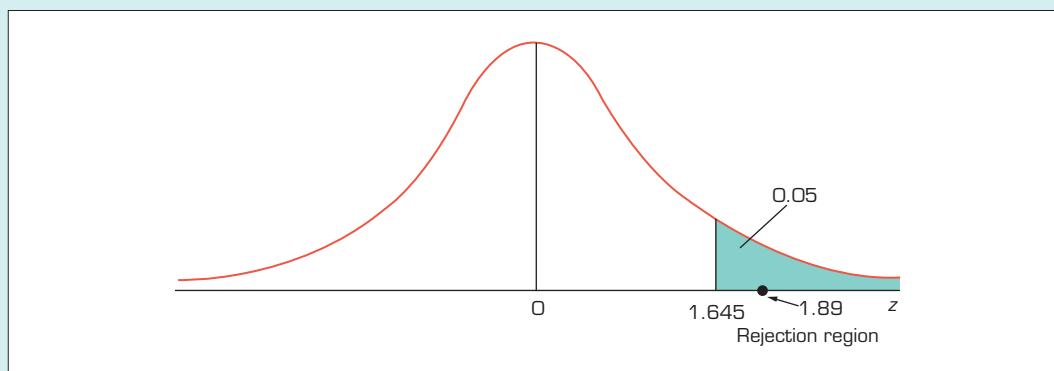
$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{19.13 - 17.85}{3.39/\sqrt{25}} = 1.89$$

Step 6. Conclusion:

Because the value of the test statistic, $Z = 1.89$, is greater than the critical value 1.645, we reject H_0 in favour of the alternative hypothesis.

Figure 12.9 describes the sampling distribution in this test.

FIGURE 12.9 Sampling distribution for Example 12.3



Interpreting the results

There is enough evidence for us to infer that the mean monthly long-distance telephone bill is greater than \$17.85. This means that the data support the belief that the advertising campaign was successful. However, bear in mind that the statistical inference is only as good as the data-gathering process. If the sample is not randomly selected, the results may be meaningless. It is also worth remembering that we have not proven the success of the advertising campaign, just that the data indicate that the advertising campaign has had a positive impact on increasing the long-distance telephone usage.

Using the computer

The Excel workbook and XLSTAT commands are the same as in Example 12.2. We present the output below.

Using Excel workbook

Excel output for Example 12.3

	A	B	C	D
1	z-Test of a Mean			
2				
3	Sample mean	19.13	z Stat	1.89
4	Population standard deviation	3.39	P[Z<=z] one-tail	0.0295
5	Sample size	25	z Critical one-tail	1.6449
6	Hypothesised mean	17.85	P[Z<=z] two-tail	0.0590
7	Alpha	0.05	z Critical two-tail	1.9600

Using XLSTAT

	B	C	D	E	F	G
1	Theoretical mean: 17.85					
2	Significance level [%]: 5					
3						
4	Summary statistics:					
5	Variable	Observations	Minimum	Maximum	Mean	Std. deviation
6	Bills	10	3.040	4.240	3.840	0.329
7						
8						
9	One-sample z-test / Two-tailed test:					
10						
11	Difference	1.282				
12	z (Observed value)	1.890				
13	z (Critical value)	1.645				
14	p-value (one-tailed)	0.029				
15	alpha	0.05				
16						
17	Test interpretation:					
18	HO: The mean is equal to 17.85.					
19	HA: The mean is greater than 17.85.					
20	As the computed p-value is greater than the significance level alpha=0.05, one cannot reject the null hypothesis HO.					

12.2c Testing hypotheses and confidence interval estimators

As you've seen, the test statistic and the confidence interval estimator are both derived from the sampling distribution. It shouldn't be a surprise then that we can use the confidence interval estimator to test hypotheses. To illustrate, consider Example 12.1. The 95% confidence interval estimate of the population mean is

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 3.84 \pm 1.96 \times \frac{0.4}{\sqrt{10}} = 3.84 \pm 0.248$$

$$\text{LCL} = 3.592 \text{ and UCL} = 4.088$$

We estimate that μ lies between 3.592 cm and 4.088 cm. Because this interval includes 4, we cannot conclude that there is sufficient evidence to infer that the population mean differs from 4.

In Example 12.2, the 90% confidence interval estimate is

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 499.76 \pm 1.645 \times \frac{10}{\sqrt{25}} = 499.76 \pm 3.29$$

$$\text{LCL} = 496.47 \text{ and UCL} = 503.05$$

The interval estimate includes 500, allowing us to conclude that there is not enough evidence to conclude that the population mean weight is less than 500g.

As you can see, the confidence interval estimator can be used to conduct tests of hypotheses. This process is equivalent to the rejection region approach. However, instead of finding the critical values of the rejection region and determining whether the test statistic falls into the rejection region, we compute the interval estimate and determine whether the hypothesised value of the mean falls into the interval.

Using the interval estimator to test hypotheses has the advantage of simplicity. Evidently, we do not need the formula for the test statistic; we need only the interval estimator. However, there are also some serious drawbacks.

The confidence interval estimator does not yield a p -value, which we will argue (see Section 12.3) is the better way to draw inferences about a parameter. Using the confidence interval estimator to test hypotheses forces the decision maker into making a ‘reject’/‘do not reject’ decision rather than providing information about how much statistical evidence exists to be judged with other factors in the decision process. Furthermore, we only postpone the point in time when a test of hypothesis must be used. In later chapters, we will present problems in which only a test produces the information we need to make decisions.

12.2d Developing an understanding of statistical concepts 1

As is the case with the confidence interval estimator, the test of hypothesis is based on the sampling distribution of the sample statistic. The result of a test of hypothesis is a probability statement about the sample statistic. We assume that the population mean is specified by the null hypothesis. We then compute the test statistic and determine how likely it is to observe this large (or small) α value when the null hypothesis is true. If the probability is small, we conclude that the assumption that the null hypothesis is true is unfounded and we reject it.

12.2e Developing an understanding of statistical concepts 2

When we (or the computer) calculate the value of the test statistic

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

we are also measuring the difference between the sample statistic and the hypothesised value of the parameter μ in terms of the standard error σ/\sqrt{n} . In Example 12.1, we found that the value of the test statistic was $Z = -1.26$. This means that the sample mean was 1.26 standard errors below the hypothesised value of μ . The standard normal probability table told us that this value is not considered unlikely. As a result, we did not reject the null hypothesis.

The concept of measuring the difference between the sample statistic and the hypothesised value of the parameter in terms of the standard errors is one that will be used throughout this book.

IN SUMMARY

Factors that identify the z-test of μ

- 1** Problem objective: to describe a single population
- 2** Data type: numerical (quantitative)
- 3** Population variance: known

EXERCISES

Learning the techniques

- 12.1** Define these terms:
- a** Type I error
 - b** Type II error
 - c** significance level
 - d** rejection region
- 12.2** For the following tests of hypotheses, determine the rejection regions.
- | | |
|--|--|
| a $H_0: \mu = 500$
$H_A: \mu > 500$
$\alpha = 0.02, n = 100 \quad \sigma = 25$ | b $H_0: \mu = 20$
$H_A: \mu \neq 20$
$\alpha = 0.07 \quad n = 250 \quad \sigma = 2$ |
| c $H_0: \mu = 1000$
$H_A: \mu \neq 1000$
$\alpha = 0.04 \quad n = 20 \quad \sigma = 50$ | |

- 12.3** Sketch the sampling distribution and indicate the rejection region for each of the tests in Exercise 12.2. *The following exercises can be solved manually or by using Excel's **Test Statistics** workbook that is provided on the companion website (accessible through <https://login.cengagebrain.com>).*

- 12.4** A random sample of 200 observations from a normal population whose standard deviation is 300 produced a mean of 450. Test the null hypothesis $H_0: \mu = 480$ against the alternative hypothesis $H_A: \mu < 480$. Use $\alpha = 0.05$.

- 12.5** Test the following hypotheses with $\alpha = 0.01$, given that a sample of size $n = 25$ from a normal population whose variance is 100 produced $\bar{X} = 115$.

$$\begin{aligned} H_0: \mu &= 110 \\ H_A: \mu &> 110 \end{aligned}$$

- 12.6** Given that $\bar{X} = 22.3$, $\sigma = 12$ and $n = 100$, test the following hypotheses with $\alpha = 0.01$:

$$\begin{aligned} H_0: \mu &= 20 \\ H_A: \mu &\neq 20 \end{aligned}$$

- 12.7 a** Compute the test statistic in order to test the following hypotheses given that $\bar{X} = 52$, $n = 9$ and $\sigma = 5$:

$$\begin{aligned} H_0: \mu &= 50 \\ H_A: \mu &> 50 \end{aligned}$$

- b** Repeat part (a) with $n = 25$.
- c** Repeat part (a) with $n = 100$.
- d** Describe what happens to the value of the test statistic when the sample size increases.

- 12.8 a** A statistics practitioner formulated the following hypotheses:

$$\begin{aligned} H_0: \mu &= 200 \\ H_A: \mu &< 200 \end{aligned}$$

He has the following information: $\bar{X} = 190$, $n = 9$ and $\sigma = 50$. Compute the test statistic.

- b** Repeat part (a) with $\sigma = 30$.
- c** Repeat part (a) with $\sigma = 10$.
- d** Discuss what happens to the value of the test statistic when the standard deviation decreases.

- 12.9 XR12-09** Given the following data drawn from a population whose standard deviation is known to be 5, test to determine if there is enough evidence at the 5% significance level to infer that the population mean is greater than 25.

35	15	15	45	40	20	40	35	40	25	45	25	35
20	40	25	30	15	45	20	35	20	30	15	45	

- 12.10 XR12-10** Suppose that the following observations were drawn from a normal population whose standard deviation is 20. Test with $\alpha = 0.10$ to determine whether there is enough evidence to conclude that the population mean differs from 50.

42	74	66	94	56	32	58	74	82	40
----	----	----	----	----	----	----	----	----	----

Applying the techniques

- 12.11 Self-correcting exercise.** A university claims that the average tertiary entry (TE) score of applicants to its business studies program has increased during the past three years. Three years ago, the mean and the standard deviation of TE scores of the university's business studies program applicants were 920 and 20 respectively. For a sample of 36 of this year's applicants for the program, the mean TE score was 925. At the 5% level of significance, can we conclude that the university's claim is true? (Assume that the standard deviation is unchanged.)

- 12.12** The average weekly wage of all the workers at a large factory is \$626.40. In a random sample of 100 male workers at the factory, it was found that the mean weekly wage is \$682.00. Assuming that the population standard deviation of weekly male wage is \$82.09, can we conclude (with $\alpha = 0.05$) that the mean weekly wage of male workers is greater than the overall mean weekly wage?

- 12.13** The manager of a department store is thinking about establishing a new billing system for the store's credit customers. After a thorough financial analysis, she determines that the new system will be cost effective only if the mean monthly account is greater than \$70. A random sample of 200 monthly accounts is drawn for which the sample mean is \$74. The manager knows that the accounts are normally distributed with a standard deviation of \$30. Is there enough evidence at the 5% significance level to conclude that the new system will be cost-effective?

- 12.14** A production line that assembles computer keyboards has been experiencing problems since a new process was instituted. The supervisor notes that there has been an increase in defective units and occasional backlogs when the productivity of one station is not matched by that of the other stations. Upon reviewing the process, the supervisor discovered that the management scientists who developed the production process assumed that the amount of time to complete a critical part of the

process is normally distributed with a mean of 130 seconds and a standard deviation of 15 seconds. The supervisor is satisfied that the process time is normally distributed with a standard deviation of 15 seconds, but he is unsure about the mean time. In order to examine the problem, he measures the times for 100 assemblies. The mean of these times is calculated to be 126.8 seconds. Can the supervisor conclude at the 5% significance level that the assumption that the mean assembly time is 130 seconds is incorrect?

Computer/manual applications

The following exercises require the use of a computer and software. The answers may be calculated manually using the sample summary statistics provided.

- 12.15 XR12-15** A manufacturer of light bulbs advertises that, on average, its long-life bulb will last for more than 5000 hours. To test the claim, a statistics practitioner took a random sample of 100 bulbs and measured the length of time until each bulb burned out. The results are recorded and some of the data are listed below. If we assume that the lifetime of this type of bulb has a standard deviation of 400 hours, can we conclude at the 5% significance level that the claim is true?

4061	5361	4531	...	4943	5820
------	------	------	-----	------	------

Sample statistics: $\bar{X} = 5064.96$, $n = 100$

- 12.16 XR12-16** In the midst of labour-management negotiations, the CEO of a company argues that the company's blue-collar workers, who are paid an average of \$60 000 per year, are well paid because the mean annual income of all blue-collar workers in the country is less than \$60 000. That figure is disputed by the union, which does not believe that the mean blue-collar income is less than \$60 000. To test the company CEO's belief, an arbitrator draws a random sample of 350 blue-collar workers from across the country and asks each to report his or her annual income. Some of the recorded results are shown below. If the arbitrator assumes that the blue-collar incomes are normally distributed with a standard deviation of \$16 000, can it be inferred at the 5% significance level that the company CEO is correct?

58218	43092	60834	...	86458	64860
-------	-------	-------	-----	-------	-------

Sample statistics: $\bar{X} = 58\ 239.04$, $n = 350$

12.3 The *p*-value of a test of hypothesis

It is important for you to realise that the result of a statistical procedure is only one of several factors considered prior to making a decision. In Example 12.3, for instance, the manager concluded that there was enough statistical evidence to show that the advertising campaign was successful. However, we did not prove that the campaign was successful by using such procedures; we merely showed that statistical evidence existed to that effect. What is really needed in this situation is a measure of how much statistical evidence exists, so that it can be weighed in relation to other factors. In this section, we present such a measure: the *p*-value of a test.

To understand this definition, review Example 12.3, in which the value of the test statistic was $z = 1.89$ and where, with $\alpha = 0.05$, the rejection region was $Z > 1.645$. In that instance, we rejected the null hypothesis. Notice that our test's conclusion depended on the choice of the significance level α . Had we chosen, say, $\alpha = 0.01$, the rejection region would have been $Z > 2.33$ and we could not reject the null hypothesis. Notice, however, that we did not have to decrease α very much in order to change the decision. Values of $\alpha = 0.02$ or 0.025 or even 0.029 lead to the same conclusion as $\alpha = 0.01$, but $\alpha = 0.03$ produces the rejection region $Z > 1.88$, which does result in rejecting the null hypothesis.

Table 12.1 summarises the relationship between the different values of α and our test conclusion. As you can see, the smallest value of α that would lead to the rejection of the null hypothesis (i.e. the *p*-value) must lie between 0.029 and 0.030. We can determine this value more accurately and more simply by realising that the *p*-value is the probability that $Z > 1.89$. From Table 3 in Appendix B, we find

$$p\text{-value} = P(Z > 1.89) = 0.0294$$

p-value

Smallest value of α that would lead to the rejection of the null hypothesis.

Figure 12.10 demonstrates how we determine this value.

FIGURE 12.10 p -value of the test in Example 12.3

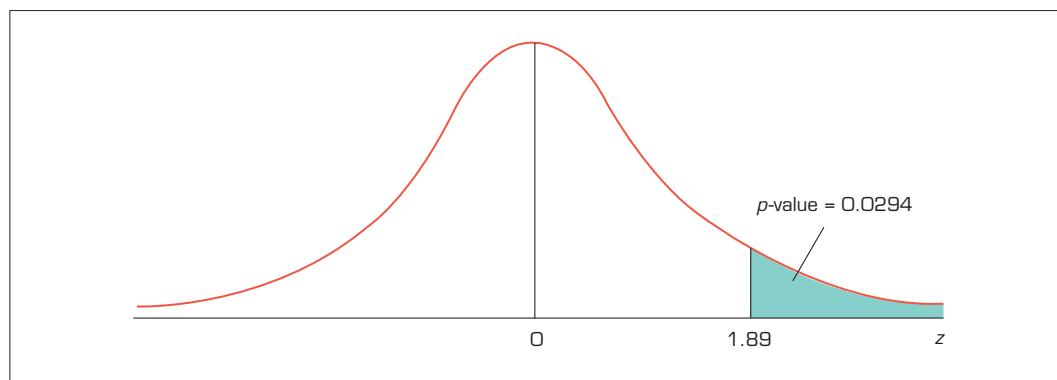


TABLE 12.1 Rejection region for a variety of values of α for Example 12.3

Value of α	Rejection region	Decision with $z = 1.89$
0.01	$Z > 2.33$	Do not reject H_0 .
0.02	$Z > 2.05$	Do not reject H_0 .
0.025	$Z > 1.96$	Do not reject H_0 .
0.029	$Z > 1.90$	Do not reject H_0 .
0.030	$Z > 1.88$	Reject H_0 .
0.05	$Z > 1.645$	Reject H_0 .

It is important to understand that the calculation of the p -value depends on, among other things, the alternative hypothesis. For Example 12.2, the hypotheses were

$$H_0: \mu = 500$$

$$H_A: \mu < 500$$

and we found $Z = -0.12$. Because the rejection region is

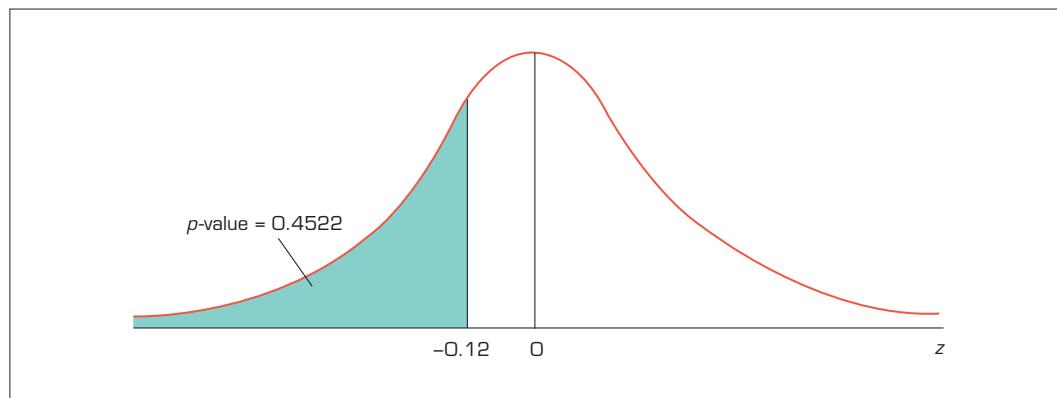
$$Z < -z_\alpha$$

the p -value is the probability that Z is less than -0.12 . That is,

$$p\text{-value} = P(Z < -0.12) = 0.4522$$

Figure 12.11 depicts this calculation.

FIGURE 12.11 p -value of the test in Example 12.2



The p -value of a two-tail test is calculated somewhat differently. As an illustration, consider Example 12.1, in which we tested

$$H_0: \mu = 0.5$$

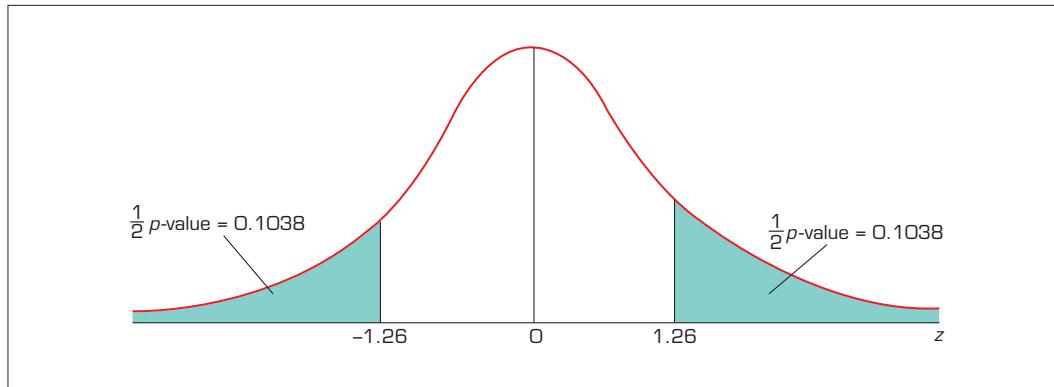
$$H_A: \mu \neq 0.5$$

and found $z = -1.26$. Because the rejection region in a two-tail test is $|Z| > z_{\alpha/2}$, the probability that Z is less than -1.26 must be doubled in order to determine the p -value. That is,

$$p\text{-value} = 2P(Z < -1.26) = 2(0.1038) = 0.2076$$

Figure 12.12 describes this calculation.

FIGURE 12.12 p -value of the test in Example 12.1



12.3a Summary of calculation of the p -value

Let z_0 be the actual value of the test statistic, and let μ_0 be the value of μ specified under the null hypothesis.

$$\begin{aligned} \text{If } H_A: \mu > \mu_0, \quad p\text{-value} &= P(Z > z_0) \\ \text{If } H_A: \mu < \mu_0, \quad p\text{-value} &= P(Z < -z_0) \\ \text{If } H_A: \mu \neq \mu_0, \quad p\text{-value} &= 2P(Z > z_0) \text{ if } z_0 > 0 \\ &\quad = 2P(Z < -z_0) \text{ if } z_0 < 0 \end{aligned}$$

or, perhaps more simply: $p\text{-value} = 2P(|Z| > |z_0|)$

12.3b Interpreting the p -value

The p -value is an important number because it measures the amount of statistical evidence that supports the alternative hypothesis. To understand this interpretation fully, again refer to Example 12.3, in which we tested

$$H_0: \mu = 17.85$$

$$H_A: \mu > 17.85$$

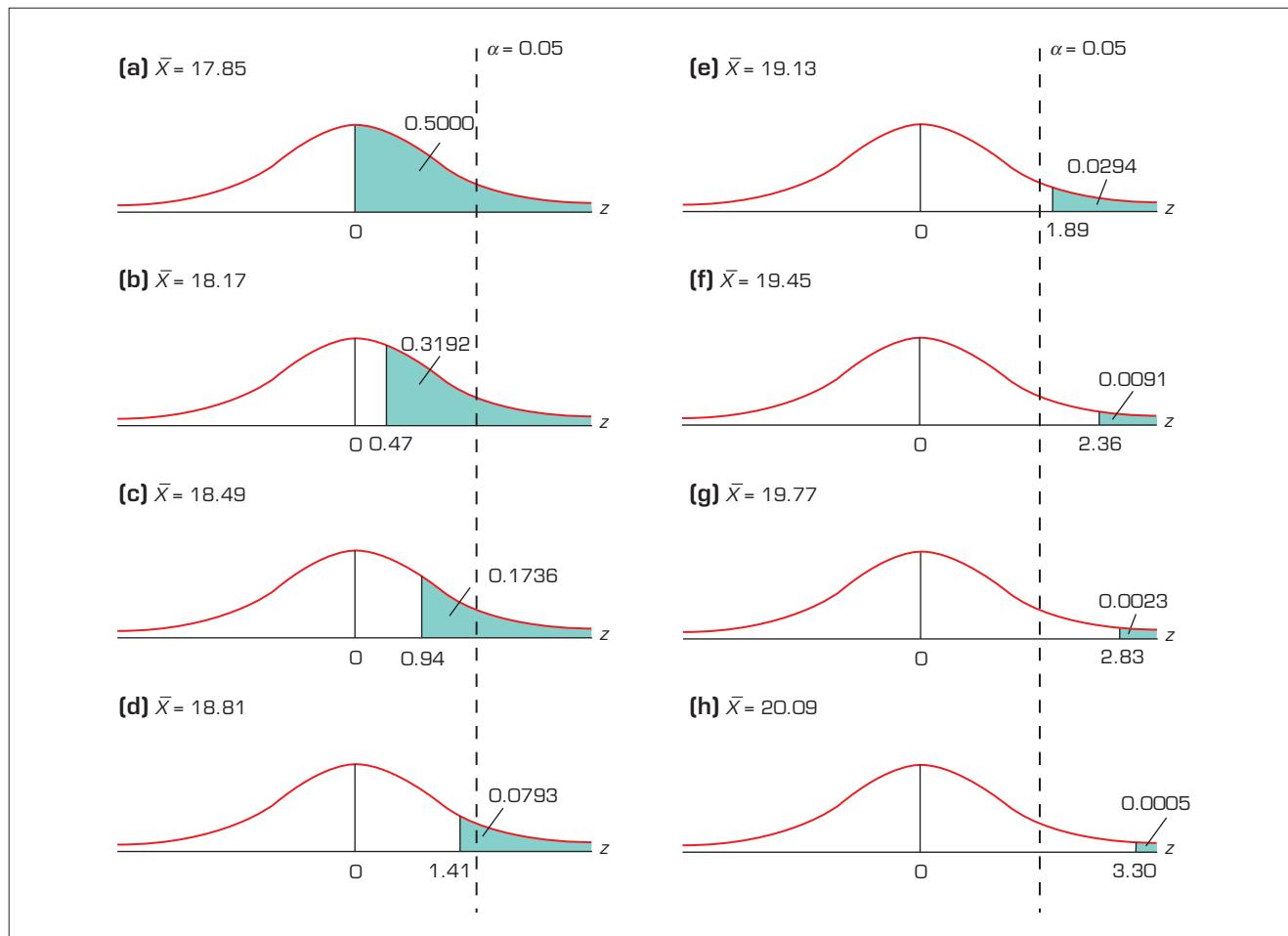
We found $\bar{X} = \$19.13$, which yielded $Z = 1.89$ and $p\text{-value} = 0.0294$. Had we observed $\bar{X} = \$17.85$, then the test statistic would be $Z = 0$ and the $p\text{-value} = 0.5$, which indicates that there is very little evidence to infer that the population mean is greater than 17.85. If \bar{X} had equalled \$20.09, the test statistic would equal 3.30 with a p -value of 0.0005, indicating that there is a great deal of evidence to infer that the mean exceeds \$17.85.

In **Table 12.2**, we list several values of \bar{X} , the resulting test statistics and p -values. Notice that as \bar{X} increases, so does the test statistic. However, as the test statistic increases, the p -value decreases. **Figure 12.13** illustrates this relationship. As \bar{X} moves further away from the value specified in the null hypothesis (17.85), there is more evidence to indicate that the alternative hypothesis is true. This is reflected in the value of the test statistic and in the p -value. That is, the more evidence that exists to reject the null hypothesis in favour of the alternative hypothesis, the *greater* is the test statistic and the *smaller* is the p -value.

TABLE 12.2 Test statistics and p -values for Example 12.3

Sample mean \bar{X}	Test statistic $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$	p -value	Decision with $\alpha = 0.05$
17.85	0.00	0.5000	Do not reject H_0 .
18.17	0.47	0.3192	Do not reject H_0 .
18.49	0.94	0.1736	Do not reject H_0 .
18.81	1.41	0.0793	Do not reject H_0 .
19.13	1.89	0.0294	Reject H_0 .
19.45	2.36	0.0091	Reject H_0 .
19.77	2.83	0.0023	Reject H_0 .
20.09	3.30	0.0005	Reject H_0 .

FIGURE 12.13 Relationship between z and p -value in Example 12.3



To illustrate further, consider Example 12.2, in which we tested

$$H_0: \mu = 500$$

$$H_A: \mu < 500$$

In this example, smaller values of \bar{X} (values smaller than 500) represent stronger support of the alternative hypothesis $H_A: \mu < 500$. Smaller values of \bar{X} produce smaller test statistics with smaller p -values. **Table 12.3** demonstrates this point. In this example, as in Example 12.3, the greater the evidence to support the alternative hypothesis, the smaller the p -value.

TABLE 12.3 Test statistics and p -values for Example 12.2

Sample mean \bar{X}	Test statistic $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$	p -value	Decision with $\alpha = 0.05$
500.00	0.00	0.5000	Do not reject H_0 .
498.74	-0.63	0.2643	Do not reject H_0 .
497.50	-1.25	0.1056	Do not reject H_0 .
496.24	-1.88	0.0301	Reject H_0 .
495.00	-2.50	0.0062	Reject H_0 .
493.74	-3.13	0.0009	Reject H_0 .

IN SUMMARY

Interpreting the p -value

- A small p -value indicates that there is ample evidence to support the alternative hypothesis.
- A large p -value indicates that there is little evidence to support the alternative hypothesis.

Because the p -value measures the statistical evidence supporting the alternative hypothesis, we can use the p -value to make the decision.

12.3c Using the p -value to draw conclusions

In order to draw conclusions about the hypotheses earlier in this section, we used Table 3 in Appendix B and a predetermined value of α to determine the rejection region in each case, and to discover whether or not the test statistic value fell into the rejection region. We will call this approach the *rejection region method*. The p -value method is simpler. All we need to do is judge whether the p -value is small enough to justify our rejecting the null hypothesis in favour of the alternative. What is considered small enough? The answer to this question is whatever the manager (for example) decides. In Example 12.3, we found the p -value to be 0.0294. If the manager decided that, for this test (taking into account all the other factors), any p -value less than $\alpha = 0.01$ was small enough to support the alternative hypothesis, then a p -value of 0.0294 would be relatively large and she would conclude that the statistical evidence did not establish that the advertising campaign was successful. However, if she felt that $\alpha = 0.05$ or less was small enough to support the alternative hypothesis, then the p -value of 0.0294 would be relatively small. It follows that the manager would conclude from the statistical evidence that the advertising campaign was successful.

12.3d Describing the p -value

Statistics practitioners often translate p -values using the following descriptive terms.

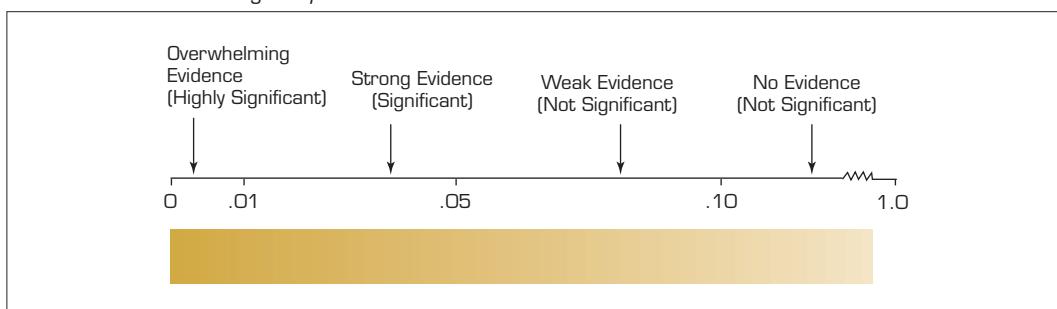
- If the p -value is less than 1%, we say that there is *overwhelming* evidence to infer that the alternative hypothesis is true. We also say that the test is *highly significant*.
- If the p -value lies between 1% and 5%, there is *strong* evidence to infer that the alternative hypothesis is true. The result is deemed to be *significant*.
- If the p -value is between 5% and 10%, we say that there is *weak* evidence to indicate that the alternative hypothesis is true. When the p -value is greater than 5%, we say that the result is *not statistically significant*.
- When the p -value exceeds 10%, we say that there is no evidence to infer that the alternative hypothesis is true.

Figure 12.14 summarises these terms.

statistically significant

There is enough evidence to reject the null hypothesis.

FIGURE 12.14 Describing the p -value of a test



12.3e The p -value method and rejection region method

If we so choose, we can use the p -value to make the same type of decisions we make in the rejection region method. The rejection region method requires the decision maker to select a significance level from which the rejection region is constructed. We then decide to reject or not reject the null hypothesis. Another way of making that type of decision is to compare the p -value with the selected value of the significance level. If the p -value is less than α , we judge the p -value to be small enough to reject the null hypothesis. If the p -value is greater than α , we do not reject the null hypothesis.

Decision rule: Reject H_0 if p -value $< \alpha$, otherwise do not reject H_0 .

We demonstrate this by providing the solution for our opening example described in this chapter's introduction.

SPOTLIGHT ON STATISTICS

SSA envelope plan: Solution

Identifying the technique L05

The objective of the study is to draw a conclusion about the mean payment period. To calculate the p -value, we proceed in the usual way to perform a hypothesis test, except that we do not specify a significance level and a rejection region.



Source: Shutterstock.com/edel

Thus, the parameter to be tested is the population mean μ . We want to know whether there is enough statistical evidence to show that the population mean is less than 22 days. Thus, the alternative hypothesis is

$$H_A: \mu < 22$$

The null hypothesis automatically follows:

$$H_0: \mu = 22$$

The test statistic is

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Calculating manually

To solve this problem manually, we need to define the rejection region, which requires us to specify a significance level. A 10% significance level is deemed to be appropriate. (We will discuss our choice later.)

We wish to reject the null hypothesis in favour of the alternative only if the sample mean and hence the value of the test statistic is small enough. As a result, we locate the rejection region in the left tail of the sampling distribution. To understand why, remember that we are trying to decide whether there is enough statistical evidence to infer that the population mean μ is less than 22 (which is the alternative hypothesis). If we observe a large sample mean (and hence a large value of Z), do we want to reject the null hypothesis in favour of the alternative? The answer is an emphatic *no*. It is illogical to think that if the sample mean is, say, 30, there is enough evidence to conclude that the mean payment period for all customers would be less than 22. Consequently, we want to reject the null hypothesis only if the sample mean (and hence the value of the test statistic Z) is small. How small is small enough? The answer is determined by the significance level and the rejection region. Thus, we set up the rejection region as

$$Z < -z_\alpha = -z_{0.10} = -1.28$$

Note that the direction of the inequality in the rejection region ($Z < -z_\alpha$) matches the direction of the inequality in the alternative hypothesis ($\mu < 22$). Also note the negative sign, since the rejection region is in the left tail (containing values of Z less than 0) of the sampling distribution.

From the data, we calculate the sum and the sample mean. They are

$$\sum x_i = 4759$$

$$\bar{X} = \frac{\sum x_i}{n} = \frac{4759}{220} = 21.63$$

We will assume that the standard deviation of the payment periods for the SSA plan is unchanged from its current value of $\sigma = 6$. The sample size is $n = 220$, and the value of μ is hypothesised to be 22 ($\mu_0 = 22$). We calculate the value of the test statistic under the null hypothesis H_0 as

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{21.63 - 22}{6/\sqrt{220}} = -0.91$$

Because the value of the test statistic, $Z = -0.91$, is not less than -1.28 , we do not reject the null hypothesis in favour of the alternative hypothesis. There is insufficient evidence to infer that the mean is less than 22 days.

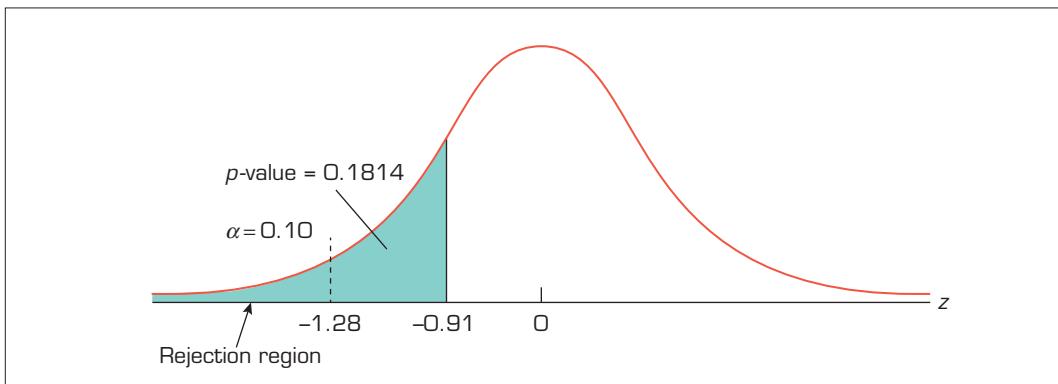
We can determine the p -value of the test:

$$p\text{-value} = P(Z < -0.91) = 0.1814$$

In this type of one-tail (left tail) test of hypothesis, we calculate the *p*-value as $P(Z < z)$ where z is the actual value of the test statistic. As the *p*-value = 0.1814 > 0.10 = α , we do not reject H_0 .

Figure 12.15 depicts the sampling distribution, rejection region and *p*-value.

FIGURE 12.15 Sampling distribution for SSA envelope example



Interpreting the results

The value of the test statistic is -0.91 and its *p*-value is 0.1814, a figure that does not allow us to reject the null hypothesis. Because we were not able to reject the null hypothesis, we say that there is not enough evidence to infer that the mean payment period is less than 22 days. Note that based on the calculated sample mean of 21.63, there was *some* evidence to indicate that the mean of the entire population of payment periods is less than 22 days. However, to reject the null hypothesis we need *enough* statistical evidence, and in this case we simply did not have enough reason to reject the null hypothesis in favour of the alternative. In the absence of evidence to show that the mean payment period for all customers sent an SSA envelope would be less than 22 days, we cannot infer that the plan would be profitable.

Using the computer

The Excel workbook and XLSTAT commands are the same as in Example 12.2. We present the output below.

Using Excel workbook

Excel output for the SSA envelope plan example

	A	B	C	D
1	z-Test of a Mean			
2				
3	Sample mean	21.63	z Stat	-0.91
4	Population standard deviation	6	P(Z<=z) one-tail	0.1802
5	Sample size	220	z Critical one-tail	1.6449
6	Hypothesised mean	22	P(Z<=z) two-tail	0.3604
7	Alpha	0.05	z Critical two-tail	1.9600

Using XLSTAT

	B	C	D	E	F	G
1	Theoretical mean: 22					
2	Significance level (%): 5					
3						
4	Summary statistics:					
5	Variable	Observations	Minimum	Maximum	Mean	Std. deviation
6	Payment	220	9.000	39.000	21.632	5.835
7						
8	One-sample z-test / Two-tailed test:					
9						
10	Difference	-0.368				
11	z (Observed value)	-0.910				
12	z (Critical value)	-1.645				
13	p-value (Two-tailed)	0.181				
14	alpha	0.05				
15						
16	Test interpretation:					
17	HO: The mean is equal to 22.					
18	HA: The mean is lower than 22.					
19	As the computed p-value is greater than the significance level alpha=0.05, one cannot reject the null hypothesis HO.					

A Type I error occurs when we conclude that the plan works when it actually does not. The cost of this mistake is not high. A Type II error occurs when we don't adopt the SSA envelope plan when it would reduce costs. The cost of this mistake can be high. As a consequence, we would like to minimise the probability of a Type II error. Thus we choose a large value for the probability of a Type I error; we set $\alpha = 0.10$.

12.3f Why do we need the *p*-value and the rejection region methods to conduct tests?

The presentation of the *p*-value as another method of conducting statistical tests raises the question, why do we need two criteria for deciding whether or not to reject the null hypothesis? Ideally, all tests should be conducted using *p*-values, because the *p*-value provides more information than does the rejection region method. Unfortunately, we are sometimes unable to determine the *p*-value of a test. The test statistic presented in this chapter is the standard normal distribution for which a table (Table 3 in Appendix B) exists that allows us to calculate the *p*-value. However, in the rest of this book, we deal with other (sampling) distributions, such as *t*, *F*, χ^2 , whose tables presented in Appendix B make calculating the *p*-value quite difficult (if not impossible). If you are using a computer to produce your statistics, the software packages will usually print the *p*-values. However, Excel, like other software packages, does not print the *p*-value for all tests. Consequently, you will have to interpret the value of the test statistic and judge its magnitude relative to some critical value that you will obtain from a table.

There is another reason why we need to understand both methods of testing. You will frequently encounter printed articles in reports and magazines that only print the value of the test statistic. To properly understand the article, you will have to find the rejection region and make your decision by calculating the value of the test statistic and determining whether it falls in the rejection region.

EXAMPLE 12.4

Tips and tax

XM12-04 A significant portion of the incomes of waiters and waitresses is derived from tips. This income, of course, must be reported on income tax forms. Based on historical data, the government tax auditors believe that the average weekly total of tips is \$100. A recently hired tax accountant at the tax department, who formerly worked as a waitress, believes that this figure underestimates the true average of the total weekly tips. As a result, she investigated the weekly total tips of a randomly selected group of 150 waiters and waitresses and found the mean to be \$104. Based on historical data, the population standard deviation is found to be \$22. Test whether there is enough evidence to support the tax accountant's belief.

Solution

Identifying the technique

The objective of this study is to draw conclusions about the population mean based on the *p*-value. The data type is numerical.

To calculate the *p*-value, we do not need the level of significance. We proceed in the usual way to perform the hypothesis test.

Step 1. Null and alternative hypotheses:

Because we want to determine whether there is sufficient evidence to show that the average weekly tip total exceeds \$100, we set up the null and alternative hypotheses as follows:

$$H_0: \mu = 100$$

$$H_A: \mu > 100 \quad (\text{Right one-tail test})$$

Step 2. Test statistic:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

As the sample size $n = 150 > 30$, using the central limit theorem, \bar{X} is approximately normally distributed. Therefore, as the population standard deviation σ is also known, the standardised test statistic Z has a standard normal distribution.

Step 3. Level of significance:

$$\alpha = 0.05$$

Step 4. Decision rule (in terms of *p*-value):

We set up the rejection region as

$$p\text{-value} < \alpha$$

Therefore, the *decision rule* for the test is:

Reject H_0 if *p*-value < $\alpha = 0.05$.

Step 5. Calculating the *p*-value:

Calculating manually

From the data, the sample mean $\bar{X} = 104$. As the population standard deviation is known to be $\sigma = \$22$, the sample size is $n = 150$, and the value of μ is hypothesised to be 100 under H_0 ($\mu_0 = 100$), we calculate the value of the test statistic as

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{104 - 100}{22/\sqrt{150}} = 2.23$$



Therefore,

$$p\text{-value} = P(Z > 2.23) = 0.0129$$

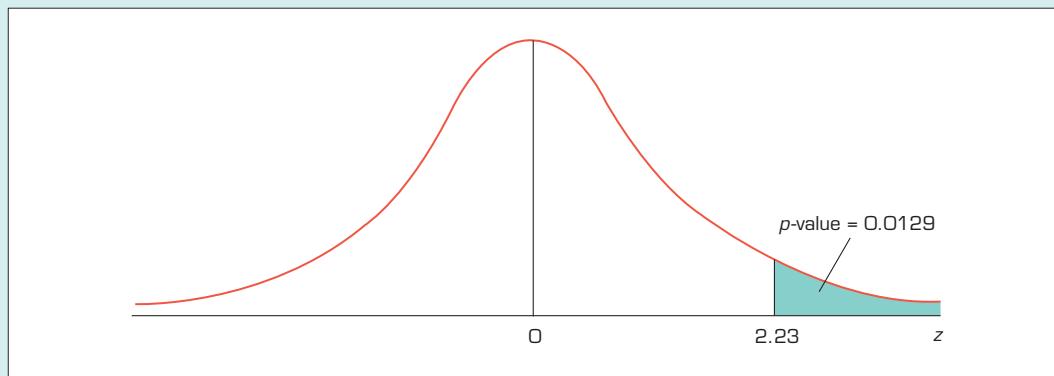
Step 6: Conclusion:

Because the p -value = 0.0129 is less than the level of significance $\alpha = 0.05$, based on the decision rule, we reject the null hypothesis. There is sufficient evidence to infer that the mean weekly total of tips exceeds \$100 at the 5% level of significance.

Interpreting the results

The decision to reject the null hypothesis and conclude that the mean weekly total of tips exceeds \$100 was made at the 5% level of significance. But, it is up to the tax accountant to judge the size of the p -value. If she decides that in conjunction with other factors this p -value is small, she will conclude that the reported weekly total of \$100 underestimates the true total. If the accountant believes that the p -value is large, she will conclude that there is not enough evidence to infer that \$100 is not an accurate estimate of the true total of tips. It is probably safe to say that most people would agree that this p -value is small, and hence, the average amount of money earned by waiters and waitresses in tips exceeds \$100. Once again, we note that the results are dubious if the sampling process is flawed or if the population standard deviation is not equal to \$22. **Figure 12.16** depicts the sampling distribution and the p -value.

FIGURE 12.16 p -value of the test



Using the computer

The Excel commands are the same as in Example 12.2 (page 485).

Using Excel workbook

Excel output

	A	B	C	D
1	z-Test of a Mean			
2				
3	Sample mean	103.9969	z Stat	2.23
4	Population standard deviation	22	P[Z<=z] one-tail	0.0130
5	Sample size	150	z Critical one-tail	1.6449
6	Hypothesised mean	100	P[Z<=z] two-tail	0.0261
7	Alpha	0.05	z Critical two-tail	1.9600

Using XLSTAT

	B	C	D	E	F	G
1	Theoretical mean: 100					
2	Significance level (%): 5					
3						
4	Summary statistics:					
5	Variable	Observations	Minimum	Maximum	Mean	Std. deviation
6	Payment	150	48.100	161.760	103.997	22.469
7						
8	One-sample z-test / Two-tailed test:					
9	Difference	3.997				
10	z (Observed value)	2.225				
11	z (Critical value)	1.645				
12	p-value (Two-tailed)	0.013				
13	alpha	0.05				
14						
15	Test interpretation:					
16	HO: The mean is equal to 100.					
17	HA: The mean is greater than 100.					
18	As the computed p-value is lower than the significance level alpha=0.05, one should reject the null hypothesis HO, and accept the alternative hypothesis Ha.					

The output also reveals that the p -value = 0.013.

EXERCISES

The following exercises can be solved manually or by using Excel's **Test Statistics** workbook that is provided on the companion website (accessible through <https://login.cengagebrain.com>).

12.17 Self-correcting exercise. Determine the p -value in Exercise 12.11.

12.18 Find the p -value of the following tests:

	Hypotheses	Sample value of the test statistic under H_0 , z_0
a	$H_0: \mu = 500$	-1.76
	$H_A: \mu \neq 500$	
b	$H_0: \mu = 200$	2.63
	$H_A: \mu > 200$	
c	$H_0: \mu = 600$	1.75
	$H_A: \mu < 600$	
d	$H_0: \mu = 0$	0.00
	$H_A: \mu > 0$	

12.19 Find the p -value of the following test:

$$H_0: \mu = 25$$

$$H_A: \mu \neq 25$$

$$\bar{X} = 29, \sigma = 15, n = 100$$

12.20 Find the p -value of the test conducted in Exercise 12.15.

12.21 Find the p -value of the test conducted in Exercise 12.16.

12.22 a Calculate the p -value in order to test the hypotheses below:

$$\bar{X} = 52, n = 9, \sigma = 5, \alpha = 0.05$$

$$H_0: \mu = 50$$

$$H_A: \mu > 50$$

b Repeat part (a) with $n = 25$.

c Repeat part (a) with $n = 100$.

d Review parts (a), (b) and (c). Describe what happens to the value of the test statistic and the p -value when the sample size increases.

e Repeat part (a) with $\sigma = 10$.

f Repeat part (a) with $\sigma = 20$.

g Review parts (a), (e) and (f), and discuss what happens to the value of the test statistic and its p -value when the standard deviation increases.

h Repeat part (a) with $\bar{X} = 54$.

i Repeat part (a) with $\bar{X} = 56$.

j Summarise parts (a), (h) and (i) by describing what happens to the value of the test statistic and its p -value when the value of \bar{X} increases.

12.23 a Test the hypotheses below by calculating the p -value given that

$$\bar{X} = 99, n = 100, \sigma = 8, \alpha = 0.05$$

$$H_0: \mu = 100$$

$$H_A: \mu < 100$$

- b** Repeat part (a) with $n = 50$.
- c** Repeat part (a) with $n = 20$.
- d** From parts (a), (b) and (c), discuss the effect on the value of the test statistic and the p -value of the test when the sample size decreases.
- e** Repeat part (a) with $\sigma = 12$.
- f** Repeat part (a) with $\sigma = 15$.
- g** Referring to parts (a), (c) and (f), describe what happens to the value of the test statistic and its p -value when the standard deviation decreases.
- h** Repeat part (a) with $\bar{X} = 98$.
- i** Repeat part (a) with $\bar{X} = 96$.
- j** Summarise parts (a), (h) and (i) by describing what happens to the value of the test statistic and the p -value of the test when the value of \bar{X} decreases.

- 12.24 a** Test these hypotheses by calculating the p -value when $\bar{X} = 21$, $n = 25$, $\sigma = 5$ and $\alpha = 0.05$.

$$H_0: \mu = 20$$

$$H_A: \mu \neq 20$$

- b** Repeat part (a) with $\bar{X} = 22$.
- c** Repeat part (a) with $\bar{X} = 23$.
- d** Describe what happens to the value of the test statistic and its p -value when the value of \bar{X} increases.

- 12.25 a** Test these hypotheses by calculating the p -value given that $\bar{X} = 99$, $n = 100$ and $\sigma = 8$.

$$H_0: \mu = 100$$

$$H_A: \mu \neq 100$$

- b** Repeat part (a) with $n = 50$.
- c** Repeat part (a) with $n = 20$.
- d** What is the effect on the value of the test statistic and the p -value of the test when the sample size decreases?

- 12.26 a** Find the p -value of the following test given that $\bar{X} = 990$, $n = 100$ and $\sigma = 25$.

$$H_0: \mu = 1000$$

$$H_A: \mu < 1000$$

- b** Repeat part (a) with $\sigma = 50$.
- c** Repeat part (a) with $\sigma = 100$.
- d** Describe what happens to the value of the test statistic and its p -value when the standard deviation increases.

- 12.27** Almost everyone who regularly drives a car in Sydney agrees that traffic is getting worse. A randomly selected sample of 50 cars had their speeds measured on a highway during rush hour. The sample mean speed was 60 km/h. Traffic engineers determined that two years ago the mean

and the standard deviation of the speeds on the same freeway during rush hour were 70 and 10 km/h respectively. Find the p -value to determine whether the sample results provide enough statistical evidence to allow the engineers to conclude that freeway traffic has worsened in the last two years.

- 12.28 XR12-28** A random sample of 12 second-year university students enrolled in a business statistics course was drawn. At the course completion, each student was asked how many hours he or she spent doing homework in statistics. The data are listed below. It is known that the population standard deviation is $\sigma = 8.0$. The instructor has recommended that students devote three hours per week for the duration of the 12-week semester, for a total of 36 hours. Test to determine whether there is evidence that the average student spent less than the recommended amount of time. Calculate the p -value of the test.

31	40	26	30	36	38
29	40	38	30	35	38

- 12.29 XR12-29** An office manager believes that the average amount of time spent by office workers reading then deleting spam email exceeds 25 minutes per day. To test this belief, he takes a random sample of 18 workers and measures the amount of time each spends reading and deleting spam. The results are recorded and listed here. If the population of times is normal with a standard deviation of 12 minutes, can the manager infer at the 1% significance level that he is correct?

35	48	29	44	17	21	32	28	34
23	13	9	11	30	42	37	43	48

Computer/manual applications

The following exercises require the use of a computer and software. The answers may be calculated manually using the sample statistics provided.

- 12.30 XR12-30** In an attempt to reduce the number of person-hours lost as a result of industrial accidents, a large production plant installed new safety equipment. In a test of the effectiveness of the equipment, a random sample of 50 departments was chosen. The number of person-hours lost in the month prior to and the month after the installation of the safety equipment was recorded. The percentage change was calculated, and recorded. Assume that the population standard deviation is $\sigma = 5$. Can we infer at the 10% significance level that the new safety equipment is effective?

Sample statistics: $\bar{X} = -1.20$, $n = 50$.

- 12.31 XR12-31** A traffic police officer believes that the average speed of cars travelling over a certain stretch of highway exceeds the posted limit of 110 km/h. The speeds of a random sample of 200 cars were recorded. Do these data provide sufficient evidence at the 1% significance level to support the officer's belief? What is the p -value of the test? (Assume that the population standard deviation is known to be 10.)

Sample statistics: $\bar{X} = 111.6$, $n = 200$.

- 12.32 XR12-32** The golf professional at a private course claims that members who have taken lessons from him have lowered their handicap by more than five strokes. The club manager decides to test the claim by randomly sampling 25 members who have had lessons and asking each to report the reduction in their handicap, where a negative number indicates an increase in the handicap, and recorded the data. Assuming that the reduction in handicap is approximately normally distributed with a standard deviation of two strokes ($\sigma = 2$), test the golf professional's claim using a 10% significance level.

Sample statistics: $\bar{X} = 5.64$, $n = 25$.

- 12.33 XR12-33** The current no-smoking regulations in office buildings require workers who smoke to take breaks and leave the building in order to satisfy their habits. A study indicates that such workers average 32 minutes per day taking smoking breaks. The standard deviation is six minutes ($\sigma = 6$). To help reduce the average break time, break rooms with powerful exhausts were installed in the buildings. To see whether these rooms serve their designed purpose, a random sample of 110 smokers was taken. The total amount of time away from their desks was measured for one day and recorded. Test to determine whether there has been a decrease in the mean time away from their desks. Calculate the p -value and judge it relative to the costs of Type I and Type II errors.

Sample statistics: $\bar{X} = 29.9$, $n = 110$.

- 12.34 XR12-34** A low-handicap golfer who uses Titleist brand golf balls observed that his average drive is 208 metres and the standard deviation 9 metres. Nike has just introduced a new ball. Nike claims that the new ball will travel farther than a Titleist ball. To test the claim the golfer hits 100 drives with a Nike ball and measures the distances. The distances are recorded. Conduct a test to determine whether Nike is correct. Use a 5% significance level.

Sample statistics: $\bar{X} = 209.45$, $n = 100$.

12.4 Testing the population mean μ when the population variance σ^2 is unknown

In Section 12.2 we described the hypothesis test for a rather unlikely situation: testing the population mean when the population variance is known. In this section we progress to the more realistic case in which the population variance is unknown. In this situation we estimate σ by s . As you are about to learn, the only difference between the two cases is the test statistic. In Section 12.2 the test statistic was

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

which has a standard normal (or z) distribution.

Test statistic for μ when σ^2 is unknown

When the population variance is unknown and the population is normally distributed, the test statistic for testing the hypothesis about μ is

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

which is t distributed with $n - 1$ degrees of freedom.

The other five steps in the test are exactly the same as those described in Section 12.2. The following example illustrates this.

EXAMPLE 12.5

LO4

Has production declined due to new government regulations?

XM12-05 A manufacturer of television screens has a production line that used to produce an average of 200 screens per day. Because of recently introduced government regulations, a new safety device is installed, which the manufacturer believes will reduce the average daily output. After installation of the safety device, a random sample of 15 days' production was recorded, as follows:

186	206	190	202	182	210	192	188	202	176	196	188	202	184	190
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Assuming that the daily output is normally distributed, is there sufficient evidence to allow the manufacturer to conclude that the average daily output has decreased following installation of the safety device? (Use $\alpha = 0.05$.)

Solution**Identifying the technique**

The problem objective is to describe the population of daily output. Since we count the number of units produced, the data type is numerical. Thus, the parameter to be tested is the population mean μ .

We specify the hypotheses as before. The alternative hypothesis is set up to answer the question. Since we want to know whether the mean production is now less than 200, we have

$$H_A: \mu < 200$$

Hence, the null hypothesis is

$$H_0: \mu = 200$$

In identifying the test statistic, we note that the population variance is not mentioned in the question, so we assume that it is unknown. The population is assumed to be normally distributed. As a consequence, the test statistic is

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

which has a t distribution with $n - 1$ degrees of freedom.

Calculating manually

From the 15 observations, we calculate

$$n = 15$$

$$\bar{X} = 192.94$$

$$s = 9.70$$

Because of the way the alternative hypothesis is set up, this is a one-tail test. The critical value is

$$-t_{\alpha, n-1} = -t_{0.05, 14} = -1.761$$

Therefore, the decision rule is

Reject H_0 if $t < -1.761$.

The complete test is as follows:

Hypotheses: $H_0: \mu = 200$ $H_A: \mu < 200$ (Left one-tail test)

Test statistic: $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$

Level of significance: $\alpha = 0.05$

Decision rule: Reject H_0 if $t < -1.761$.

Value of the test statistic: $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{192.94 - 200}{9.7/\sqrt{15}} = -2.82$

Conclusion: As $t = -2.82 < -1.761$, reject H_0 .





There is enough evidence to conclude that the mean daily production has decreased since the installation of the safety device.

Interpreting the results

The manufacturer would be advised to look for ways to restore productivity with the safety device in place. Perhaps developing another safety device would help. We note that the results are valid only if the assumption that daily output is normally distributed is true.

Using the computer

If the sample mean and sample standard deviation are already known or calculated, we can use the **t-test_Mean** worksheet in **Test Statistics** workbook to perform the test. The Excel output provides the value of the test statistic and critical values as well as the *p*-values for a one-tail and two-tail tests. The *p*-value method (with decision rule, Reject H_0 if p -value < α) can also be used to derive the conclusion. Both methods would arrive at exactly the same conclusion for a given level of significance.

Excel output for Example 12.5

	A	B	C	D
1	t-Test of a Mean			
2				
3	Sample mean	192.94	t Stat	-2.82
4	Sample standard deviation	9.7	P[T<=t] one-tail	0.0068
5	Sample size	15	t Critical one-tail	1.7613
6	Hypothesised mean	200	P[T<=t] two-tail	0.0137
7	Alpha	0.05	t Critical two-tail	2.1448

COMMANDS

Open the worksheet **t-Test_Mean** available in the **Test Statistics** workbook. Type the values of the sample mean \bar{X} (**192.94**), sample standard deviation s (**9.7**), sample size n (**15**), hypothesised mean μ (**200**) and alpha α (**0.05**).

Using XLSTAT

Alternatively, when raw data are available, we can use XLSTAT to perform this task.

	B	C	D	E	F	G
1	Theoretical mean: 200					
2	Significance level [%]: 5					
3						
4	Summary statistics:					
5	Variable	Observations	Minimum	Maximum	Mean	Std. deviation
6	Payment	15	176.000	210.000	192.933	9.706
7						
8	One-sample z-test / Two-tailed test:					
9	Difference	-7.067				
10	t (Observed value)	-2.820				
11	t (Critical value)	-1.761				
12	DF	14				
13	p-value (one-tailed)	0.007				
14	alpha	0.05				
15						
16	Test interpretation:					
17	HO: The mean is equal to 200.					
18	HA: The mean is lower than 200.					
19	As the computed p-value is lower than the significance level alpha=0.05, one should reject the null hypothesis HO, and accept the alternative hypothesis Ha.					

(Note: This is only a partial output).



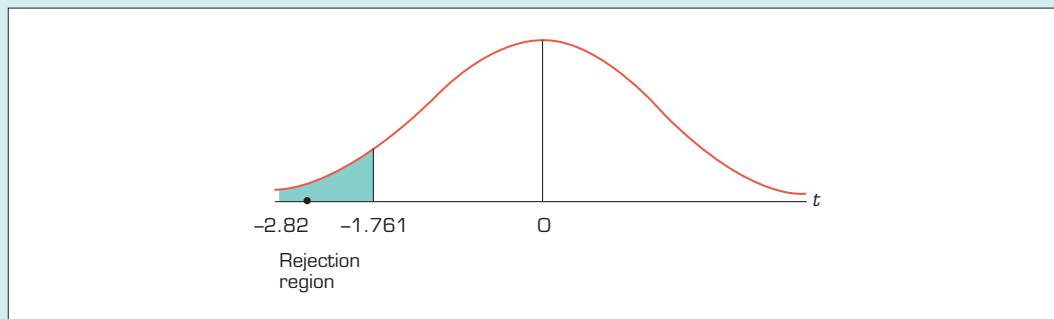


COMMANDS

- 1 Type the data in one column or open the data file (**XM12-05**).
- 2 Click **XLSTAT Parametric Tests** and **One-sample t-test** and **z-test**.
- 3 In the **Data:** dialog box type the input range (**A1:A16**). Click **Column labels** if the first row contains the name of the variable (as in this example). Check **Student's t-test** and do not check z-test.
- 4 Click the **Options** tab and choose **Mean < Theoretical mean** in the **Alternative hypothesis:** box. Type the value under the null hypothesis in the **Theoretical mean:** box (**200**) and the value of α (**5**) in the **Significance:** box. If there are blanks in the column (usually used to represent missing data) click **Missing data, Remove the observations**. Click **OK** and then **Continue**.

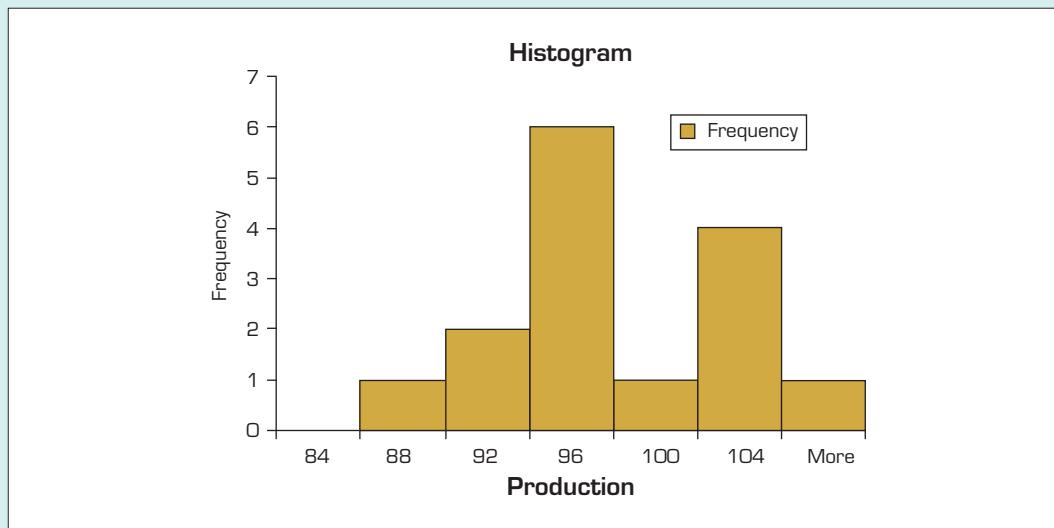
Figure 12.17 describes the sampling distribution of this test.

FIGURE 12.17 Sampling distribution for Example 12.5



Checking the required condition

As before in Chapter 10, we check the required condition that the population is normal. The histogram of the data below indicates that the population is approximately normal. Hence, the t -test and the conclusion is valid.



IN SUMMARY

Factors that identify the *t*-test of μ

- 1 *Problem objective:* to describe a single population
- 2 *Data type:* numerical (quantitative)
- 3 *Descriptive measurement:* central location
- 4 *Population variance:* unknown

EXERCISES

Learning the techniques

12.35 In a random sample of 15 observations from a normal population, we found that $\bar{X} = 150$ and $s = 10$. Using $\alpha = 0.01$, test the hypotheses:

$$H_0: \mu = 160$$

$$H_A: \mu < 160$$

12.36 For each of the following tests of hypotheses about the mean of a normal population, determine whether or not the null hypothesis should be rejected:

a $H_0: \mu = 10\ 000$

$H_A: \mu > 10\ 000$

$n = 10, \bar{X} = 11\ 500, s = 3000, \alpha = 0.05$

b $H_0: \mu = 75$

$H_A: \mu > 75$

$n = 29, \bar{X} = 77, s = 1, \alpha = 0.01$

c $H_0: \mu = 200$

$H_A: \mu < 200$

$n = 25, \bar{X} = 175, s = 50, \alpha = 0.10$

12.37 A random sample of 75 observations from a normal population produced the following statistics:

$\bar{X} = 239.6, s^2 = 1637.5$. Test the following

hypotheses with $\alpha = 0.05$:

$$H_0: \mu = 230$$

$$H_A: \mu \neq 230$$

12.38 Do the following data (drawn from a normal population) allow us to conclude with $\alpha = 0.10$, that $\mu > 7$?

4	8	12	11	14	6	12	8	9	5
---	---	----	----	----	---	----	---	---	---

12.39 A random sample of 10 observations was drawn from a large population and shown below.

7	12	8	4	9	3	4	9	5	2
---	----	---	---	---	---	---	---	---	---

- a Test to determine if we can infer at the 5% significance level that the population mean is not equal to 5.
- b What is the required condition of the technique used in part (a)?

The following problems can be solved manually or by using Excel's Test Statistics workbook that is provided on the companion website (accessible through <https://login.cengagebrain.com>).

12.40 a The sample mean and standard deviation for a random sample of 20 observations from a normal population were calculated as $\bar{X} = 23$ and $s = 9$. Calculate the *t*-statistic (and for Excel users, the *p*-value) of the test required to determine whether there is enough evidence to infer at the 5% significance level that the population mean is greater than 20.

b Repeat part (a) with $n = 10$.

c Repeat part (a) with $n = 50$.

d Refer to parts (a), (b) and (c). Describe the effect on the *t*-statistic (and for Excel users, the *p*-value) of increasing the sample size.

e Repeat part (a) with $s = 5$.

f Repeat part (a) with $s = 20$.

g Refer to parts (a), (e) and (f). Discuss what happens to the *t*-statistic (and for Excel users, the *p*-value) when the standard deviation decreases.

h Repeat part (a) with $\bar{X} = 21$.

i Repeat part (a) with $\bar{X} = 26$.

j Review the results of parts (a), (h) and (i). What happens to the *t*-statistic (and for Excel users, the *p*-value) when the sample mean increases?

12.41 a A random sample of eight observations was taken from a normal population. The sample mean and standard deviation are $\bar{X} = 75$ and $s = 50$. Can we infer at the 10% significance level that the population mean is less than 100?

b Repeat part (a) assuming that you know that the population standard deviation is $\sigma = 50$.

c Review parts (a) and (b). Explain why the test statistics differed.

12.42 a A statistics practitioner is in the process of testing to determine whether there is enough evidence to infer that the population mean is different from 180. She calculated the mean and standard deviation of a sample of 200 observations as $\bar{X} = 175$ and $s = 22$. Calculate the value of the test statistic (and for Excel users, the p -value) of the test required to determine whether there is enough evidence at the 5% significance level.

- b** Repeat part (a) with $s = 45$.
- c** Repeat part (a) with $s = 60$.
- d** Discuss what happens to the t -statistic (and for Excel users, the p -value) when the standard deviation increases.

12.43 a Calculate the test statistic (and for Excel users, the p -value) when $\bar{X} = 145$, $s = 50$ and $n = 100$. Use a 5% significance level.

$$H_0: \mu = 150$$

$$H_A: \mu < 150$$

- b** Repeat part (a) with $\bar{X} = 140$.
- c** Repeat part (a) with $\bar{X} = 135$.
- d** What happens to the t -statistic (and for Excel users, the p -value) when the sample mean decreases?

12.44 a A random sample of 25 observations was drawn from a normal population. The sample mean and sample standard deviation are $\bar{X} = 52$ and $s = 15$. Calculate the test statistic (and for Excel users, the p -value) of a test to determine whether there is enough evidence at the 10% significance level to infer that the population mean is not equal to 50.

- b** Repeat part (a) with $n = 15$.
- c** Repeat part (a) with $n = 5$.
- d** Discuss what happens to the t -statistic (and for Excel users, the p -value) when the sample size decreases.

12.45 a To test the following hypotheses, a statistics practitioner randomly sampled 100 observations and found $\bar{X} = 106$ and $s = 35$. Calculate the test statistic (and for Excel users, the p -value) of a test to determine whether there is enough evidence at the 1% significance level to infer that the alternative hypothesis is true.

$$H_0: \mu = 100$$

$$H_A: \mu > 100$$

- b** Repeat part (a) with $s = 25$.
- c** Repeat part (a) with $s = 15$.

d Discuss what happens to the t -statistic (and for Excel users, the p -value) when the standard deviation decreases.

12.46 a A random sample of 11 observations was taken from a normal population. The sample mean and standard deviation are $\bar{X} = 74.5$ and $s = 9$. Can we infer at the 5% significance level that the population mean is greater than 70?

- b** Repeat part (a) assuming that you know that the population standard deviation is $\sigma = 9$.
- c** Explain why the conclusions produced in parts (a) and (b) differ.

Applying the techniques

12.47 Self-correcting exercise. A doctor claims that the average Australian is more than 5 kg overweight. To test this claim, a random sample of 50 Australians were weighed, and the difference between their actual weight and their ideal weight was calculated. The mean and the standard deviation of that difference were 6.5 and 2.2 kg respectively. Can we conclude with $\alpha = 0.05$, that enough evidence exists to show that the doctor's claim is true?

12.48 Ecologists have long advocated recycling newspapers as a way of saving trees and reducing landfills. A number of companies have gone into the business of collecting used newspapers from households and recycling them. A financial analyst for one such company has recently calculated that the firm would make a profit if the mean weekly newspaper collection from each household exceeded 1 kg. In a study to determine the feasibility of a recycling plant, a random sample of 100 households showed that the mean and the standard deviation of the weekly weight of newspapers discarded for recycling are

$$\bar{X} = 1.1 \text{ kg}, s = 0.35 \text{ kg}$$

Do these data provide sufficient evidence at the 1% significance level to allow the analyst to conclude that the recycling plant would be profitable?

12.49 A courier service in Brisbane advertises that its average delivery time is less than six hours for local deliveries. A random sample of the amount of time this courier takes to deliver packages to an address across town produced the following delivery times (rounded to the nearest hour):

7	3	4	6	10	5	6	4	3	8
---	---	---	---	----	---	---	---	---	---

- a** Is this sufficient evidence to support the courier's advertisement, at the 5% level of significance?

- b** What assumption must be made in order to answer part (a)?

12.50 One of the critical factors in choosing a location for a new men's clothing store is the mean clothing expenditure per household in the surrounding neighbourhood. A survey of 20 households reveals that the mean and the standard deviation of annual expenditure on clothes are \$387 and \$60 respectively. Can we conclude at the 5% significance level that the population mean annual expenditure is less than \$400?

Computer/manual applications

The following exercises require the use of a computer and software. The answers may be calculated manually using the sample statistics provided.

12.51 XR12-51 The following observations were drawn from a large population.

22	18	25	28	19	20	24	26	19	26	27	22	23
25	25	18	20	26	18	26	27	24	20	19	18	

- a** Test to determine if we can infer at the 10% significance level that the population mean is greater than 20.
b What is the required condition of the techniques used in part (a)? Use a graphical technique to check that required condition is satisfied.

Sample statistics: $\bar{X} = 22.6$, $s = 3.416$, $n = 25$.

12.52 XR12-52 A random sample of 75 observations obtained from a normal population is recorded. Test the following hypotheses with $\alpha = 0.05$.

$$\begin{aligned} H_0: \mu &= 103 \\ H_A: \mu &\neq 103 \end{aligned}$$

Sample statistics: $\bar{X} = 99.45$, $s = 21.25$, $n = 75$.

12.53 XR12-53 A diet doctor claims that the average Australian is more than 10 kg overweight. To test

his claim, a random sample of 100 Australians were weighed, and the difference between their actual weight and their ideal weight was calculated and recorded. Some of the data are listed below.

16	4	4	4.5	...	8.5	16.5	17
----	---	---	-----	-----	-----	------	----

Do these data allow us to infer at the 5% significance level that the doctor's claim is true?

Sample statistics: $\bar{X} = 12.175$, $s = 7.9$, $n = 100$.

12.54 XR12-54 A pizza outlet advertises that its average waiting time is less than 12 minutes from the time an order is placed. A random sample of waiting times (in minutes) for 50 orders was recorded. A portion of this sample is shown below.

10.4	10.8	10.0	13.2	...	14.8	8.8	8.2
------	------	------	------	-----	------	-----	-----

- a** Is this sufficient evidence to support the pizza outlet's advertisement, at the 5% level of significance?
b What assumption must be made in order to answer part (a)? Use whatever graphical technique you deem appropriate to confirm that the required condition is satisfied.

Sample statistics: $\bar{X} = 11.74$, $s = 2.04$, $n = 50$.

12.55 XR12-55 Companies that sell groceries over the internet are called e-grocers. Customers enter their orders, pay by credit card, and receive delivery by truck. A potential e-grocer analysed the market and determined that to be profitable the average order would have to exceed \$85. To determine whether an e-grocery would be profitable in one large city, the service was offered to a random sample of customers and the size of the orders recorded. Can we infer from these data that an e-grocery will be profitable in this city?

Sample statistics: $\bar{X} = 89.27$, $s = 17.30$, $n = 85$.

12.5 Calculating the probability of a Type II error

As you have seen, to properly interpret the results of a test of hypothesis requires that you be able to judge the p -value of the test. However, to do so also requires that you have an understanding of the relationship between Type I and Type II errors. In this section, we describe how the probability of a Type II error is calculated.

Recall Example 12.1, in which we conducted the test using the sample mean as the test statistic and we calculated the rejection region as

$$\bar{X} < 3.752 \quad \text{or} \quad \bar{X} > 4.248$$

Thus, if \bar{X} falls between 3.752 and 4.248, we will not reject the null hypothesis. A Type II error occurs when a false null hypothesis is not rejected. Therefore, if the null hypothesis is false, the probability of a Type II error is defined as

$$\beta = P(3.752 < \bar{X} < 4.248 \text{ given that } H_0 \text{ is false})$$

The condition that the null hypothesis is false only tells us that the mean is not equal to 4. If we want to calculate β , we need to specify a value for μ . Suppose that we want to determine the probability of making a Type II error when, in actual fact, $\mu = 4.44$. That is,

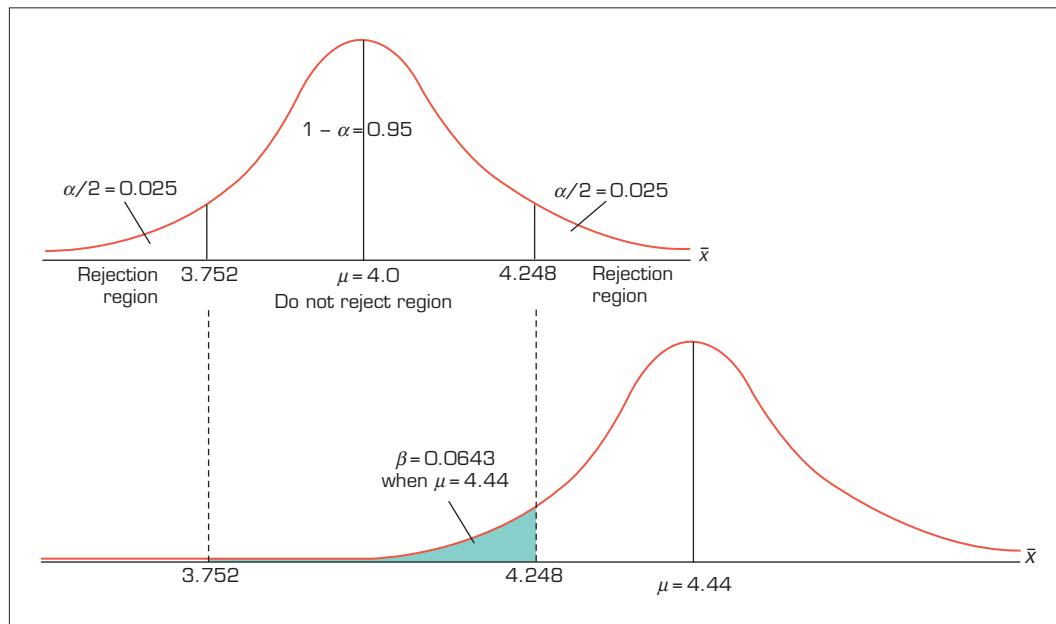
$$\beta = P(3.752 < \bar{X} < 4.248 \text{ given that } \mu = 4.44)$$

We know that \bar{X} is normally distributed with mean μ and standard deviation σ/\sqrt{n} . To proceed, we standardise \bar{X} and use Table 3 in Appendix B as follows:

$$\begin{aligned}\beta &= P\left(\frac{3.752 - 4.44}{0.4/\sqrt{10}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{4.248 - 4.44}{0.4/\sqrt{10}}\right) \\ &= P(-5.44 < Z < -1.52) = 0.0643\end{aligned}$$

This means that, if μ is actually equal to 4.44, the probability of incorrectly not rejecting the null hypothesis is 0.0643. **Figure 12.18** graphically represents the calculation of β . Notice that, in order to calculate the probability of a Type II error, we had to express the rejection region in terms of the unstandardised test statistic \bar{X} , and we had to specify a value of μ other than the one shown in the null hypothesis. The one we used above was arbitrarily selected. In a practical setting, we would choose a value of interest to us. The following example illustrates how to choose that value and how to calculate β .

FIGURE 12.18 Calculation of β for Example 12.1



12.5a Effect on β of changing α

Suppose that in the previous illustration we had used a significance level of 1% instead of 5%. The rejection region expressed in terms of the standardised test statistic would be

$$-2.575 < Z < 2.575$$

or

$$-2.575 < \frac{\bar{X} - 4.0}{0.4/\sqrt{10}} < 2.575$$

Solving for \bar{X} , we find the rejection region in terms of the unstandardised test statistic:

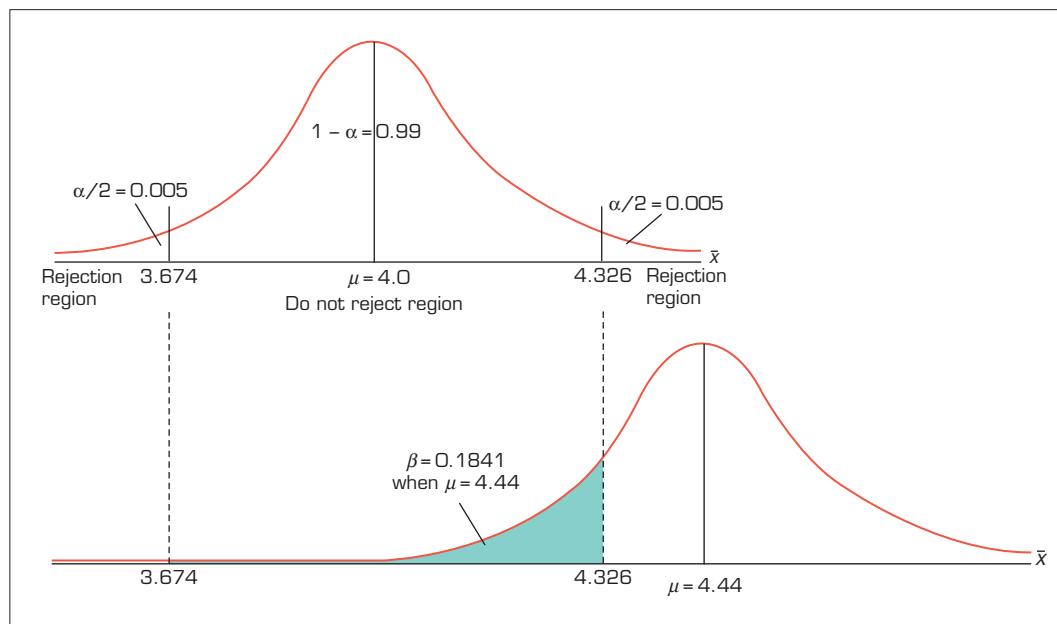
$$3.674 < \bar{X} < 4.326$$

The probability of a Type II error when $\mu = 4.44$ is

$$\begin{aligned}\beta &= p\left(\frac{3.674 - 4.44}{0.4/\sqrt{10}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{4.326 - 4.44}{0.4/\sqrt{10}}\right) \\ &= p(-6.06 < Z < -0.90) = 0.1841\end{aligned}$$

Figure 12.19 depicts this calculation. Compare this figure with **Figure 12.18**. As you can see, by decreasing the significance level from 5% to 1%, we have shifted the critical value of the rejection region to the right and thus enlarged the area where the null hypothesis is not rejected. The probability of a Type II error increases from 0.0643 to 0.1841.

FIGURE 12.19 Calculating β for $\mu = 4.44$, $\alpha = 0.01$ and $n = 10$



The above calculation of β illustrates the inverse relationship between the probabilities of Type I and Type II errors alluded to in Section 12.1. It is important to understand this relationship. From a practical point of view, it tells us that if you want to decrease the probability of a Type I error (by specifying a small value of α), you increase the probability of a Type II error. In applications where the cost of a Type I error is considerably larger than the cost of a Type II error, this is appropriate. In fact a significance level of 1% or less is probably justified. However, when the cost of a Type II error is relatively large, a significance level of 5% or more may be appropriate.

Unfortunately, there is no simple formula to determine what the significance level should be. It is necessary for the manager to consider the costs of both mistakes in deciding what to do. Judgement and knowledge of the factors in the decision are crucial.

12.5b Judging the test

There is another important concept to be derived from this section. A statistical test of hypothesis is effectively defined by the significance level and the sample size, both of which are selected by the statistics practitioner. We can judge how well the test functions by calculating the probability of a Type II error at some value of the parameter. To illustrate, consider Example 12.6.

EXAMPLE 12.6

LO7

Cost of making a wrong decision

The feasibility of constructing a profitable electricity-producing wind turbine depends on the average velocity of the wind. For a certain type of wind turbine, the average wind speed would have to exceed 32 km/h in order for construction of that turbine to be feasible. To test whether or not a particular site is appropriate for this wind turbine, 50 readings of the wind velocity are taken, and the average calculated. The test is designed to answer the question: Is the site feasible? That is, is there sufficient evidence to conclude that the average wind velocity exceeds 32 km/h? As a result, we wish to test the following hypotheses:

$$H_0: \mu = 32$$

$$H_A: \mu > 32 \quad (\text{Right one-tail test})$$

If, when the test is conducted, a Type I error is committed (rejecting H_0 when it is true), we would conclude mistakenly that the average wind velocity exceeds 32 km/h. The consequence of this decision is that the wind turbine would be built on an inappropriate site. Since this error is quite costly, we set $\alpha = 0.01$. If a Type II error is committed (not rejecting H_0 when it is false), we would conclude mistakenly that the wind velocity does not exceed 32 km/h. As a result, we would not build the wind turbine on that site, even though the site is a good one. The cost of this error is not very large since, if the site under consideration is judged to be inappropriate, the search for a good site would simply continue. But suppose that a site where the wind velocity is greater than or equal to 40 km/h is extremely profitable. To judge the effectiveness of this test (to determine if our selection of $\alpha = 0.01$ and $n = 50$ is appropriate), we calculate the probability of committing this error. Our task is to calculate β when $\mu = 40$. (Assume that we know that $\sigma = 19.2$ km/h.)

Solution**Identifying the technique**

Our first step is to set up the rejection region in terms of \bar{X} . Since the rejection region for $\alpha = 0.01$ is

$$Z > z_\alpha = 2.33$$

we have

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - 32}{19.2/\sqrt{50}} > 2.33$$

Solving the inequality, we express the rejection region in terms of the sample statistic \bar{X} as

$$\bar{X} > 38.33$$

The second step is to describe the region where H_0 is not rejected as

$$\bar{X} \leq 38.33$$

Calculating the Type II error manually

Since the 'do not' region is $\bar{X} \leq 38.33$, we have

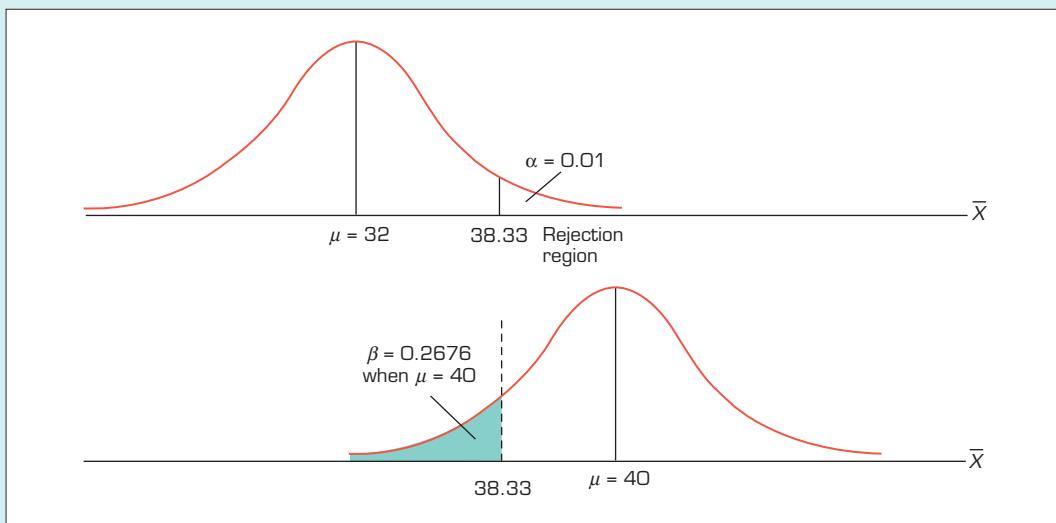
$$\begin{aligned} \beta &= P(\bar{X} \leq 38.33 | \mu = 40) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{38.33 - 40}{19.2/\sqrt{50}}\right) \\ &= P(Z \leq -0.62) \\ &= 0.2676 \end{aligned}$$

Thus the probability of making a Type II error when the true mean $\mu = 40$ is 26.76%.

Interpreting the results

The probability of accepting H_0 when $\mu = 40$ is 0.2676 (see **Figure 12.20**). This means that when the mean wind velocity is 40 km/h, there is a 26.76% probability of erroneously concluding that the site is not profitable. If this probability is considered too large, we can reduce it by either increasing α or increasing n .



FIGURE 12.20 Calculation of β for Example 12.6: $\mu = 40$, $\alpha = 0.01$ and $n = 50$ 

For example, if we increase α to 0.10 and leave $n = 50$, then $\beta = 0.0475$. With $\alpha = 0.10$, however, the probability of building on a site that is not profitable is too large. If we let $\alpha = 0.01$ but increase n to 100, then $\beta = 0.0336$. Now both α and β are quite small, but the cost of sampling has increased. Nonetheless, the cost of sampling is quite small in comparison to the costs of making Type I and Type II errors in this situation.

Using the computer

We have made it possible to utilise Excel to calculate β for any test of hypothesis using the workbooks.

	A	B	C	D
1	Type II Error			
2				
3	HO: MU	32	Critical value	38.32
4	Sigma	19.2	Prob(Type II error)	0.2677
5	Sample size	50	Power of the test	0.7323
6	Alpha	0.01		
7	HA: MU	40		

COMMANDS

Open the **Beta-mean** workbook. Within it are three worksheets: Right-tail test, Left-tail test and Two-tail test. Find the appropriate worksheet for the test of hypothesis you are analysing and type values for μ (under the null hypothesis), σ , n and μ (actual value under the alternative hypothesis).

The accompanying output was produced by selecting the **Right-tail test** worksheet and substituting μ_0 (under the null hypothesis) (32), σ (19.2), n (50), α (0.01) and μ (under the alternative hypothesis) (40).

You can use the left-tail test worksheet to calculate the probability of Type II errors when the alternative hypothesis states that the mean is less than a specified value (e.g. the SSA envelope plan opening example). The two-tail test worksheet is used to calculate β for two-tail tests (e.g. Example 12.1).

In this section we have discussed calculating the probability of a Type II error only when testing μ with σ^2 known, because only the standard normal table (Table 3 in Appendix B) allows us to perform this calculation. In the following section you will encounter another test statistic that is approximately normally distributed; and, in that case too, you will be able to calculate β .

12.5c Power of a test

Another way of judging a test is to measure its **power** – the probability of its leading us to reject H_0 when it is false, rather than measuring the probability of a Type II error, which is the probability of not rejecting H_0 when it is false.

power

The probability of correctly rejecting a false null hypothesis.

$$\begin{aligned} \text{Power} &= P(\text{Reject } H_0 | H_0 \text{ is false}) \\ &= 1 - P(\text{do not reject } H_0 | H_0 \text{ is false}) \end{aligned}$$

Thus, the power of the test is equal to $(1 - \beta)$. In the present example, the power of the test with $n = 50$ and $\alpha = 0.01$ is $1 - 0.2676 = 0.7324$ when $\mu = 40$.

When more than one test can be performed in a given situation, we would naturally prefer to use the one that is correct more frequently. If (given the same alternative hypothesis, sample size and significance level) one test has a higher power than another test, the first test is said to be *more powerful*.

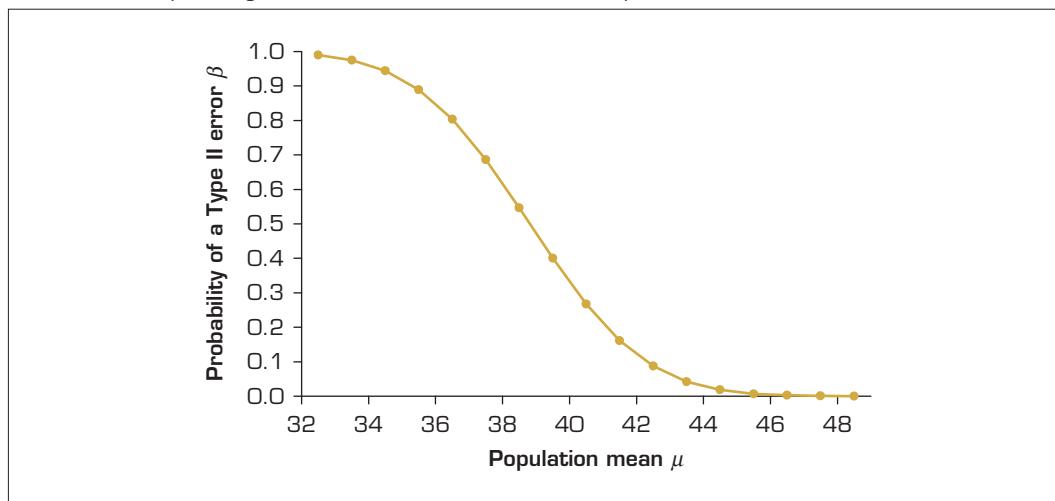
12.5d Operating characteristic curve and the power curve

To compute the probability of a Type II error, we must specify the significance level, the sample size, and an alternative value of the population mean. One way to keep track of all these components is to draw the **operating characteristic (OC) curve**, which plots the values of β versus the values of μ . Because of the time-consuming nature of these calculations, the computer is a necessity. To illustrate, we will draw the OC curve for Example 12.6. We used Excel to compute the probability of Type II error in Example 12.6 for $\mu = 32, 33, \dots, 48$, with $n = 50$. **Figure 12.21** depicts this curve. Notice that as the alternative value of μ increases, the value of β decreases. This tells us that, as the alternative value of μ moves further from μ under the null hypothesis, the probability of a Type II error decreases. In other words, it becomes easier to distinguish between $\mu = 32$ and other values of μ when μ is further from 32. Note that when $\mu = 32$ (the hypothetical value of μ), $\beta = 1 - \alpha$.

operating characteristic (OC) curve

A plot of the probability of making Type II error against the values of the parameter.

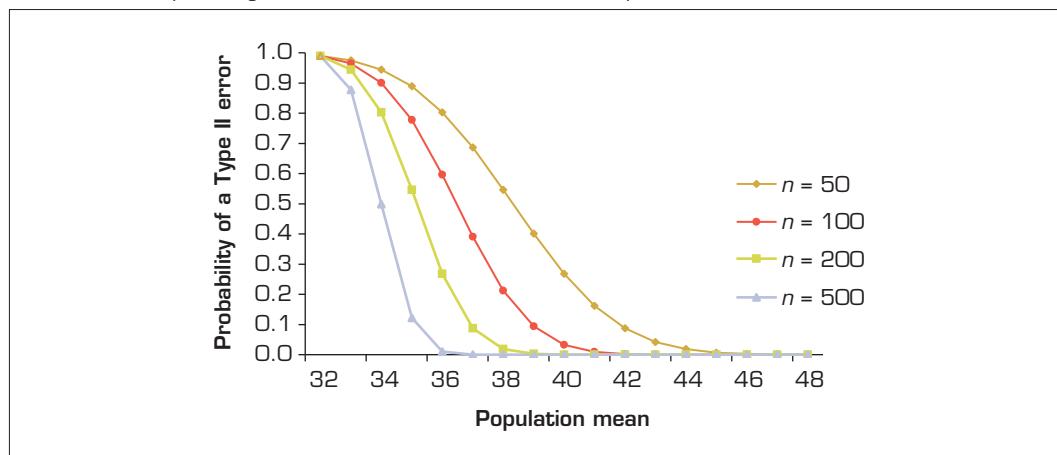
FIGURE 12.21 Operating characteristic (OC) curve for Example 12.6



The OC curve can also be useful in selecting a sample size. **Figure 12.22** shows the OC curve for Example 12.6 with $n = 50, 100, 200$ and 500 . An examination of the chart sheds some light concerning the effect that increasing the sample size has on how well the test performs at different values of μ . For example, we can see that smaller sample sizes will work well to distinguish between 32 and values of μ larger than 40. However, to distinguish between 32 and smaller values of μ requires larger sample sizes. Although the information is

imprecise, it does allow us to select a sample size that is suitable for our purpose. The graph that plots the values of $(1 - \beta)$ versus the values of μ is called the *power curve*.

FIGURE 12.22 Operating characteristic (OC) curve for Example 12.6 for $n = 50, 100, 200$ and 500



EXERCISES

Learning the techniques

- 12.56** For the test of hypothesis, the following apply:

$$H_0: \mu = 1000$$

$$H_A: \mu \neq 1000$$

$$\alpha = 0.05, \sigma = 200, n = 100$$

Find the probability of a Type II error, β , when $\mu = 900, 940, 980, 1020, 1060, 1100$.

- 12.57** For Exercise 12.56, graph the operating characteristic curve with μ on the horizontal axis and β on the vertical axis. If necessary, calculate β for additional values of μ . Also graph the power curve μ versus $1 - \beta$.

- 12.58** Repeat Exercises 12.56 and 12.57 with $n = 25$. What do you notice about these graphs when compared with the graphs drawn in Exercise 12.57?

- 12.59** A statistics practitioner wants to test the following hypotheses with $\sigma = 20$ and $n = 100$.

$$H_0: \mu = 100$$

$$H_A: \mu > 100$$

- a Using $\alpha = 0.10$, find the probability of a Type II error, β , when $\mu = 102$.
- b Repeat part (a) with $\alpha = 0.02$.
- c Describe the effect on β of decreasing α .

- 12.60 a** Calculate the probability of a Type II error for the following hypotheses when $\mu = 37$.

$$H_0: \mu = 40$$

$$H_A: \mu < 40$$

The significance level is $\alpha = 0.05$, the population standard deviation is 5, and the sample size is 25.

- b Repeat part (a), with $\alpha = 0.15$.
- c Describe the effect on β of increasing α .

- 12.61 a** Find the probability of a Type II error for the following test of hypothesis, given that $\mu = 196$.

$$H_0: \mu = 200$$

$$H_A: \mu < 200$$

The significance level is $\alpha = 0.10$, the population standard deviation is 30, and the sample size is 25.

- b Repeat part (a) with $n = 100$.
- c Describe the effect on μ of increasing n .

Applying the techniques

- 12.62 Self-correcting exercise.** For Exercise 12.11, determine the probability of concluding that the average TE score did not increase when the population mean score μ actually equals 930.

- 12.63** Suppose we want to test a null hypothesis that the mean of a population is 145 against an alternative hypothesis that the mean is less than 145. A sample of 100 measurements drawn from the population (whose standard deviation is 20) yields a mean of 140. If the probability of a Type I error is chosen to be 0.05, calculate the probability of a Type II error, assuming that the true population mean equals 142.

12.6 Testing the population proportion p

As you have seen in Chapter 9 and in Section 10.4, we cannot calculate the mean when the data type is nominal. Instead, the parameter of interest is the population proportion p . The point estimator of this parameter is the sample proportion \hat{p} , which under some rather reasonable conditions has an approximately normal sampling distribution.

Sampling distribution of \hat{p}

The sample proportion \hat{p} is approximately normally distributed, with mean p and the standard deviation $\sqrt{pq/n}$ provided that n is large ($np \geq 5$ and $nq \geq 5$).

12.6a Testing hypothesis about a population proportion p

The test statistic is based on the point estimator of the parameter to be tested. As the sample proportion \hat{p} is a point estimator of the population proportion p , \hat{p} or its standardised value will be used as the test statistic to test hypotheses about the population proportion p .

Test statistic for a population proportion

The test statistic used to test a population proportion is

$$Z = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

In order for this test statistic to be valid, the sample size must be large enough so that $np \geq 5$ and $nq \geq 5$. As was the case with previous tests of hypothesis, the value of the parameter in the test statistic is that specified in the null hypothesis. In this test statistic, the parameter p appears in both the numerator and the denominator; remember too, that $q = 1 - p$.

EXAMPLE 12.7

LO8

Is the market share large enough to introduce a new product?

XM12-07 After careful analysis, a company contemplating the introduction of a new product has determined that it must capture a market share of 10% to break even. Anything greater than 10% will result in a profit for the company. In a survey, 400 potential customers are asked whether or not they would purchase the product and their responses were recorded. If 52 people respond affirmatively, is this enough evidence to enable the company to conclude that the product will produce a profit? (Use $\alpha = 0.05$.)

Solution

Identifying the technique

The problem objective is to describe the population of shoppers. The data type is nominal, since the possible responses to the survey questions are 'Yes, I would purchase this product' and 'No, I would not purchase this product'. The parameter of interest is the population proportion p . We want to know if there is enough evidence to allow us to conclude that the company will make a profit, so

$$H_A: p > 0.10$$



Using the z method

As before we follow the six steps. The complete test is as follows:

Hypotheses: $H_0: p = 0.10$

$H_A: p > 0.10$ (Right one-tail test)

$$\text{Test statistic: } Z = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

is standard normally distributed as $np \geq 5$ and $nq \geq 5$.

Significance level: $\alpha = 0.05$

Decision rule: Critical value $z_{0.05} = 1.645$. Reject H_0 if $Z > 1.645$.

Value of the test statistic:

Calculating manually

$$\hat{p} = 52/400 = 0.13; \hat{q} = 1 - 0.13 = 0.87$$

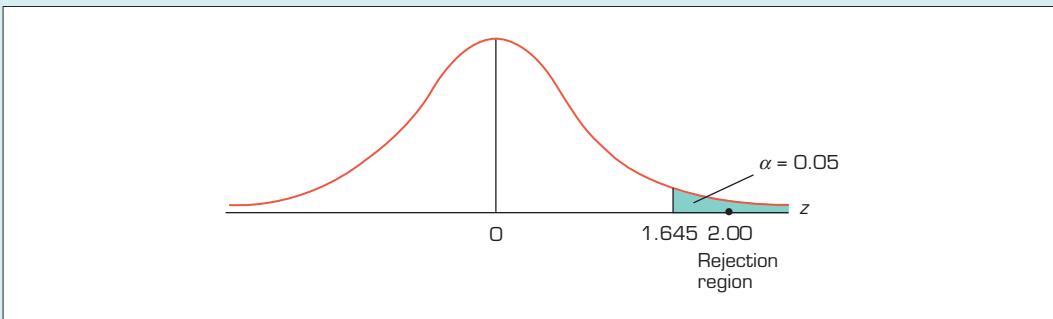
$$Z = \frac{0.13 - 0.10}{\sqrt{(0.10)(0.90)/(400)}} = 2.00$$

Conclusion: As $Z = 2.00 > 1.645$, we reject H_0 .

Consequently, there is enough evidence to allow us to conclude that the product will contribute a profit to the company.

Observe that, based on the sample value of $\hat{p} = 0.13$, we did find some evidence to support a conclusion that the population proportion p is greater than 0.10. Our test results formally confirm this. **Figure 12.23** describes this test.

FIGURE 12.23 Sampling distribution for Example 12.7



Using the p -value method

In Section 12.3 we pointed out that the p -value of a test is the smallest value of α that leads to rejection of the null hypothesis. As a result, a small p -value allows the statistics practitioner to conclude that there is enough evidence to justify rejecting the null hypothesis in favour of the alternative hypothesis.

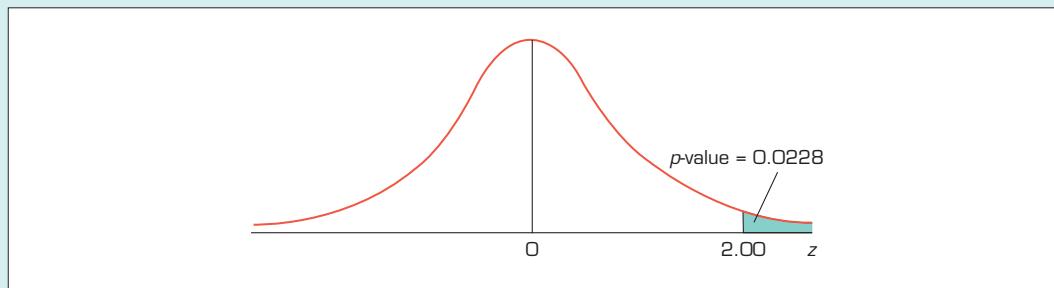
Calculating manually

The p -value of a test of hypothesis about a population proportion is determined in the same way as was shown in Section 12.3. For this example we find

$$p\text{-value} = P(Z > 2.00) = 0.0228$$

Thus, if the decision maker chooses a significance level larger than 0.0228, the p -value will be considered small, and H_0 will be rejected. If a significance level less than 0.0228 is chosen, the p -value will be considered large, and the null hypothesis will not be rejected. In this example, $\alpha = 0.05$, which is greater than 0.0228, therefore H_0 is rejected.

Figure 12.24 describes the sampling distribution, the test statistic and its p -value.

FIGURE 12.24 Sampling distribution for Example 12.7

Interpreting the results

A p -value of 2.28% provides strong evidence to infer that the proportion of customers who would purchase the product exceeds 10%. However, before introducing the new product, the company should make certain that the market analysis was properly performed. Was the sample of 400 potential customers randomly selected? If a substantial number of potential customers who were asked refused to respond, then the sample may be partially self-selected, which may invalidate the conclusion. Did all of the respondents understand the operation of the new product? If the description was unclear, some of the responses are useless. However, if the company is satisfied with the market analysis, the new product should be introduced.

Using the computer

Most statistical software packages do not perform tests of hypothesis about p . However, it is often the case that the survey responses are stored on computer files. For example, suppose that in Example 12.7 the data are stored using the following codes:

0 = No, I would not purchase the product.

1 = Yes, I would purchase the product.

We could use the computer to count the number of each type of response. From these results, we can calculate \hat{p} and the value of the test statistic (as we did above). (For the Excel commands, see Example 10.4, pages 406–08.)

We can also instruct the computer to calculate the value of the test statistic from raw data. Because we frequently encounter this type of problem, we created an Excel workbook to output the required statistics.

Using Excel workbook

If the sample mean and sample standard deviation are already known or calculated, we can use the **z-test_Proportion** worksheet in **Test Statistics** workbook to perform the test.

Excel output for Example 12.7

	A	B	C	D
1	z-Test of a Proportion			
2				
3	Sample Proportion	0.130	z Stat	2.00
4	Sample size	400	$P(Z \leq z)$ one-tail	0.0228
5	Hypothesized proportion	0.10	z Critical one-tail	1.6449
6	Alpha	0.05	$P(Z \leq z)$ two-tail	0.0455
7			z Critical two-tail	1.9600

COMMANDS

Open the worksheet **z-Test_Proportion** available in the **Test Statistics** workbook. Calculate the sample proportion \hat{p} from the raw data. Type the values of the sample proportion (**0.13**), sample size n (**400**), hypothesised proportion p (**0.1**) and alpha α (**0.05**).

Using XLSTAT

Alternatively, when raw data are available, we can use XLSTAT, an EXCEL Add-in, to perform this task.

	B	C	D	E	F	G
1	Proportion: 0.13					
2	Sample size: 400					
3	Test proportion: 0.1					
4	Hypothesized difference (D): 0					
5	Significance level (%): 5					
6						
7	z-test for one proportion / Upper-tailed test:					
8	Difference	0.030				
9	z (Observed value)	2.000				
10	z (Critical value)	1.645				
11	p-value (one-tailed)	0.023				
12	alpha	0.05				
13						
14	Test interpretation:					
15	HO: The difference between the proportions is equal to 0.					
16	HA: The difference between the proportions is greater than 0.					
17	As the computed p-value is lower than the significance level alpha=0.05, one should reject the null hypothesis HO, and accept the alternative hypothesis Ha.					

(Note: This is only a partial output).

COMMANDS

- Type the data in one column or open the data file (**XM12-07**). In any empty cell, calculate the number of 'successes' (**=COUNTIF(A2:A401,1)**). Divide that number by the sample size (**400**) to obtain the sample proportion.
- Click **XLSTAT**, **Parametric tests**, and **Tests for one proportion**.
Author's note: We find the XLSTAT terminology confusing. However, these instructions will produce the correct result.
- Type the sample **Proportion: (0.13)**, the **Sample size: (400)**, and the value of p under the null hypothesis – **Test proportion: (0.1)**. Under **Data format**: check **Proportion**. Click **z test**.
- Click the **Options** tab and choose **Proportion – Test proportion > D**. Type the **Hypothesized difference (D): (0)** and type the **Significance level (%) (5)**.

IN SUMMARY

Factors that identify the test of p

- Problem objective:** to describe a single population
- Data type:** nominal (categorical)

12.6b Probability of a Type II error involving p

In Example 12.7 we specified that the probability of a Type I error was 0.05, meaning that there was a 5% chance we might conclude that the product would be profitable when in fact it would not. If we are interested in determining the probability of our concluding that the product would not be profitable when in fact it would be, we can calculate β , the probability of a Type II error. As in Section 12.5, the process of calculating β is broken into two steps. The first step is to determine the rejection region in terms of the unstandardised test statistic \hat{p} . The rejection region in terms of the standardised statistic Z is $Z > 1.645$; therefore,

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{\hat{p} - 0.10}{\sqrt{\frac{(0.10)(0.90)}{400}}} > 1.645$$

which holds whenever $\hat{p} > 0.125$. Therefore, we do not reject H_0 when $\hat{p} < 0.125$.

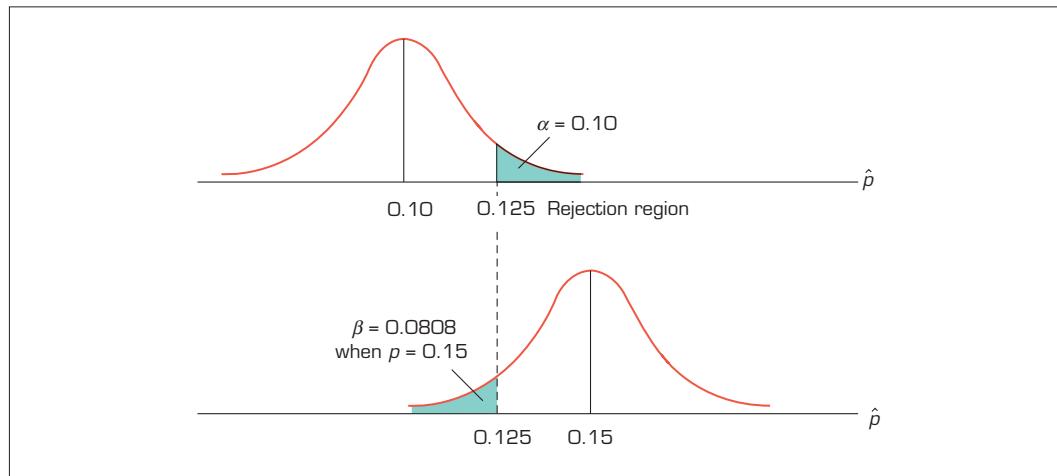
The second step is to calculate the probability that \hat{p} will not fall into the rejection region when p is actually greater than 10% (since $H_A: p > 0.10$). Suppose that we want to know β when p is really 0.15. That is,

$$\begin{aligned}\beta &= P(\hat{p} < 0.125 | p = 0.15) = P\left(\frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} < \frac{0.125 - 0.15}{\sqrt{\frac{(0.15)(0.85)}{400}}}\right) \\ &= P(Z < -1.4) \\ &= 0.0808\end{aligned}$$

In this calculation we assume that $p = 0.15$ and, as a result, the standard deviation of \hat{p} changes.

The probability of mistakenly concluding that the product would not be profitable when 15% of the potential customers actually would buy the product is 0.0808. **Figure 12.25** depicts this calculation.

FIGURE 12.25 Calculation of the probability of a Type II error



EXERCISES

Learning the techniques

12.64 If $\hat{p} = 0.57$ and $n = 100$, can we conclude at the 5% level of significance that the population proportion p is greater than 0.50?

12.65 Test each of the following hypotheses given the sample size n , the level of significance α and the sample proportion \hat{p} .

	Hypotheses	n	α	\hat{p}
a	$H_0: p = 0.45$ $H_A: p \neq 0.45$	100	0.05	0.40
b	$H_0: p = 0.7$ $H_A: p > 0.7$	1000	0.01	0.75
c	$H_0: p = 0.25$ $H_A: p < 0.25$	2000	0.10	0.23

12.66 In Exercise 12.65 determine the p -value of each test.

12.67 In Exercise 12.65 determine β for the following:

- a** True value of $p = 0.50$
- b** True value of $p = 0.73$
- c** True value of $p = 0.22$

12.68 Repeat Exercise 12.67(a) with $n = 500$.

12.69 In a sample of 200, we observe 140 successes.

- a** Is this sufficient evidence at the 1% significance level to indicate that the population proportion of successes exceeds 65%?
- b** Find the p -value of the test described in part (a).
- c** For the test described in part (a), find the probability of erroneously concluding that p does not exceed 65% when in fact p is 68%.

12.70 a Calculate the p -value of the test of the following hypotheses given that $\hat{p} = 0.63$ and $n = 100$.

$$\begin{aligned} H_0: p &= 0.60 \\ H_A: p &> 0.60 \end{aligned}$$

- b** Repeat part (a), with $n = 200$.
- c** Repeat part (a), with $n = 400$.
- d** Describe the effect on the p -value of increasing the sample size.

12.71 a A statistics practitioner wants to test the following hypotheses:

$$\begin{aligned} H_0: p &= 0.80 \\ H_A: p &< 0.80 \end{aligned}$$

A random sample of 100 produced $\hat{p} = 0.73$.

Calculate the p -value of the test.

- b** Repeat part (a) with $\hat{p} = 0.72$.
- c** Repeat part (a) with $\hat{p} = 0.71$.
- d** Describe the effect on the z -statistic and its p -value of decreasing the sample proportion.

Applying the techniques

12.72 **Self-correcting exercise.** A tyre manufacturer claims that more than 90% of the company's tyres will last at least 80000km. In a random sample of 200 tyres, 10 wore out before reaching 80000km. Do the data support the manufacturer's claim with $\alpha = 0.01$?

12.73 In a random sample of 100 units from an assembly line, 15 were defective. Does this constitute sufficient evidence at the 10% significance to conclude that the defective rate among all units exceeds 10%?

12.74 A university bookstore claims that 50% of its customers are satisfied with the service and prices.

- a** If this claim is true, what is the probability that in a random sample of 600 customers, less than 45% are satisfied?
- b** Suppose that in a random sample of 600 customers, 270 express satisfaction with the bookstore. What does this tell you about the bookstore's claim?

12.75 A psychologist believes that 80% of male drivers when lost continue to drive, hoping to find the location they seek rather than ask directions. To examine this belief, he took a random sample of 350 male drivers and asked each what they did when lost. Seventy-five per cent said they continued driving. Does this support the psychologist's belief?

12.76 A restaurant chain regularly surveys its customers. On the basis of these surveys, the management of the chain claims that at least 75% of its customers rate the food as excellent. A consumer testing service wants to examine the claim by asking 460 customers to rate the food. Eighty per cent rated the food as excellent. Does this support the management's claim?

12.77 An accounting lecturer claims that no more than one-quarter of undergraduate business students will major in accounting.

- a** If this claim is true, what is the probability that in a random sample of 1200 undergraduate business students, 336 or more will major in accounting?
- b** A survey of a random sample of 1200 undergraduate business students indicates that there are 336 students who plan to major in accounting. What does this tell you about the lecturer's claim?
- 12.78** In some countries the law requires drivers to turn on their headlights when driving in the rain. A highway patrol officer in one such country believes that less than one-quarter of all drivers follow this rule. As a test, he randomly samples 200 cars driving in the rain and counts the number whose headlights are turned on. He finds this number to be 41. Does the officer have enough evidence at the 10% significance level to support his belief?
- 12.79** A large airline bragged that more than 92% of its flights were on time. A random sample of 165 flights completed in 2019 revealed that 153 were on time. Can we conclude at the 5% significance level that the airline's claim about its on-time performance is valid?

Computer/manual applications

The following exercises require the use of a computer and software. The answers may be calculated manually using the sample statistics provided.

- 12.80 XR12-80** In a television commercial, the manufacturer of a toothpaste claims that more than four out of five dentists recommend the ingredients in his product. To test that claim, a consumer-protection group randomly samples 400 dentists and asks each one whether he or she would recommend a toothpaste that contained the ingredients. The responses are recorded coded as 0 = No and 1 = Yes. At the 5% significance level, can the consumer group infer that the claim is true?

Sample statistics: $n(0) = 71$; $n(1) = 329$.

- 12.81** What is the p -value of the test in Exercise 12.80?

- 12.82 XR12-82** A lecturer in business statistics recently adopted a new textbook. At the completion of the course, 100 randomly selected students were asked to assess the book. The responses are as follows:

Excellent (1), Good (2), Adequate (3), Poor (4)

The results are recorded using the codes in parentheses.

- a** Do these results allow us to conclude at the 5% significance level that more than 50% of all business students would rate it as excellent?
- b** Do these results allow us to conclude at the 5% significance level that more than 90% of all business students would rate it as at least adequate?

Sample statistics: $n(1) = 57$, $n(2) = 35$, $n(3) = 4$, $n(4) = 4$.

- 12.83 XR12-83** An insurance company boasts that 90% of its customers who make claims are satisfied with the service. To check the accuracy of this declaration, the company conducts an annual Claimant Satisfaction Survey in which customers are asked whether they were satisfied with the quality of the service (1 = Satisfied and 2 = Unsatisfied). Their responses are recorded. Can we infer at the 5% level that the satisfaction rate is less than 90%?

Sample statistics: $n(1) = 153$, $n(2) = 24$.

- 12.84 XR12-84** An increasing number of people are giving gift vouchers as Christmas presents. To measure the extent of this practice, a random sample of 120 people were asked (survey conducted 26–29 December) whether they had received a gift voucher for Christmas. The responses are recorded as 1 = No and 2 = Yes. Can we infer that the proportion of people who received a gift voucher for Christmas is more than 20% (use $\alpha = 0.05$)?

Sample statistics: $n(1) = 92$, $n(2) = 28$.

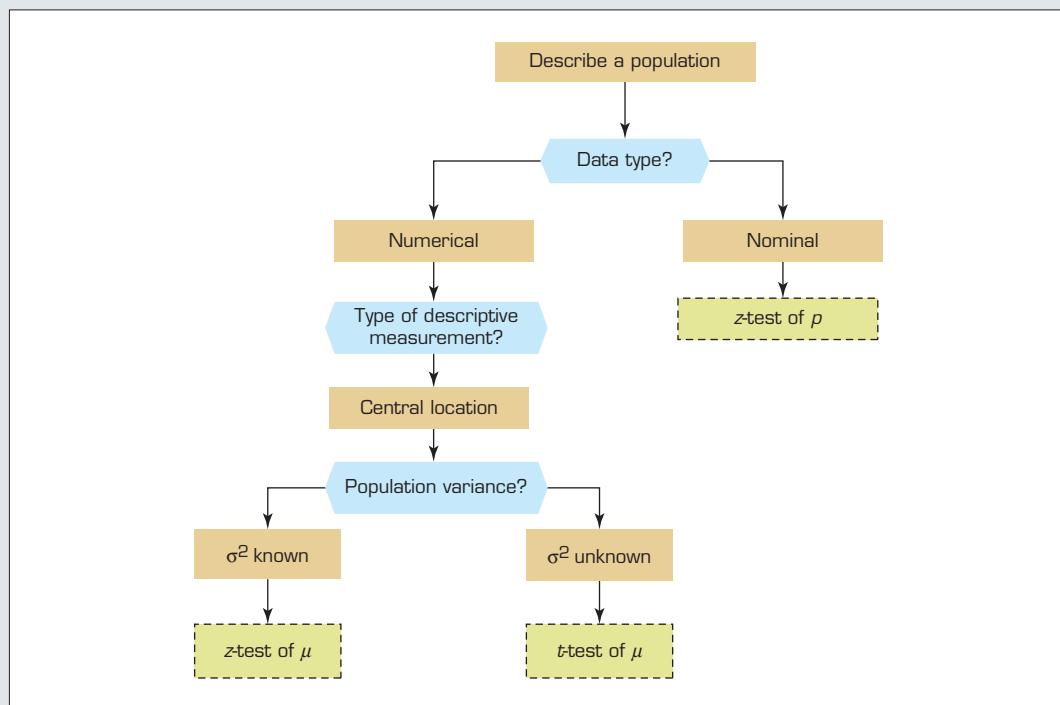
Study Tools

CHAPTER SUMMARY

In this chapter we presented the hypothesis testing techniques that are used in describing a single population. As was the case with estimation, the statistics practitioner must identify the parameter to be tested and its test statistic. Other important steps in the process include setting up the hypotheses and specifying the decision rule. In this chapter, the parameters μ and p are tested by reference to the sampling distributions of \bar{X} and \hat{p} , respectively. The alternative hypothesis is set up to answer the question, while the null hypothesis states that the parameter equals a fixed value. The test is conducted by specifying the significance level α .

Table 12.4 shows the test statistics and the required conditions.

This chapter also presented p -values, the calculation of β and the power of a test.



The data files for Examples, Exercises and Cases are provided on the companion website (accessible through <https://login.cengagebrain.com/>).

SUMMARY OF FORMULAS

TABLE 12.4 Summary of test statistics for μ and p

Parameter	Test statistic	Required conditions
μ (Numerical)	$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$	σ^2 is known; X is normally distributed or $n \geq 30$
μ (Numerical)	$t = \frac{\bar{X} - \mu}{s / \sqrt{n}}$	σ^2 is unknown and estimated by s^2 ; X is normally distributed
p (Nominal)	$z = \frac{\hat{p} - p}{\sqrt{pq/n}}$	$n\hat{p} \geq 5$ and $n\hat{q} \geq 5$, where $\hat{q} = 1 - \hat{p}$

SUPPLEMENTARY EXERCISES

- 12.85** Suppose that hourly wages in the chemical industry are normally distributed, with a mean of \$50 and a standard deviation of \$12. A large company in this industry took a random sample of 50 of its workers and determined that their average hourly wage was \$48. Can we conclude at the 10% level of significance that this company's average hourly wage is less than that of the entire industry?
- 12.86** Refer to Exercise 12.85. What is the *p*-value of this test?
- 12.87** For the past few years, the number of customers of a drive-in bottle shop has averaged 20 per hour, with a standard deviation of three per hour. This year, another bottle shop 1 km away opened a drive-in service. The manager of the first shop believes that this will result in a decrease in the number of customers. A random sample of 15 hours showed that the mean number of customers per hour was 18.7. Can we conclude at the 5% level of significance that the manager's belief is correct?
- 12.88** A manufacturer of computer chips claims that at least 90% of the product conforms to specifications. In a random sample of 1000 chips drawn from a large production run, 125 were defective. Do the data provide sufficient evidence at the 1% level of significance to conclude that the manufacturer's claim is false? What is the *p*-value of this test?
- 12.89 XR12-89** An automotive expert claims that the conversion of petrol stations to self-serve petrol stations has resulted in poor car maintenance, and that the average tyre pressure is at least 28 kilopascals below its manufacturer's specification. As a quick test, 10 tyres are examined, and the number of kilopascals by which each tyre is below specification is recorded. The resulting data (in kilopascals) are as follows:
- | | | | | | | | | | |
|----|----|----|---|----|----|----|----|----|----|
| 48 | 62 | 14 | 0 | 34 | 41 | 21 | 34 | 55 | 62 |
|----|----|----|---|----|----|----|----|----|----|
- Is there sufficient evidence, with $\alpha = 0.05$, to support the expert's claim?
- 12.90** A fast-food franchiser is considering building a restaurant at a certain location. On the basis of financial analysis, a site is acceptable only if the number of pedestrians passing the location averages at least 100 per hour. A random sample of 50 hours produced $\bar{X} = 110$ and $s = 12$ pedestrians per hour. Do these data provide sufficient evidence to establish that the site is acceptable? (Use $\alpha = 0.05$.)
- 12.91** Officials of a private bus company operating between Brisbane and Sydney claim that less than 10% of all its buses are late. If a random sample of 70 buses shows that only 60 of them are on schedule, can we conclude that the claim is false? (Use $\alpha = 0.10$.)
- 12.92** In a wealthy suburb in Sydney, 22% of the households had a Sunday newspaper delivered to their doors. After an aggressive marketing campaign to increase that figure, a random sample of 200 households was taken, and it was found that 61 households now have the paper delivered. Can we conclude at the 5% significance level that the campaign was a success?
- 12.93 XR12-93** The owner of a city car park suspects that the person she hired to run the car park is stealing money. The receipts as provided by the employee indicate that the average number of cars parked is 125 per day and that, on average, each car is parked for 3.5 hours. In order to determine whether or not the employee is stealing, the owner watches the car park for five days. On those days the number of cars parked is as follows:
- | | | | | |
|-----|-----|-----|-----|-----|
| 120 | 130 | 124 | 127 | 128 |
|-----|-----|-----|-----|-----|
- For the 629 cars that the owner observed during the five days, the mean and the standard deviation of the time spent at the car park were 3.6 and 0.4 hours, respectively. Can the owner conclude at the 5% level of significance that the employee is stealing? (*Hint:* As there are two ways to steal, two tests should be performed.)

Computer/manual applications

The following exercises require the use of a computer and software. The answers may be calculated manually using the sample statistics provided.

- 12.94 XR12-94** The television networks often compete on the evening of an election day to be the first to correctly identify the winner of the election. One commonly used technique is the random sampling of voters as they exit the polling booths. Suppose that, in a two-candidate race, 500 voters were asked for whom they voted. The results were recorded using the code 1 = Liberal/National Party and 2 = Labor. Can we conclude at the 5% level of significance that the Labor Party candidate will win?

Sample statistics: $n(1) = 232$, $n(2) = 268$.

- 12.95 XR12-95** Suppose that in a large university (with numerous campuses) the marks in an introductory statistics course are normally distributed with a mean of 68%. To determine the effect of requiring students to pass a calculus test (which at present is not a prerequisite), a random sample of 50 students who have taken calculus is given a statistics course. The marks out of 100 are recorded. Do these data provide sufficient evidence to infer that students with a calculus background would perform better in statistics than students with no calculus background?

Sample statistics: $\bar{X} = 71.88$; $s = 10.03$, $n = 50$.

Case Studies

CASE 12.1 Singapore Airlines has done it again

Based on communications with his customers, a travel agent believes that Singapore airlines has a customer satisfaction rate of at least 85%. However, he has seen a new paper quoting a Roy Morgan survey that the customer satisfaction rate is 89%. The Roy Morgan website stated (24 May 2019, Finding Number 7985, Topic: Roy Morgan International Airline Customer Satisfaction Press Release, Country: Australia):

Michele Levine, CEO, Roy Morgan, says Singapore Airlines is on track to win another Annual Roy Morgan International Airline Customer Satisfaction Award after five victories in the annual category in 2012, 2013, 2014, 2015 and 2018.

Singapore Airlines has won the latest international airline customer satisfaction award with an exceptional customer satisfaction rating of 89% in April ahead of Emirates on 87%, Qatar Airways on 85% and Etihad Airways on 83%.

This result was based on a sample of $n = 3657$ travellers. Test whether there is any support for the travel agent's belief based on the Roy Morgan survey?

CASE 12.2 Australian rate of real unemployment

Roy Morgan believes that the true rate of unemployment is higher than the rate of 5.3% used by the Australian Bureau of Statistics (December 2019). The average Roy Morgan Research unemployment figure for the month of September 2019 was 8.7%. The Roy Morgan unemployment estimates are based on a face-to-face survey of an Australia-wide cross-section of 4000 Australians aged 14 and over. Do the data support Roy Morgan's belief?

Source: <https://www.roymorgan.com/findings/8159-australian-unemployment-estimates-september-2019-201910110149>

CASE 12.3 The republic debate: What Australians are thinking

In November 1999, a constitutional referendum was held to determine whether or not Australia should become a republic by the centenary of Federation, 1 January 2001. More than 50% of Australians voted to reject the question, 'Should Australia become a republic?'. In a survey carried out by ReachTEL during 30 January 2014, the following question was asked: 'Would you support Australia becoming a republic?' On the basis of 2100 respondents, the results were:

Yes	827
No	874
Undecided	399

Do you support the claim that more than 75% of Australians support Australia becoming a republic?

CASE 12.4 Has Australian Business Confidence improved since the May 2019 election?

It is widely believed that Business Confidence in Australia during the recent post-election phase (June–August 2019) improved well above the 105 mark of the pre-election phase (June–August 2018). In a recent statement, Roy Morgan stated:

... the improvement in Business Confidence in recent months shows that significant Government income tax cuts as well as the RBA's post-election interest rate cuts have given businesses renewed confidence about the future.... Analysing Business Confidence by State shows that the smaller States are performing best and Western Australia now has the highest Business Confidence of any State in August just ahead of South Australia and Tasmania.

Business Confidence was just above the national average in New South Wales and dropped in both Queensland and Victoria which now has the lowest Business Confidence of any State although still in positive territory above 100.

Roy Morgan published the following information on Business Confidence by leading industries:

Education and Training	132.6
Wholesale	128.0
Agriculture	121.6
Property and Business Services	118.9
Retail	116.1
Construction	115.5
Accommodation and Food Services	115.4
Professional Scientific and Technical Services	115.3

Source: <https://www.roymorgan.com/findings/8121-roy-morgan-business-confidence-august-2019-201909090047>.

Test whether overall post-election Australian Business Confidence has passed the level of 120.

CASE 12.5 Is there a gender bias in the effect of COVID-19 infection?

C12-05 The coronavirus pandemic, which was first identified in December 2019, had infected more than 5 million people worldwide by 20 May 2020. It is widely believed that the virus is affecting more men than women. The confirmed number of cases and deaths and the number of male cases and deaths in randomly selected countries were recorded. Using the data, test whether there is a gender bias in the confirmed cases and deaths. (Hint: Test whether the proportion of male confirmed cases is greater than 0.5 and whether the proportion of male deaths due to coronavirus is greater than 0.5.)

Source: <https://globalhealth5050.org/covid19/sex-disaggregated-data-tracker/>

Appendix 12.A

Excel instructions

Testing the population mean when the variance σ^2 is known

Instead of using the macro we created, you can use one of Excel's built-in ZTESTs.

- 1 Open the Excel file (e.g. XM12-02.xlsx).
- 2 Click the **Insert function fx**, **Statistical**, **ZTEST** and **OK**.
- 3 Specify the location of the data (array), e.g. **A1:A26** (for Example 12.2).
- 4 Specify the value of the parameter under the null hypothesis (x), e.g. **500**.
- 5 Specify the value of the population standard deviation (sigma), e.g. **10**.
- 6 Click **OK**.

Excel will output a one-tail p -value. It is actually $P(Z > \text{calculated value of the test statistic}, z)$. In Example 12.2, $z = -0.12$. Thus, Excel outputs $P(Z > -0.12)$, which is 0.5478. Since the alternative hypothesis states that $\mu < 500$, we calculate the p -value by subtracting 0.5478 from 1. Thus, the p -value = 0.4522.

Hypothesis testing: Two populations

Learning objectives

This chapter extends the approaches developed in Chapter 12 to test the differences in two population parameters involving numerical and nominal data.

At the completion of this chapter, you should be able to:

- L01** test hypotheses about differences in two population means with independent samples and known population variances
- L02** test hypotheses about differences in two population means with independent samples and unknown and unequal population variances
- L03** test hypotheses about differences in two population means with independent samples and unknown but equal population variances
- L04** test hypotheses about differences in two population means with dependent samples
- L05** test hypotheses about differences in two population proportions.

CHAPTER OUTLINE

Introduction

13.1 Testing the difference between two population means: Independent samples

13.2 Testing the difference between two population means: Dependent samples – matched pairs experiment

13.3 Testing the difference between two population proportions

SPOTLIGHT ON STATISTICS

Selecting a location for a new department store

Often when an upmarket department store chain such as Myer or David Jones wishes to open a new store in a region, management needs to make a decision regarding the location of the new store based on a number of consumer characteristics. One of the main location-based characteristics used in such decision making is the average income of residents living in the surrounding areas.

The management of a chain of department stores wants to know if there is a difference in the average annual income of potential customers at two possible sites for a new store. In the first location, a random sample of 100 households showed a mean annual income of \$166000. In the second location, the mean annual income of 75 households was \$134000. Assume that annual income in both locations is normally distributed with a standard deviation of \$10000. At the 5% significance level, can it be concluded that the mean household income of the first location exceeds that of the second location? The solution is shown on pages 535–6.



Source: Shutterstock.com/Alexzel

Introduction

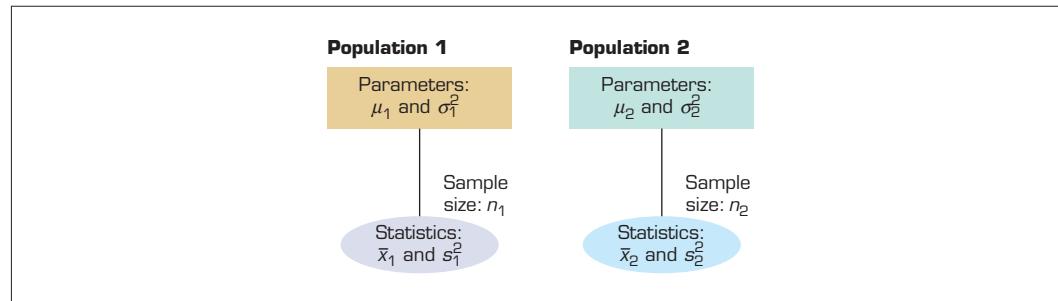
In this chapter we continue exploring statistical inference by extending the range of techniques and discussing how to perform tests of hypothesis when the problem objective involves comparing two populations. When the data type is numerical, we will test hypotheses about the difference between two population means, $\mu_1 - \mu_2$. When the data type is nominal, we will test hypotheses about the difference between two population proportions, $p_1 - p_2$. Examples of the use of these methods include the following:

- 1 Consumers who purchase televisions, major household appliances and cars consider reliability a major factor in their brand choice. Reliability can be measured by the length of time the product lasts. To compare two brands of televisions, we would test $\mu_1 - \mu_2$, the difference between the mean lifetimes of the two brands.
- 2 Production supervisors and quality-control engineers are responsible for measuring, controlling and minimising the number of defective units that are produced at a plant. Frequently, more than one method or machine can be used to perform the manufacturing function. The decision about which one of two machines to acquire and use often depends on which machine produces the smaller proportion of defective units – or, in other words, on the parameter $p_1 - p_2$, the difference in the proportions of defective units from each machine.

13.1 Testing the difference between two population means: Independent samples

In this section, we consider independent samples – samples that are completely unrelated to one another. In Section 13.2, we consider the matched pairs experiment.

FIGURE 13.1 Independent samples from two populations



13.1a Six-steps of testing hypothesis

In Chapter 12 we presented the six-step process used to test hypotheses for a single population parameter. In this chapter, we will go through these same six steps to test hypotheses regarding two populations.

Step 1: Specify the null and alternative hypotheses.

In this section, of course, all hypotheses will feature $\mu_1 - \mu_2$. The null hypothesis will again specify that $\mu_1 - \mu_2$ is equal to some value D (usually zero), while the alternative hypothesis takes one of the following three formats, depending on what the question asks:

- 1 $H_A: \mu_1 - \mu_2 \neq D$ (Two-tail test)
- 2 $H_A: \mu_1 - \mu_2 > D$ (Right one-tail test)
- 3 $H_A: \mu_1 - \mu_2 < D$ (Left one-tail test)

Step 2: Determine the test statistic.

As you will recall from Chapter 11, we had access to three different confidence interval estimators of the difference between two population means; and the choice depended on several factors. Those same factors combine to produce three different *test statistics*. The test statistics and the factors that identify their use are as follows.

known variances **t -test statistic of $\mu_1 - \mu_2$**

The statistic used to test for the equality of means when both population variances are known.

unequal variances **t -test statistic of $\mu_1 - \mu_2$**

The statistic used to test for the equality of two population means when the variances of the two populations are unknown and presumed, or tested, to be unequal.

equal variances **t -test statistic for $\mu_1 - \mu_2$**

The statistic used to test for the equality of means when both variances of two populations are unknown and presumed, or tested, to be equal.

Test statistics and factors that identify their use

Case 1 Known variances t -test statistic

σ_1^2 and σ_2^2 are known, and the populations are normally distributed or n_1 and n_2 are large.

$$\text{Test statistic: } Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has a standard normal distribution.

Case 2 Unequal variances t -test statistic

σ_1^2 and σ_2^2 are unknown and unequal ($\sigma_1^2 \neq \sigma_2^2$) and the populations are normally distributed.

$$\text{Test statistic: } t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

has a t distribution with degrees of freedom given by

$$\text{d.f.} = \frac{\left[(s_1^2/n_1) + (s_2^2/n_2) \right]^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

Case 3 Equal variances t -test statistic

σ_1^2 and σ_2^2 are unknown but equal ($\sigma_1^2 = \sigma_2^2$) and the populations are normally distributed.

$$\text{Test statistic: } t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

has a t distribution with degrees of freedom d.f. = $n_1 + n_2 - 2$ and s_p^2 is the pooled variance estimate, defined as

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$

Note that in cases 2 and 3 above, if d.f. ≥ 200 , the test statistic is approximately normally distributed, allowing us to find the critical value of the rejection region by using either Table 3 (standard normal random variable) or Table 4 (t -random variable) in Appendix B.

The remaining parts of the test are the same as in Chapter 12. That is, we proceed as follows:

Step 3: Specify the significance level.

Step 4: Define the decision rule.

Step 5: Calculate the value of the test statistic.

Step 6: Conclusion. Make a decision and answer the question.

The next three examples illustrate the procedure for testing $\mu_1 - \mu_2$.

EXAMPLE 13.1

L01

CEO salaries: Is there a gender difference?

Until recently, most Chief Executive Officers (CEOs) of companies in Australia have been men. However, as government encourages companies to increase the number of female CEOs, a number of companies now have female CEOs. A researcher decided to examine the success of female CEOs of medium-size companies by comparing their base salaries (excluding bonuses and outliers) with those of their male counterparts. She took a random sample of 100 female and 100 male CEOs and recorded their salaries for the preceding year. The average salary for the female CEOs was \$402 500 and for the male CEOs was \$430 000. From past records, we have information that CEO incomes are normally distributed with a standard deviation of \$7000 for both females and males. At the 5% level of significance, is there enough evidence to support the claim that average salaries of female CEOs are lower than average salaries of male CEOs, of medium-size companies?

Solution**Identifying the technique**

The problem objective is to compare two populations whose data type is numerical (salaries of female CEOs and male CEOs of medium-size companies). This tells us that the parameter to be tested is $\mu_1 - \mu_2$, where μ_1 = mean annual salary of female CEOs of medium-size companies, and μ_2 = mean annual salary of male CEOs of medium-size companies. Because we want to know if we can conclude that μ_1 is lower than μ_2 , the alternative hypothesis is

$$H_A: \mu_1 - \mu_2 < 0$$

The population standard deviations are known ($\sigma_1 = 7000$ and $\sigma_2 = 7000$) and n_1 and n_2 are large. The correct test statistic is

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The complete test follows:

Step 1. Hypotheses:

$$\begin{aligned} H_0: \mu_1 - \mu_2 &= 0 \\ H_A: \mu_1 - \mu_2 &< 0 \end{aligned} \quad (\text{Left one-tail test})$$

Step 2. Test statistic:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \text{ has a standard normal distribution.}$$

Step 3. Significance level:

$$\alpha = 0.05$$

Step 4. Decision rule:

Reject H_0 if $Z < -z_{\alpha} = -z_{0.05} = -1.645$.
Alternatively, reject H_0 if $p\text{-value} < \alpha = 0.05$.

Calculating manually

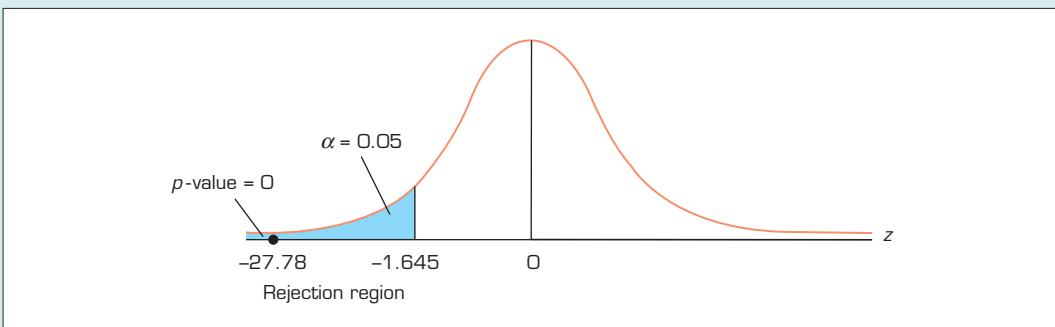
Step 5. Value of the test statistic:

$$Z = \frac{(402500 - 430000) - 0}{\sqrt{\frac{7000^2}{100} + \frac{7000^2}{100}}} = -27.78$$

Step 6. Conclusion: As $Z = -27.78 < -1.645$, reject H_0 .

Figure 13.2 depicts the distribution of the standardised test statistic.



FIGURE 13.2 Sampling distribution for Example 13.1.

Interpreting the results

There is enough evidence to infer that the mean salary of female CEOs is lower than that of male CEOs at the 5% significance level. Note, however, that the conclusion will be the same even if we use the *p*-value method.

Using the computer

As the sample means and population standard deviations are already known (given), use the **z-test_2Means** worksheet in the **Test Statistics** workbook that is provided on the companion website (accessible through <https://login.cengagebrain.com/>) and enter the required sample means, population variances, sample sizes, hypothesised difference in means and level of significance. The output is shown below.

Excel output for Example 13.1

	A	B	C	D	E
1	z-Test of the Difference Between Two Means (known Variances)				
2					
3		Sample 1	Sample 2	Confidence Interval Estimate	
4	Sample mean	402500	430000	z Stat	-27.78
5	Population variance	49000000	49000000	P[Z<=z] one-tail	0.0000
6	Sample size	100	100	z Critical one-tail	1.6449
7	Hypothesized difference	0		P[Z<=z] two-tail	0.0000
8	Alpha	0.05		z Critical two-tail	1.9600

Excel outputs the statistical results for the test of hypothesis. The value of the test statistic *z* is -27.78, with a one-tail *p*-value of 0. Excel also outputs the critical values (using a 5% significance level) for a one-tail and a two-tail test. It also provides the two-tail test *p*-value, which is not needed in this example.

Known variances

z-test of $\mu_1 - \mu_2$

This test assesses the significance of the difference between two populations when the variances are known.

Because three different possible test statistics can be used in testing the difference between two population means, you must be particularly careful to identify factors that determine when to use the *z*-test of $\mu_1 - \mu_2$. Here is a summary of how we recognise when to use the **known-variances z-test** of $\mu_1 - \mu_2$.

IN SUMMARY

Factors that identify the z-test of $\mu_1 - \mu_2$

- 1 *Problem objective:* to compare two populations
- 2 *Data type:* numerical (quantitative)
- 3 *Descriptive measurement:* central location
- 4 *Experimental design:* independent samples
- 5 *Population variances:* known

Now we will discuss the problem posed in this chapter's introduction.

SPOTLIGHT ON STATISTICS

Selecting a location for a new department store: Solution

Identifying the technique

The problem objective is to compare two populations whose data type is numerical (household income in locations 1 and 2). This tells us that the parameter to be tested is $\mu_1 - \mu_2$ (where μ_1 mean annual household income in location 1, and μ_2 mean annual household income in location 2). Because we want to know if we can conclude that μ_1 exceeds μ_2 , the alternative hypothesis is

$$H_A: \mu_1 - \mu_2 > 0$$

The population standard deviations are known ($\sigma_1 = 10000$ and $\sigma_2 = 10000$) and n_1 and n_2 are large. The correct test statistic is

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The complete test follows:

Step 1. Hypotheses: $H_0: \mu_1 - \mu_2 = 0$

$H_A: \mu_1 - \mu_2 > 0$ (Right one-tail test)

Step 2. Test statistic:

$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ is distributed as a standard normal distribution.

Step 3. Significance level: $\alpha = 0.05$

Step 4. Decision rule: Reject H_0 if $Z > z_{\alpha} = z_{0.05} = 1.645$

Alternatively, reject H_0 if $p\text{-value} < \alpha = 0.05$.

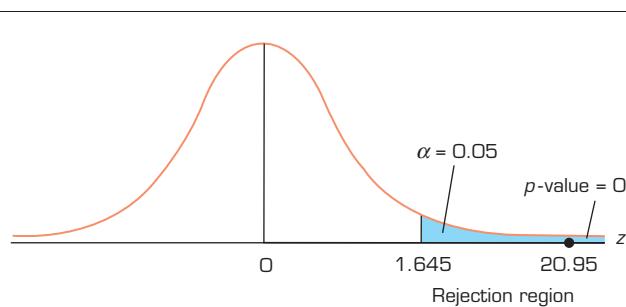
Calculating manually

Step 5. Value of the test statistic:

$$Z = \frac{(166000 - 134000) - 0}{\sqrt{\frac{10000^2}{100} + \frac{10000^2}{75}}} = 20.95$$

Step 6. Conclusion: As $Z = 20.95 > 1.645$, reject H_0 .

The figure below depicts the distribution of the standardised test statistic.



Source: Shutterstock.com/Alexzel

Interpreting the results

There is enough evidence to infer that the mean household income in location 1 exceeds that of location 2. Hence, we recommend locating the new store in location 1.

Using the computer

As the sample means and population standard deviations are already known (given), use the **z-test_2Means** worksheet in the **Test Statistics** workbook.

Excel output for the opening example

	A	B	C	D	E
1	z-Test of the Difference Between Two Means (known Variances)				
2					
3		Sample 1	Sample 2	Confidence Interval Estimate	
4	Sample mean	166000	134000	z Stat	20.95
5	Population variance	1000000000	1000000000	P(Z<=z) one-tail	0.0000
6	Sample size	100	75	z Critical one-tail	1.6449
7	Hypothesized difference	0		P(Z<=z) two-tail	0.0000
8	Alpha	0.05		z Critical two-tail	1.9600

Excel outputs the statistical results related to the test of hypothesis. The value of the test statistic z is 20.95, with a one-tail p -value of 0. As p -value = 0 < 0.05, we reject H_0 in favour of H_A and conclude that $\mu_1 - \mu_2 > 0$.

EXAMPLE 13.2

LO2

Dietary effects of high-fibre breakfast cereals: Part II

XM13-02 Refer to Example 11.3. At the 5% significance level, test the scientist's claim that people who eat high-fibre cereal for breakfast will consume, on average, fewer kilojoules for lunch than people who don't eat high-fibre cereal for breakfast. Assume that the two populations are normal.

Solution

Identifying the technique

In order to assess the claim, the scientists need to compare the mean kilojoule intake of the population of consumers of high-fibre cereal for breakfast (μ_1) with the mean kilojoule intake of the population of non-consumers of high-fibre cereal for breakfast (μ_2). The data are numerical (obviously, as we have recorded real numbers). This problem objective–data type combination tells us that the parameter to be tested is the difference between two means: $\mu_1 - \mu_2$.

The claim to be tested is that the mean kilojoule intake of consumers (μ_1) is less than that of non-consumers (μ_2).

Hypotheses: $H_0: \mu_1 - \mu_2 = 0$

$H_A: \mu_1 - \mu_2 < 0$ (Left one-tail test)

Test statistic: The population variances are unknown. To identify the test statistic, the scientists calculate the sample standard deviations:

$$s_1 = 142.75 \quad \text{and} \quad s_2 = 462.61$$

There is reason to believe that the population variances are unequal. We are also given that the two populations are normally distributed. Thus, we use the unequal-variances test statistic:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$



which has a t distribution with

$$\text{d.f.} = \frac{\left(\frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\left(\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}\right)}\right)}{}$$

Level of significance: $\alpha = 0.05$

Calculating manually

Decision rule: From the data we calculated the following statistics:

$$\bar{X}_1 = 2383.2$$

$$\bar{X}_2 = 2644.4$$

$$s_1 = 142.75$$

$$s_2 = 462.61$$

The number of degrees of freedom of the test statistic is:

$$\begin{aligned} \text{d.f.} &= \frac{\left(\frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\left(\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}\right)}\right)}{=} \\ &= \frac{\left[\frac{(142.75)^2/10 + (462.61)^2/20}{10-1}\right]^2}{\left[\frac{(142.75)^2/10}{10-1} + \frac{(462.61)^2/20}{20-1}\right]} \\ &= 25.01 \approx 25 \text{ (rounded)} \end{aligned}$$

Therefore, the decision rule is:

Reject H_0 if $t < -t_{\alpha, \text{d.f.}} = -t_{0.05, 25} = -1.708$.

Alternatively, reject H_0 if $p\text{-value} < \alpha = 0.05$.

Value of the test statistic:

$$\begin{aligned} t &= \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\ &= \frac{(2383.2 - 2644.4) - 0}{\sqrt{\frac{(142.75)^2}{10} + \frac{(462.61)^2}{20}}} = -2.31 \end{aligned}$$

Conclusion: As $t = -2.31 < -1.708$, reject the null hypothesis.

Interpreting the results

As discussed in Chapter 12, the p -value ($p = 0.0146$) of the test (from the Excel output) is smaller than the level of significance $\alpha = 0.05$, indicating that we should reject the null hypothesis. This leads us to conclude that these data provide enough evidence to infer that consumers of high-fibre cereal do eat fewer kilojoules at lunch than do non-consumers. However, there are several reasons to be cautious about concluding that high-fibre cereals constitute an effective contribution to weight loss. First, the sample sizes are small. (Conclusions based on larger sample sizes are more realistic and reliable.) Second, the data were likely to be self-reported, which means that each person determined the number of kilojoules that he or she consumed. Such data are often unreliable. Ideally, a less subjective method of counting kilojoules should be used. Finally, the way in which the experiment was performed may lead to several contradictory interpretations of the data. We will discuss this important issue in the next section.



Using the computer

Using Excel Data Analysis

Excel output for Example 13.2

	A	B	C
1	t-Test: Two-Sample Assuming Unequal Variances		
2			
3		Consumers	Non-consumers
4	Mean	2383.2	2644.4
5	Variance	20376.2	214004.0
6	Observations	10	20
7	Hypothesized Mean Difference	0	
8	df	25	
9	t Stat	-2.31	
10	P(T<=t) one-tail	0.0146	
11	t Critical one-tail	1.7081	
12	P(T<=t) two-tail	0.0292	
13	t Critical two-tail	2.0595	

COMMANDS

- 1 Type the data in two columns or open the data file (**XM13-02**).
- 2 Click **DATA, Data Analysis**, and **t-Test: Two-Sample Assuming Unequal Variances**. Click **OK**.
- 3 Specify the **Variable 1 Range (A1:A11)** and the **Variable 2 Range (B1:B21)**. Click **Labels** (if necessary). Type the value of the **Hypothesized Mean Difference**¹ (**0**) and type a value for α (**0.05**).

Using XLSTAT

Alternatively, when raw data are available, we can use XLSTAT to perform this task.

	B	C	D	E	F	G
1	Hypothesized difference (D): 0					
2	Significance level (%): 5					
3						
4	Summary statistics:					
5	Variable	Observations	Minimum	Maximum	Mean	Std. deviation
6	Consumers	10	2116.0	2560.0	2383.2	142.7
7	Non-consumers	20	2008.0	3804.0	2644.4	462.6
8						
9	t-test for two independent samples / Lower-tailed test:					
10	Difference	-261.200				
11	t (Observed value)	-2.314				
12	t (Critical value)	-1.708				
13	DF	25.011				
14	p-value (one-tailed)	0.015				
15	alpha	0.05				

¹ This term is technically incorrect. Because we're testing $\mu_1 - \mu_2$, Excel should ask for and output the 'Hypothesized Difference between Means'.



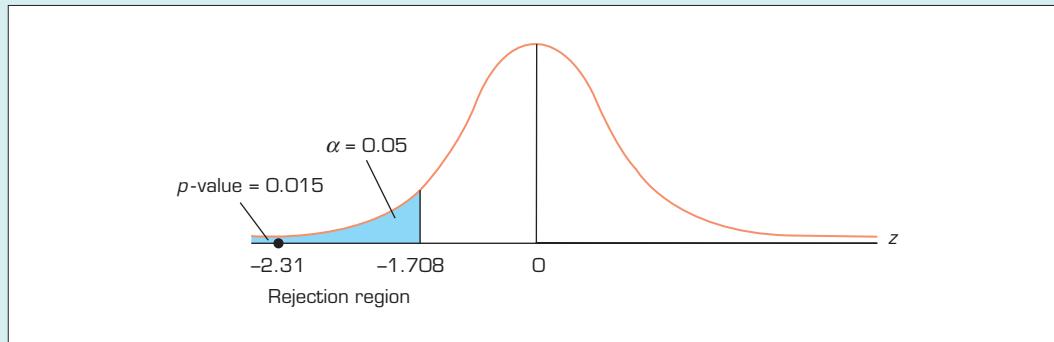
COMMANDS

- 1 Type the data in two columns or open the data file (XM13-02).
- 2 Click **XLSTAT**, **Parametric test**, and **Two-sample t-test and z-test**.
- 3 Check **One-column per sample**. Type the input range for both samples: **Sample 1 (A1:A11)**, **Sample 2 (B1:B21)**. Click **Student's t-test**. Do not click **z-test**.
- 4 Click the **Options** tab and choose **Mean 1 – Mean 2 < D** in the **Alternative hypothesis** box. Type the **Hypothesized difference (D) (0)**. Do not click **Assume equality** under **Population variances for the t-test**. (If you click **Use an F-test** you do not need to conduct a separate F-test of the two variances as a first step to testing the difference between two means.) Type the value of α in the **Significance level (%)** box (5). Click **OK**.

Excel outputs sample means, variances and sizes. It also outputs the statistical results related to the test of hypothesis. The value of the test statistic t Stat is -2.31 , with a one-tail p -value of 0.015 . Excel also outputs the critical values (using a 5% significance level) for a one-tail and a two-tail test. As p -value = $0.015 < 0.05 = \alpha$, reject the null hypothesis. It also provides the two-tail test p -value, which is not needed in this example.

Figure 13.3 depicts the sampling distribution of the test statistic.

FIGURE 13.3 Sampling distribution for Example 13.2



Here is a summary of how we recognise when to use the **unequal-variances t-test** of $\mu_1 - \mu_2$.

IN SUMMARY

Factors that identify the unequal-variances t-test of $\mu_1 - \mu_2$

- 1 *Problem objective:* to compare two populations
- 2 *Data type:* numerical (quantitative)
- 3 *Descriptive measurement:* central location
- 4 *Experimental design:* independent samples
- 5 *Population variances:* unknown and not equal

unequal variances t-test of $\mu_1 - \mu_2$

This test assesses the significance of the difference between two population means when the variances are unknown but are not expected to be equal.

REAL-LIFE APPLICATIONS

Operations management: Production design

During the production design stage of new products, an operations manager determines how a product is to be manufactured. The objective is to produce

the highest quality product at a reasonable cost. This objective is achieved by choosing the machines, materials, methods and 'manpower' (personnel),

the so-called 4 Ms. The manager can often employ statistical tools to help make this decision. Various experiments can be conducted to determine the lowest cost or fastest production schedule. The experiments use different materials, machines, methods or personnel. There are several ways to judge differences in processes. The manager can determine whether differences in quality or cost exist. If no differences exist, the manager may decide on the basis of some other criterion, such as the process that requires the least new training of workers.



Source: Shutterstock.com/Pressmaster

EXAMPLE 13.3

LO3

Direct and broker-purchased mutual funds: Part I

XM13-03 Millions of investors buy mutual funds, choosing from thousands of possibilities. Some funds can be purchased directly from banks or other financial institutions, whereas others must be purchased through brokers, who charge a fee for this service. This raises the question, can investors do better by buying mutual funds directly than by purchasing mutual funds through brokers? To help answer this question, a group of researchers randomly sampled the annual returns from mutual funds that can be acquired directly and mutual funds that are bought through brokers and recorded the net annual returns, which are the returns on investment after deducting all relevant fees. These are listed in **Table 13.1**. Can we conclude at the 5% significance level that directly purchased mutual funds outperform mutual funds bought through brokers? Assume that the two populations of annual returns are normally distributed.

TABLE 13.1 Annual returns (%)

Direct					Broker				
9.33	4.68	4.23	14.69	10.29	3.24	3.71	16.4	4.36	9.43
6.94	3.09	10.28	-2.97	4.39	-6.76	13.15	6.39	-11.07	8.31
16.17	7.26	7.1	10.37	-2.06	12.8	11.05	-1.9	9.24	-3.99
16.97	2.05	-3.09	-0.63	7.66	11.1	-3.12	9.49	-2.67	-4.44
5.94	13.07	5.6	-0.15	10.83	2.73	8.94	6.7	8.97	8.63
12.61	0.59	5.27	0.27	14.48	-0.13	2.74	0.19	1.87	7.06
3.33	13.57	8.09	4.59	4.8	18.22	4.07	12.39	-1.53	1.57
16.13	0.35	15.05	6.38	13.12	-0.8	5.6	6.54	5.23	-8.44
11.2	2.69	13.21	-0.24	-6.54	-5.75	-0.85	10.92	6.87	-5.72
1.14	18.45	1.72	10.32	-1.06	2.59	-0.28	-2.15	-1.69	6.95

Solution

Identifying the technique

To answer the question, we need to compare the population of returns from directly purchased funds and the returns from broker-bought mutual funds. The data are obviously numerical. This problem objective–data type combination tells us that the parameter to be tested is the difference between two population means, $\mu_1 - \mu_2$. The hypothesis to be tested is that the mean net annual return from directly purchased mutual funds (μ_1) is larger than the mean of broker-purchased funds (μ_2). Hence, the alternative hypothesis: $H_A: \mu_1 - \mu_2 > 0$. As usual, the null hypothesis automatically follows: $H_0: (\mu_1 - \mu_2) = 0$



As a result, the null and alternative hypotheses are

Hypotheses:

$$H_0: (\mu_1 - \mu_2) = 0$$

$$H_A: (\mu_1 - \mu_2) > 0 \quad (\text{Right one-tail test})$$

Test statistic:

The population variances are unknown. To identify the correct test statistic, we need to calculate the sample variances:

$$s_1^2 = 37.49 \quad \text{and} \quad s_2^2 = 43.34$$

Because these sample variances, s_1^2 and s_2^2 , do not differ much, we can infer that the population variances are approximately equal. The two population annual returns are normally distributed. Thus, we employ the equal variances test statistic:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{n_1+n_2-2}$$

Level of significance: $\alpha = 0.05$

Decision rule: The number of degrees of freedom is

$$\text{d.f.} = n_1 + n_2 - 2 = 50 + 50 - 2 = 98$$

The decision rule is

Reject H_0 if $t > t_{\alpha, \text{d.f.}} = t_{0.05, 98} \neq t_{0.05, 100} = 1.66$

or, using the *p*-value method, Reject H_0 if *p*-value $< \alpha = 0.05$.

(Note that the critical values from Table 4 in Appendix B does not have d.f. = 98 and we have to use the approximate critical value for d.f. = 100).

Calculating manually

Value of the test statistic: We determined the following statistics:

$$\bar{X}_1 = 6.63$$

$$\bar{X}_2 = 3.72$$

$$s_1^2 = 37.49$$

$$s_2^2 = 43.34$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$= \frac{(50 - 1)(37.49) + (50 - 1)(43.34)}{50 + 50 - 2} = 40.41$$

The value of the test statistic is

$$t = \frac{(6.63 - 3.72) - 0}{\sqrt{40.41 \left(\frac{1}{50} + \frac{1}{50} \right)}} = 2.29$$

Conclusion: As $t = 2.29 > 1.66$ (or, from the Excel output, *p*-value = $0.0122 < 0.05 = \alpha$), reject the null hypothesis.

Interpreting the results

We conclude that at the 5% level of significance there is sufficient evidence to infer that the mean net annual return from directly purchased mutual funds is larger than the mean net annual return of broker-purchased funds. As a result, we conclude that there is sufficient evidence to infer that, on average, directly purchased mutual funds outperform broker-purchased mutual funds.

Using the computer

Using Excel Data Analysis

When raw data are available, we can use Data Analysis in EXCEL to perform this task.



Excel output for Example 13.3

	A	B	C
1	t-Test: Two-Sample Assuming Equal Variances		
2			
3		Direct	Broker
4	Mean	6.63	3.72
5	Variance	37.49	43.34
6	Observations	50	50
7	Pooled Variance	40.41	
8	Hypothesized Mean Difference	0	
9	df	98	
10	t Stat	2.29	
11	P(T<=t) one-tail	0.0122	
12	t Critical one-tail	1.6606	
13	P(T<=t) two-tail	0.0243	
14	t Critical two-tail	1.9845	

COMMANDS

- 1 Type the data in two columns or open the data file (**XM13-03**).
- 2 Click **DATA**, **Data Analysis**, and **t-Test: Two-Sample Assuming Equal Variances**.
- 3 Specify the **Variable 1 Range** (**A1:A51**) and the **Variable 2 Range** (**B1:B51**). Type the value of the **Hypothesized Mean Difference** (**0**) and type a value for α (**0.05**).

Using XLSTAT

Alternatively, when raw data are available, we can use XLSTAT to perform this task.

	B	C	D	E	F	G
1	Hypothesized difference (D): 0					
2	Significance level (%): 5					
3						
4	Summary statistics:					
5	Variable	Observations	Minimum	Maximum	Mean	Std. deviation
6	Direct	50	-6.540	18.450	6.631	6.123
7	Broker	50	-11.070	18.220	3.723	6.583
8						
9	t-test for two independent samples / Lower-tailed test:					
10	Difference	2.908				
11	t (Observed value)	2.287				
12	t (Critical value)	1.661				
13	DF	97.49				
14	p-value (one-tailed)	0.012				
15	alpha	0.05				

COMMANDS

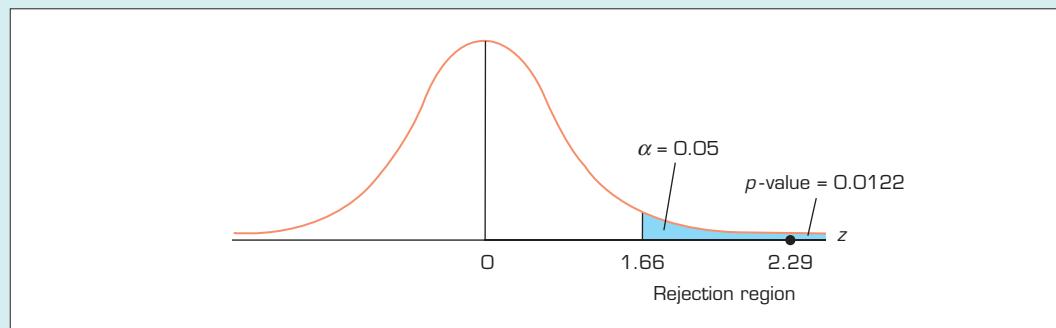
- 1 Type the data in two columns or open the data file (**XM13-03**).
- 2 Click **XLSTAT**, **Parametric test**, and **Two-sample t-test and z-test**.
- 3 Check **One-column per sample**. Type the input range for both samples: **Sample 1** (**A1:A51**), **Sample 2** (**B1:B51**). Click **Student t-test**. Do not click z-test.
- 4 Click the **Options** tab and choose **Mean 1 – Mean 2 > D** in the **Alternative hypothesis** box. Type the **Hypothesized difference (D)** (**0**). Click **Assume equality** under **Population variances for the t-test**. (If you click **Use an F-test** you do not need to conduct a separate F-test of the two variances as a first step to testing the difference between two means.) Type the value of α in the **Significance level (%)** box (**5**). Click **OK**.



The value of the test statistic t Stat is 2.2872. Because this is a one-tail test, the p -value is 0.0122. As p -value = 0.0122 < 0.05, reject the null hypothesis.

Figure 13.4 describes the sampling distribution of the test statistic.

FIGURE 13.4 Sampling distribution for Example 13.3



Here is a summary of how we recognise when to use the **equal-variances t -test** of $\mu_1 - \mu_2$.

IN SUMMARY

Factors that identify the equal-variances t -test of $\mu_1 - \mu_2$

- 1 *Problem objective:* to compare two populations
- 2 *Data type:* numerical (quantitative)
- 3 *Descriptive measurement:* central location
- 4 *Experimental design:* independent samples
- 5 *Population variances:* unknown but equal

equal variances t -test for $\mu_1 - \mu_2$

This test assesses the significance of the difference between two populations when the variances are unknown but are expected to be equal.

13.1b Checking the required condition

The techniques for both equal variances and unequal variances require that the populations are normally distributed. As before, we can check to see if the requirement is satisfied, by drawing histograms of the data. To illustrate, we used Excel to create the histograms for Example 13.3 (**Figures 13.5** and **13.6**). Although the histograms are not bell shaped, it appears that the annual returns are at least approximately normal. Because this technique is robust, we can be confident of the validity of the results.

FIGURE 13.5 Histogram of rates of return for directly purchased mutual funds

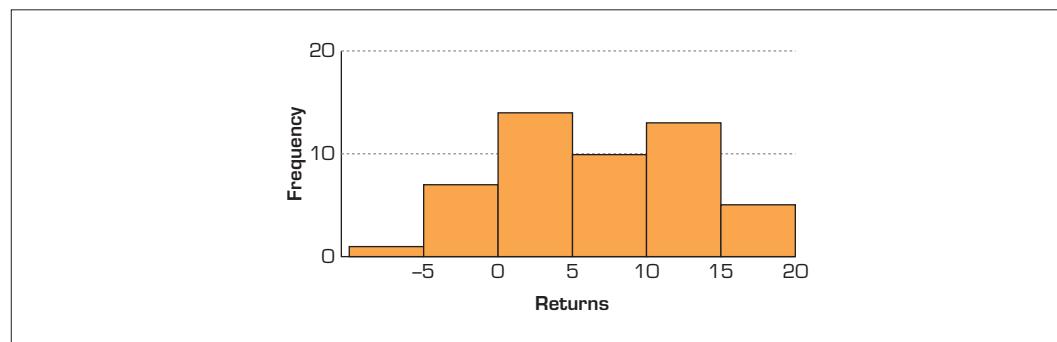


FIGURE 13.6 Histogram of rates of return for broker purchased mutual funds

13.1c Violation of the required condition

When the normality requirement is unsatisfied, we can use a nonparametric technique – the Wilcoxon rank sum test for independent samples – to replace the equal-variances test of $\mu_1 - \mu_2$. We have no alternative to the unequal-variances test of $\mu_1 - \mu_2$ when the populations are very non-normal.

13.1d Data formats

There are two formats for sorting the data when drawing inferences about the difference between two means. As we discuss in Appendix 13.A, we can store the observations of one sample in one column and the observations of the second sample in another column. When the data are stored in this manner, we say that the data are *unstacked*. Alternatively, we can *stack* the data by storing all the observations from both samples in one column and use a second column to store codes identifying the sample from which the observation is drawn. Here is an example of unstacked data:

Column 1 (sample 1)	Column 2 (sample 2)
12	18
19	23
13	25

Here are the same data in stacked form:

Column 1	Column 2
12	1
19	1
13	1
18	2
23	2
25	2

It should be understood that the data need not be in order. Hence, they could have been stored in this way:

Column 1	Column 2
18	2
25	2
13	1
12	1
23	2
19	1

If there are two populations to compare and only one variable, it is probably better to record the data in unstacked form. However, it is frequently the case that we want to observe several variables and compare them. For example, suppose that we survey male and female MBAs and ask each to report his or her income, number of years of education and number of years of experience. These data are usually stored in stacked form using the following format:

Column 1: code identifying female (1) and male (2)

Column 2: income

Column 3: years of education

Column 4: years of experience

To compare incomes between females and males, we would use columns 1 and 2. Columns 1 and 3 are used to compare education level and gender, and columns 1 and 4 are used to compare experience levels and gender.

Most statistical software requires one form or the other. Excel (with the exception of some of our macros) demands that the data must be unstacked. Some of the XLSTAT procedures allow either format, whereas others specify only one. Fortunately, both our software packages allow the statistics practitioner to alter the format. (See Appendix 13.A for details.) We say ‘fortunately’ because this allowed us to store the data in either form on the companion website (accessible through <https://login.cengagebrain.com>). In fact, we have used both forms to allow you to practise manipulating the data as necessary. You will need this ability to perform statistical techniques in this and other chapters in this book.

13.1e Developing an understanding of statistical concepts 1

The formulas in this section are relatively complicated. However, conceptually both test statistics are based on the techniques we introduced in Chapter 11 and repeated in Chapter 12: The value of the test statistic is the difference between the statistic $\bar{X}_1 - \bar{X}_2$ and the hypothesised value of the parameter $\mu_1 - \mu_2$ measured in terms of the standard error.

13.1f Developing an understanding of statistical concepts 2

The standard error must be estimated from the data for all inferential procedures introduced here. The method we use to compute the standard error of $\bar{X}_1 - \bar{X}_2$ depends on whether the population variances are equal. When they are equal, we calculate and use the pooled variance estimator s_p^2 . We are applying an important principle here, and we will do so again in Section 13.3 and in later chapters. The principle can be loosely stated as follows: Where possible, it is advantageous to pool sample data to estimate the standard error. In Example 13.1, we were able to pool because we assume that the two samples were drawn from populations with a common variance. Combining both samples increases the accuracy of the estimate. Thus, s_p^2 is a better estimator of the common variance than either s_1^2 or s_2^2 separately. When the two population variances are unequal, we cannot pool the data and produce a common estimator. We must compute s_1^2 and s_2^2 and use them to estimate σ_1^2 and σ_2^2 respectively.

EXERCISES

Learning the techniques

The following exercises can be solved manually or by using Excel's **Test Statistics** workbook, which is available from the companion website (accessible through <https://login.cengagebrain.com>).

- 13.1** You are given the following information about random samples drawn from two populations:

Sample 1	$n_1 = 50$	$\bar{X}_1 = 52.3$	$\sigma_1 = 6.1$
Sample 2	$n_2 = 100$	$\bar{X}_2 = 49.0$	$\sigma_2 = 7.9$

Test the following hypotheses:

$$H_0: (\mu_1 - \mu_2) = 0$$

$$H_A: (\mu_1 - \mu_2) \neq 0$$

$$\alpha = 0.05$$

- 13.2** In random samples drawn from two normal populations, we found the following statistics:

Sample 1	$n_1 = 15$	$\bar{X}_1 = 140$	$\sigma_1 = 10$
Sample 2	$n_2 = 10$	$\bar{X}_2 = 150$	$\sigma_2 = 15$

Test the following hypotheses:

$$H_0: (\mu_1 - \mu_2) = 0$$

$$H_A: (\mu_1 - \mu_2) < 0$$

$$\alpha = 0.01$$

- 13.3** You are given the following information on random samples drawn from two populations:

Sample 1	$n_1 = 40$	$\bar{X}_1 = 27.3$	$\sigma_1 = 7.2$
Sample 2	$n_2 = 70$	$\bar{X}_2 = 24.6$	$\sigma_2 = 6.9$

Test the following hypotheses:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 - \mu_2 > 0$$

$$\alpha = 0.05$$

- 13.4** What is the *p*-value of the test in Exercise 13.3?

- 13.5** Test the following hypotheses (assume that the two random samples are taken from populations that are normally distributed):

a	Sample 1	$n_1 = 10$	$\bar{X}_1 = 200$	$s_1 = 20$
	Sample 2	$n_2 = 8$	$\bar{X}_2 = 185$	$s_2 = 6.9$

Can we conclude at the 5% level of significance that μ_1 is greater than μ_2 ?

b	Sample 1	$n_1 = 50$	$\bar{X}_1 = 21$	$s_1 = 2$
	Sample 2	$n_2 = 70$	$\bar{X}_2 = 20$	$s_2 = 3$

Can we conclude at the 1% level of significance that μ_1 is not equal to μ_2 ?

- 13.6** Suppose that samples of size $n_1 = 20$ and $n_2 = 15$ are drawn from two normal populations. The sample statistics are as follows:

Sample 1	$n_1 = 20$	$\bar{X}_1 = 100$	$s_1^2 = 225$
Sample 2	$n_2 = 15$	$\bar{X}_2 = 125$	$s_2^2 = 150$

Can we conclude at the 5% level of significance that μ_1 is less than μ_2 ?

- 13.7** Samples of $n_1 = 60$ and $n_2 = 80$ were drawn from two normal populations. The following statistics were produced:

Sample 1	$n_1 = 60$	$\bar{X}_1 = 70.6$	$s_1 = 14.9$
Sample 2	$n_2 = 80$	$\bar{X}_2 = 68.3$	$s_2 = 12.3$

Do these results allow us to conclude that μ_1 is greater than μ_2 ? (Use $\alpha = 0.10$.)

- 13.8** What is the *p*-value of the test in Exercise 13.7?

- 13.9** The following two samples were drawn from two normal populations:

Sample 1	$n_1 = 20$	$\bar{X}_1 = 2.5$	$s_1 = 9.8$
Sample 2	$n_2 = 25$	$\bar{X}_2 = -1.6$	$s_2 = 7.8$

Do these data provide sufficient evidence at the 5% significance level to allow us to conclude that μ_1 is not equal to μ_2 ?

- 13.10** **XR13-10** Samples of size 8 were drawn independently from two normal populations. These data are listed below. Test to determine whether the means of the two populations differ. (Use $\alpha = 0.05$.)

Sample 1	7	4	6	3	7	5	8	7
Sample 2	6	4	5	3	6	5	7	5

- 13.11** In random samples of size 12 from each of two normal populations, we found the following statistics:

Sample 1	$n_1 = 12$	$\bar{X}_1 = 74$	$s_1 = 18$
Sample 2	$n_2 = 12$	$\bar{X}_2 = 71$	$s_2 = 16$

- a** Test with $\alpha = 0.05$ to determine whether we can infer that the population means differ.
- b** Repeat part (a) increasing the standard deviations to $s_1 = 210$ and $s_2 = 198$.
- c** Describe what happens as the sample standard deviations get larger.

- d** Repeat part (a) with samples of size 150.
- e** Discuss the effects of increasing the sample size.
- f** Repeat part (a), changing the mean of sample 1 to $\bar{X}_1 = 76$.
- g** Discuss the effect of increasing \bar{X}_1 .

13.12 Random sampling from two normal populations produced the following results:

Sample 1	$n_1 = 150$	$\bar{X}_1 = 412$	$s_1 = 128$
Sample 2	$n_2 = 150$	$\bar{X}_2 = 405$	$s_2 = 54$

- a** Can we infer at the 5% significance level ($\alpha = 0.05$) that μ_1 is greater than μ_2 ?
- b** Repeat part (a) decreasing the standard deviations to $s_1 = 31$ and $s_2 = 16$.
- c** Describe what happens when the sample standard deviations get smaller.
- d** Repeat part (a) with samples of size 20.
- e** Discuss the effects of decreasing the sample size.
- f** Repeat part (a), changing the mean of sample 1 to $\bar{X}_1 = 409$.
- g** Discuss the effect of decreasing \bar{X}_1 .

13.13 **a** For each of the following, determine the number of degrees of freedom assuming equal population variances and unequal population variances.

- i** $n_1 = 15, n_2 = 15, s_1^2 = 25, s_2^2 = 15$
- ii** $n_1 = 10, n_2 = 16, s_1^2 = 100, s_2^2 = 15$
- iii** $n_1 = 50, n_2 = 50, s_1^2 = 8, s_2^2 = 14$
- iv** $n_1 = 60, n_2 = 45, s_1^2 = 75, s_2^2 = 10$

b Confirm that in each case, the number of degrees of freedom for the equal-variances test statistic is larger than that for the unequal-variances test statistic.

Applying the techniques

13.14 XR13-14 Self-correcting exercise. A baby-food producer claims that her product is superior to that of her leading competitor, in that babies gain weight faster with her product. As an experiment, 10 healthy newborn infants are randomly selected. For two months, five of the babies are fed the producer's product and the other five are fed the competitor's product. Each baby's weight gain (in grams) is shown in the table.

Weight gain (grams)	
Producer's product	Competitor's product
900	960

1080	720
840	900
1110	870
1200	810

Can we conclude that the average weight gain for babies fed on the producer's baby food is greater than the average weight gain for babies fed on the competitor's food? (Use $\alpha = 0.05$.)

13.15 Kool Kat, a manufacturer of vehicle air-conditioners, is considering switching its supplier of condensers. Supplier A, the current supplier of condensers for Kool Kat, prices its product 5% higher than that of supplier B. Kool Kat wants to maintain its reputation for quality, so it wants to be sure that supplier B's condensers last at least as long as those of supplier A. The management of Kool Kat has decided to retain supplier A if there is sufficient statistical evidence that supplier A's condensers last longer, on average, than those of supplier B. In an experiment, 10 mid-size cars were equipped with air-conditioners using supplier A's condensers, while another 10 mid-size cars were equipped with supplier B's condensers. The number of kilometres driven (with the air conditioner on continuously) by each car before the condenser broke down was recorded, and the relevant statistics are listed below:

Supplier A	$n_A = 10$	$\bar{X}_A = 75000$	$s_A = 6000$
Supplier B	$n_B = 10$	$\bar{X}_B = 70000$	$s_B = 5000$

Assuming that the distance travelled is normally distributed, should Kool Kat retain supplier A? (Use $\alpha = 0.10$.)

13.16 Do students doing three-year degrees at a university work harder than those doing two-year degrees at a TAFE college? To help answer this question, 47 randomly selected university students and 36 TAFE college students were asked how many hours per week they spent doing homework. The means and variances for both groups are shown below.

University students	$n_1 = 47$	$\bar{X}_1 = 18.6$	$s_1^2 = 22.4$
TAFE college students	$n_2 = 36$	$\bar{X}_2 = 14.7$	$s_2^2 = 20.9$

- a** Do these results allow us to answer the opening question affirmatively? (Use $\alpha = 0.01$.)
- b** What is the p -value of the test?

- 13.17** High blood pressure is a leading cause of strokes. Medical researchers are constantly seeking ways to treat patients suffering from this condition. A specialist in hypertension claims that regular aerobic exercise can reduce high blood pressure just as successfully as drugs, with none of the adverse side effects. To test the claim, 50 patients who suffer from high blood pressure were chosen to participate in an experiment. For 60 days, half the sample exercised three times per week for one hour, and the other half took the standard medication. The percentage reduction in blood pressure was recorded for each individual, and the resulting data are shown in the accompanying table. Can we conclude at the 1% significance level that exercise is at least as effective as medication in reducing hypertension?

Percentage reduction in blood pressure

Exercise	$n_1 = 25$	$\bar{X}_1 = 14.31$	$s_1 = 1.63$
Medication	$n_2 = 25$	$\bar{X}_2 = 13.28$	$s_2 = 1.82$

- 13.18** Disposable batteries are expensive and they release dangerous chemicals when discarded. Consequently, many people are now using rechargeable batteries. Some rechargeable batteries, however, do not accept a full charge, and as a result do not function as well as others. In an experiment to determine which batteries accept charge better, 100 D-cells made by firm A and another 100 D-cells made by firm B were randomly selected. Each battery was charged for 14 hours. The number of volts of power each was capable of producing was measured (D-cells are supposed to produce 1.25 volts). The results are summarised in the following table. Do these results allow us to conclude at the 1% significance level that the mean power differs between the two brands of batteries?

Firm A	$n_A = 100$	$\bar{X}_A = 1.16$	$s_A = 0.08$
Firm B	$n_B = 100$	$\bar{X}_B = 1.21$	$s_B = 0.10$

- 13.19** **XR13-19** A human resources manager for a car company wanted to know whether production-line workers are absent on more days than office workers. He took a random sample of eight workers from each category and recorded the number of days they had been absent the previous year. Can we infer that there is a difference in days absent between the two groups of workers?

Production-line workers	4	0	6	8	3	11	13	5
Office workers	9	2	7	1	4	7	9	8

- 13.20** **XR13-20** The owner of a small book-publishing company is concerned about the declining number of people who read books. To learn more about the problem, she takes a random sample of customers in a retail book store and asked each how many books they read in the last 12 months. The following figures were recorded. Is there enough evidence to conclude that there are differences in the number of books purchased by females and males?

Females	5	18	11	3	7	5	9	13	15
Males	9	7	9	3	6	5			

- 13.21** **XR13-21** Every month, a clothing store conducts an inventory and calculates the losses due to theft. The store would like to reduce these losses and is considering two methods to use. The first is to hire a security guard, and the second is to install cameras. To help decide which method to choose, they hired a security guard for six months, and then installed cameras for the following six months. The monthly losses were as follows:

Security guard	355	284	401	398	477	254
Cameras	486	303	270	386	411	435

As the cameras are cheaper than the guard, the manager would prefer to install the cameras unless there was enough evidence to infer that the guard was better. What should the manager do? (Use $\alpha = 0.10$.)

- 13.22** **XR13-22** Many people who own digital cameras prefer to have their photos printed. In a preliminary study to determine spending patterns, a random sample of 8 digital camera owners (Sample 1) and 8 standard camera owners (Sample 2) were asked how many photos they had printed in the past month. The results are as follows:

Sample 1	15	12	23	31	20	14	12	19
Sample 2	0	24	36	24	0	48	0	0

Can we infer at the 10% level of significance that the two groups differ in the number of photos that they have printed?

- 13.23** **XR13-23** Random samples were drawn from each of two populations. The data are stored in rows 1

(sample 1) and 2 (sample 2). A partial listing of the data is exhibited below. Is there sufficient evidence at the 5% significance level to infer that the mean of population 1 is greater than the mean of population 2?

Sample 1	110	115	115	118	...	105	114	115
Sample 2	67	82	46	120	...	108	89	73

Sample statistics: $n_1 = 25$, $\bar{X}_1 = 101.68$, $s_1 = 19.07$; $n_2 = 25$, $\bar{X}_2 = 80.32$, $s_2 = 25.14$.

- 13.24 XR13-24** The data obtained from sampling two populations are recorded. (Column 1 contains the data, and column 2 specifies the sample.) Some of these data are shown below (in rows).

Observations	25	15	38	...	39	-3	26
Sample	1	1	1	...	2	2	2

- a Conduct a test to determine whether the population means differ. (Use $\alpha = 0.05$.)
- b What is the required condition(s) of the techniques employed in part (a)?
- c Check to ensure that the required condition(s) is satisfied.

Sample statistics: $n_1 = 100$, $\bar{X}_1 = 19.07$, $s_1 = 9.57$; $n_2 = 140$, $\bar{X}_2 = 16.38$, $s_2 = 25.16$.

- 13.25 XR13-25** A statistics practitioner wants to compare the relative success of two large department-store chains in Brisbane. She decides to measure the sales per square metre of space each store uses. She takes a random sample of 50 stores from chain 1 and 50 stores from chain 2 and records the data. Do these data provide sufficient evidence to indicate that the average sales per square metre for chain 1's stores are \$3 more than those of chain 2's stores at the 10% level of significance?

Sample statistics: $n_1 = 50$, $\bar{X}_1 = 75.51$, $s_1^2 = 133.65$; $n_2 = 50$, $\bar{X}_2 = 71.66$, $s_2^2 = 52.73$.

- 13.26 XR13-26** Is eating oat bran an effective way to reduce cholesterol? Early studies indicated that eating oat bran daily reduces cholesterol levels by 5–10%. Reports of this study resulted in the introduction of many new breakfast cereals with various percentages of oat bran as an ingredient. However, an experiment performed by a team of medical researchers cast doubt on the effectiveness of oat bran. In that study, 120 volunteers ate oat bran for breakfast, and

another 120 volunteers ate another grain cereal for breakfast. At the end of 6 weeks, the percentage of cholesterol reduction was computed for both groups. Can we infer that oat bran is different from other cereals in terms of cholesterol reduction?

Sample statistics: $n_1 = 120$, $\bar{X}_1 = 10.01$, $s_1 = 4.43$; $n_2 = 120$, $\bar{X}_2 = 9.12$, $s_2 = 4.45$.

- 13.27 XR13-27** Automobile insurance companies take many factors into consideration when setting rates, including the age of the driver, their marital status and kilometres driven per year. In order to determine the effect of gender, 100 male and 100 female drivers were surveyed and asked how many kilometres they had driven in the past year. The distances (in thousands of kilometres) are stored in stacked format (code 1 = male and code 2 = female). A partial listing of the data is shown below.

Kilometres driven	11.2	9.2	6.4	...	10.3	15.1	7.1
Gender	1	1	1	...	2	2	2

- a Can we conclude at the 5% significance level that male and female drivers differ in the numbers of kilometres driven per year?
- b Check to ensure that the required condition(s) of the techniques used in part (a) is satisfied.

Sample statistics: $n_1 = 100$, $\bar{X}_1 = 10.23$, $s_1 = 2.87$; $n_2 = 100$, $\bar{X}_2 = 9.66$, $s_2 = 2.90$.

- 13.28 XR13-28** The director of a company that manufactures automobile air-conditioners is considering switching his supplier of condensers. The product of supplier A, the current producer of condensers for the manufacturer, is priced 5% higher than that of supplier B. Since the director wants to maintain his company's reputation for quality, he wants to be sure that supplier B's condensers last at least as long as those of supplier A. After a careful analysis, the director decided to retain supplier A if there is sufficient statistical evidence that supplier A's condensers last longer on average than supplier B's condensers. In an experiment, 30 mid-size cars were equipped with air-conditioners using type A condensers, while another 30 mid-size cars were equipped with type B condensers. The number of kilometres (in thousands) driven by each car before the condenser broke down was recorded, and the data stored in unstacked format (column 1 = supplier A, and column 2 = supplier B).

Some of these data are shown below. Should the director retain supplier A? (Use $\alpha = 0.10$.)

Supplier A	156	146	93	...	106	83	125
Supplier B	109	86	75	...	88	115	103

Sample statistics: $n_1 = 30$, $\bar{X}_1 = 115.5$, $s_1 = 21.7$;

$n_2 = 30$, $\bar{X}_2 = 109.4$, $s_2 = 22.4$.

13.29 XR13-29 A statistics lecturer is about to select a statistical software package for her course. One of the most important features, according to the lecturer, is the ease with which students learn to use the software. She has narrowed the selection to two possibilities: software A, a menu-driven statistical package with some high-powered techniques; and software B, a spreadsheet that has the capability of performing most techniques. To help make her decision, she asks 40 statistics students selected at random to choose one of the two packages. She gives each student a statistics problem to solve by computer and the appropriate manual. The amount of time (in minutes) each student needs to complete the assignment was recorded and stored in unstacked format (column 1 = package A, and column 2 = package B). A partial listing of the data is provided below.

- a Can the lecturer conclude from these data that the two software packages differ in the amount of time needed to learn how to use them? (Use a 1% significance level.)
- b Check to see if the required conditions are satisfied.

Package A	88	83	70	105	82	75
Package B	55	57	49	67	

Sample statistics: $n_1 = 24$, $\bar{X}_1 = 74.71$, $s_1 = 24.02$;

$n_2 = 16$, $\bar{X}_2 = 52.50$, $s_2 = 9.04$.

13.30 XR13-30 In assessing the value of radio advertisements, sponsors not only measure the total number of listeners, but also record their ages. The 18–34 age group is considered to spend the most money. To examine the issue, the manager of an FM station commissioned a survey. One objective was to measure the difference in listening habits between the 18–34 and 35–50 age groups. The survey asked 250 people in each age category how much time they spent listening to FM radio per day. The results (in minutes) were recorded (column

1: listening times, and column 2 identifies the age group: 1 = 18–34 and 2 = 35–50). A partial listing of the data is shown below.

Listening times	75	30	50	...	135	50	0
Age group	1	1	1	...	2	2	2

- a Can we conclude at the 5% significance level that a difference exists between the two groups?
- b Are the required conditions satisfied for the techniques you used in part (a)?

Sample statistics: $n_1 = 250$, $\bar{X}_1 = 59.0$, $s_1 = 30.8$;

$n_2 = 250$, $\bar{X}_2 = 53.0$, $s_2 = 43.3$.

13.31 XR13-31 One factor in low productivity is the amount of time wasted by workers. Wasted time includes time spent cleaning up mistakes, waiting for more material and equipment, and performing any other activity not related to production. In a project designed to examine the problem, an operations management consultant took a survey of 200 workers in companies that were classified as successful (on the basis of their latest annual profits) and another 200 workers from unsuccessful companies. The amount of time (in hours) wasted during a standard 40-hour work week was recorded for each worker. These data are stored in columns 1 (successful companies) and 2 (unsuccessful companies) of the data file. A partial listing of the data follows. Do these data provide enough evidence at the 1% significance level to infer that the amount of time wasted in unsuccessful firms exceeds that of successful ones?

Successful company	5.8	2.0	6.5	...	4.1	2.0	5.3
Unsuccessful company	7.6	2.7	10.1	...	5.8	8.3	0.8

Sample statistics: $n_1 = 200$, $\bar{X}_1 = 5.02$, $s_1 = 1.39$;

$n_2 = 200$, $\bar{X}_2 = 7.80$, $s_2 = 3.09$.

13.32 XR13-32 The data obtained from sampling two populations are recorded (caution: missing data).

- a Conduct a test to determine whether the population means differ (use $\alpha = 0.05$).
- b Estimate the difference in population means with 95% confidence.
- c What is/are the required condition(s) of the techniques employed in parts (a) and (b)?
- d Is/are the required condition(s) satisfied?

Sample statistics: $n_1 = 165$, $\bar{X}_1 = 99.30$, $s_1 = 23.80$;
 $n_2 = 217$, $\bar{X}_2 = 95.77$, $s_2 = 23.74$.

13.33 XR13-33 Random samples were drawn from each of two populations. The data are stored in stacked format.

- a Is there sufficient evidence at the 10% significance level to infer that the mean of population 1 is greater than the mean of population 2?
- b Estimate with 90% confidence the difference between the two population means.
- c What is/are the required condition(s) of the techniques employed in parts (a) and (b)?
- d Is/are the required condition(s) satisfied?

Sample statistics: $n_1 = 121$, $\bar{X}_1 = 21.51$, $s_1 = 4.76$;
 $n_2 = 84$, $\bar{X}_2 = 19.76$, $s_2 = 4.13$.

13.34 XR13-34 It is often useful for companies to know who their customers are and how they became customers. In a study of credit card use, a random sample of customers who applied for the credit card and a random sample of credit card holders who were contacted by telemarketers were drawn. The total purchases made by each in the past month were recorded. Can we conclude from these data that differences exist between the two types of customers? (Use $\alpha = 0.05$.)

Sample statistics: $n_1 = 100$, $\bar{X}_1 = 130.93$, $s_1 = 31.99$;
 $n_2 = 100$, $\bar{X}_2 = 126.14$, $s_2 = 26.00$.

13.35 XR13-35 A number of studies seem to indicate that using a mobile phone while driving is dangerous. One reason for this is that a driver's reaction times may be slower while he or she is talking on the phone. Researchers at a university measured the reaction times of a sample of 270 drivers who

owned a mobile phone. Of these drivers, the driving ability of a sample of 125 drivers was tested while on the phone, while 145 drivers were tested when not on the phone. The reaction times are recorded. Can we conclude that reaction times are slower for drivers using mobile phones (use $\alpha = 0.05$)?

Sample statistics:

On the phone: $n_1 = 125$, $\bar{X}_1 = 0.646$, $s_1 = 0.045$.

Not on the phone: $n_2 = 145$, $\bar{X}_2 = 0.601$, $s_2 = 0.053$.

13.36 XR13-36 Refer to exercise 13.35. To determine whether the type of phone usage affects reaction times when driving, another study was launched. A group of drivers was asked to participate in a discussion while driving. Half the group engaged in simple chitchat on the phone while driving, and the other half participated in a political discussion on the phone while driving. Once again, reaction times were measured and recorded. Can we infer at the 5% level of significance that the type of telephone discussion affects reaction times?

Sample statistics:

Chitchat: $n_1 = 95$, $\bar{X}_1 = 0.654$, $s_1 = 0.048$.

Political: $n_2 = 90$, $\bar{X}_2 = 0.662$, $s_2 = 0.045$.

13.37 XR13-37 Which fast-food drive-through window is faster – that of chain A or chain B? To answer the question, a random sample of service times (in seconds) for each restaurant was measured and recorded. Can we infer from these data, at the 5% level of significance, that there are differences in service times between the two chains?

Sample statistics:

Chain A: $n_A = 213$, $\bar{X}_A = 149.85$, $s_A = 21.82$.

Chain B: $n_B = 202$, $\bar{X}_B = 154.43$, $s_B = 23.64$.

13.2 Testing the difference between two population means: Dependent samples – matched pairs experiment

We continue our presentation of statistical techniques that address the problem of comparing two populations of numerical data. In Section 13.1, the parameter of interest was the difference between two population means, for which the data were generated from independent samples. In Section 11.3, we considered matched pairs experiments, in which samples are not independent, to calculate interval estimates for $\mu_1 - \mu_2$. In this section, the data gathered from a *matched pairs experiment* are used to test hypotheses about $\mu_1 - \mu_2$.

13.2a Experiment with independent samples

To illustrate why matched pairs experiments are needed and how we deal with data produced in this way, consider the following example.

EXAMPLE 13.4

LO3

Comparing new and old tyre designs: Part I

XM13-04 Tyre manufacturers are constantly researching ways to produce tyres that last longer. New innovations are tested by professional drivers on race tracks, as well as by 'ordinary' drivers (as these tests are closer to the conditions experienced by the tyre companies' customers). In order to determine whether a new steel-belted radial tyre lasts longer than the company's current model, two new-design tyres were installed on the rear wheels of 20 randomly selected cars and two existing-design tyres were installed on the rear wheels of another 20 cars. All drivers were told to drive in their usual way until the tyres wore out. The number of kilometres driven by each driver were recorded and are shown in **Table 13.2**. Can the company infer that the new-design tyre will last longer on average than their existing design? Assume that the two populations of tyre lifetimes are normal.

TABLE 13.2 Distance (in thousands of kilometres) until wear-out

New-design tyre										Existing-design tyre									
70	83	78	46	74	56	74	52	99	57	47	65	59	61	75	65	73	85	97	84
77	84	72	98	81	63	88	69	54	97	72	39	72	91	64	63	79	74	76	43

Solution

Identifying the technique

The objective is to compare two populations of numerical data that are normally distributed. The parameter is the difference between two means $\mu_1 - \mu_2$ (μ_1 = mean distance to wear-out for the new-design tyre, and μ_2 = mean distance to wear-out for the existing-design tyre). Because we want to determine whether the new-design tyre lasts longer, the alternative hypothesis will specify that μ_1 is greater than μ_2 . Calculation of the sample variances ($s_1^2 = 243.4$ and $s_2^2 = 226.8$) allows us to use the equal-variances test statistic.

Hypotheses:

$$H_0: (\mu_1 - \mu_2) = 0$$

$$H_A: (\mu_1 - \mu_2) > 0 \quad (\text{Right one-tail test})$$

Test statistic:

$$t = \frac{(\bar{X}_2 - \bar{X}_1) - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{n_1+n_2-2} \quad (\text{d.f.} = n_1 + n_2 - 2 = 20 + 20 - 2 = 38)$$

Level of significance: $\alpha = 0.05$

Decision rule: Reject H_0 if $t > t_{0.05,38} \approx t_{0.05,40} = 1.684$ or if $p\text{-value} < \alpha = 0.05$

Value of the test statistic and p -value:

Using the computer

Using Excel Data Analysis

The Excel commands are exactly the same as those in Example 13.3.





Excel output for Example 13.4

	A	B	C
1	t-Test: Two-Sample Assuming Equal Variances		
2			
3		New-Design	Existing-Design
4	Mean	73.60	69.20
5	Variance	243.41	226.80
6	Observations	20	20
7	Pooled Variance	235.11	
8	Hypothesized Mean Difference	0	
9	df	38	
10	t Stat	0.91	
11	P(T<=t) one-tail	0.1849	
12	t Critical one-tail	1.6860	
13	P(T<=t) two-tail	0.3699	
14	t Critical two-tail	2.0244	

Using XLSTAT

Alternatively, when raw data are available, we can use XLSTAT to perform this task.

The XLSTAT commands are exactly the same as those in Example 13.3.

	B	C	D	E	F	G
1	Hypothesized difference [D]: 0					
2	Significance level (%): 5					
3						
4	Summary statistics:					
5	Variable	Observations	Minimum	Maximum	Mean	Std. deviation
6	New-Design	20	46.00	99.00	73.60	15.60
7	Existing-Design	20	39.00	97.00	69.20	15.06
8						
9	t-test for two independent samples / Lower-tailed test:					
10	Difference	4.400				
11	t (Observed value)	0.907				
12	t (Critical value)	1.686				
13	DF	37.953				
14	p-value (one-tailed)	0.185				
15	alpha	0.05				

Interpreting the results

The value of the test statistic t Stat = 0.91 < 1.684 and its p -value = 0.18 > 0.05 = α . Therefore we do not reject H_0 , indicating that there is very little evidence to support the hypothesis that the new-design tyre lasts longer on average than the existing-design tyre at the 5% level of significance.

As was the case with some earlier examples, we have some evidence to support the alternative hypothesis, but not enough. Note that the difference in sample means is $(\bar{X}_1 - \bar{X}_2) = (73.6 - 69.2) = 4.4$ (in thousands). However, we judge the difference in sample means in relation to the standard deviation of the sampling distribution. As you can easily calculate,

$$s_p^2 = 235.1$$

and

$$\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = 4.85$$

Consequently, the value of the test statistic is $t = 4.4/4.85 = 0.91$, a value that does not allow us to reject the null hypothesis. We can see that although the difference between the

sample means was quite large (about 4400 km), the variability of the data, as measured by s_p^2 , was also large, resulting in a small test statistic value that led to the conclusion not to reject the null hypothesis.

13.2b Experiment with matched pairs

Now we perform the same experiment in Example 13.4 and test using matched pairs data from dependent samples.

EXAMPLE 13.5

LO5

Comparing new and old tyre designs: Part II

XM13-05 Suppose we repeat the experiment in the following way. On 20 randomly selected cars, one of each type of tyre is installed on the rear wheels and, as before, the cars are driven until the tyres wear out. The number of kilometres until wear-out occurred is shown in **Table 13.3**. Can we conclude from these data that the new-design tyre is superior?

TABLE 13.3 Distance (in thousands of kilometres) until wear-out

Car	New design	Existing design	Car	New design	Existing design
1	65	56	11	108	106
2	72	58	12	98	94
3	110	97	13	91	86
4	70	64	14	92	98
5	90	87	15	94	106
6	95	83	16	70	66
7	69	58	17	75	66
8	70	57	18	48	49
9	82	78	19	79	69
10	70	74	20	86	91

Solution

Identifying the technique

The experiment described in Example 13.4 is one in which the samples are independent. That is, there was no relationship between the observations in one sample and the observations in the second sample. However, in this example the experiment was designed in such a way that each observation in one sample is matched with an observation in the other sample. The matching is conducted by using the same set of cars for each sample. Thus, it is logical to compare the distance until wear-out for both types of tyres for each car. This type of experiment is similar to the *matched pairs experiments* introduced in Chapter 11. Here is how we conduct the test.

For each car, we calculate the matched pairs difference between the distances obtained with each type of tyre, as shown in **Table 13.4**.

**TABLE 13.4** Matched pairs differences

Car	Difference ('000)	Car	Difference ('000)
1	9	11	2
2	14	12	4
3	13	13	5
4	6	14	-6
5	3	15	-12
6	12	16	4
7	11	17	9
8	13	18	-1
9	4	19	10
10	-4	20	-5

The experimental design tells us that the parameter of interest is the *mean of the population of differences*, which we label μ_D .

Hypotheses:

$$H_0: \mu_D = 0$$

$H_A: \mu_D > 0$ (Right one-tail test)

Test statistic:

We have already presented inferential techniques about a single population mean.

Recall that in Chapter 11 we introduced the *t*-test of μ when the population variance is unknown and the population is normally distributed. Thus, to test hypotheses about μ_D , we use the test statistic

$$t = \frac{\bar{X}_D - \mu_D}{s_D / \sqrt{n_D}}, \quad n_D = n_1 = n_2$$

which has a *t* distribution with $n_D - 1$ degrees of freedom, provided that the differences are normally distributed. (Aside from the subscript D, this test statistic is identical to the one presented in Chapter 11.) We conduct the test in the usual way.

Level of significance: $\alpha = 0.05$

Decision rule: Reject H_0 if $t > t_{\alpha, n_{D-1}} = t_{0.05, 19} = 1.729$.

Alternatively, using the *p*-value method, reject H_0 if *p*-value $< \alpha = 0.05$.

Calculating manually

Value of the test statistic: Using the differences calculated above, we found the following statistics:

$$\bar{X}_D = 4.55$$

$$s_D = 7.22$$

Therefore, the value of the test statistic is:

$$t = \frac{4.55 - 0}{7.22 / \sqrt{20}} = 2.82$$

Conclusion: As $t = 2.82 > 1.729$, reject the null hypothesis.

Incidentally, data from matched pairs experiments are almost always stored in unstacked form. We sometimes include a third column identifying the matched pair.

Interpreting the results

With this experimental design we have enough statistical evidence to infer that the new type of tyre is superior to the existing type. As is the case with other techniques, we must check that the required conditions are satisfied, and that the sampling procedure is reliable. If the technique is valid, we can use this result to authorise a larger matched pairs experiment. If a similar conclusion is reached, we can use it to launch an effective advertising campaign.



Using the computer

Using Excel Data Analysis

When raw data are available, **Data Analysis** in Excel can be used to perform this task.

Excel output for Example 13.5

	A	B	C
1	t-Test: Paired Two Sample for Means		
2			
3		New-Design	Existing-Design
4	Mean	81.70	77.15
5	Variance	244.01	318.77
6	Observations	20	20
7	Pearson Correlation	0.915	
8	Hypothesized Mean Difference	0	
9	df	19	
10	t Stat	2.818	
11	P(T<=t) one-tail	0.005	
12	t Critical one-tail	1.729	
13	P(T<=t) two-tail	0.011	
14	t Critical two-tail	2.093	

COMMANDS

- 1 Type the data in two columns or open the data file (**XM13-05**).
- 2 Click **DATA**, **Data Analysis**, and **t-Test: Paired Two-Sample for Means**.
- 3 Specify the **Variable 1 Range (A1:A21)** and the **Variable 2 Range (B1:B21)**. Type the value of the **Hypothesized Mean Difference (0)** and type a value for α (**0.05**).

Alternatively, one can first calculate the difference between X_1 and X_2 to create the difference variable X_D . Calculate the mean and standard deviation of X_D . Then use the **t-test_Mean** worksheet in the **Test Statistics** workbook (see Example 12.5) to test whether $\mu_D = 0$.

Using XLSTAT

Alternatively, when raw data are available, we can use XLSTAT to perform this task.

COMMANDS

Follow the commands for Example 13.2 (page 539). Under **Data format**: click **Paired samples**.

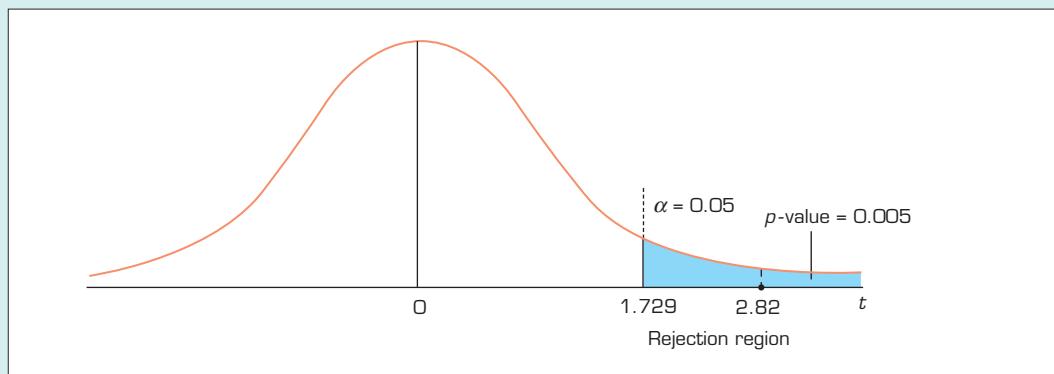
	B	C	D	E	F	G
1	Hypothesized difference (D): 0					
2	Significance level (%): 5					
3						
4	Summary statistics:					
5	Variable	Observations	Minimum	Maximum	Mean	Std. deviation
6	New-Design	20	48.00	110.00	81.70	15.62
7	Existing-Design	20	49.00	106.00	77.15	17.85
8						
9	t-test for two paired samples / Upper-tailed test:					
10	Difference	4.550				
11	t (Observed value)	2.818				
12	t (Critical value)	1.729				
13	DF	19				
14	p-value (one-tailed)	0.005				
15	alpha	0.05				

The value of the test statistic is $2.82 > 1.729$, and its p -value (one-tail) is 0.005, which is less than the level of significance, $\alpha = 0.05$. Thus, we reject the null hypothesis. We can conclude at the 5% level that there is enough statistical evidence to infer that the new type of tyre is superior to the current type.



Figure 13.7 depicts the sampling distribution.

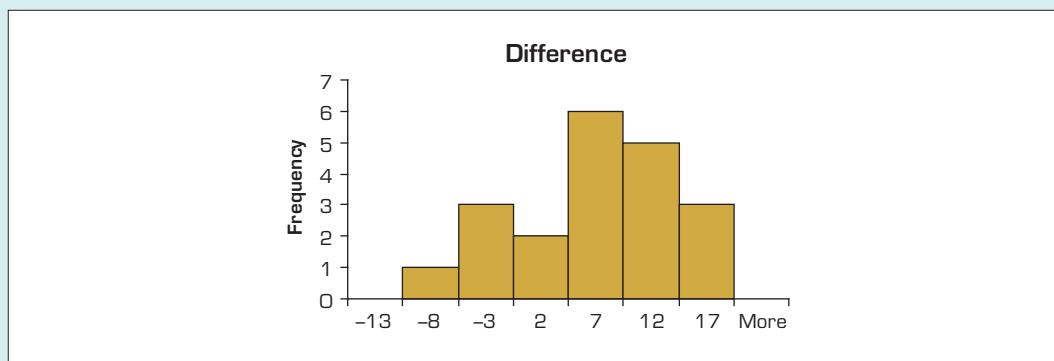
FIGURE 13.7 Sampling distribution for Example 13.5



Checking the required condition

The validity of the results of the t -test of μ_D depends on the normality of the differences. The histogram shown in **Figure 13.8** confirms that assuming normality in this example is reasonable.

FIGURE 13.8 Histogram of matched pairs differences in Example 13.5



13.2c Independent samples or matched pairs: Which experimental design is better?

Examples 13.4 and 13.5 demonstrated that the experimental design is an important factor in statistical inference. However, these two examples raise several questions about experimental designs.

- 1 Why does the matched pairs experiment result in rejecting the null hypothesis, whereas the independent samples experiment could not?
 - 2 Should we always use the matched pairs experiment? In particular, are there disadvantages to its use?
 - 3 How do we recognise when a matched pairs experiment has been performed?
- Here are our answers to these questions:

- 1 The matched pairs experiment in Example 13.5 worked by reducing the variation in the data. To understand this point, examine the statistics for the two examples. In Example 13.4, we found $\bar{X}_1 - \bar{X}_2 = 4.4$. In Example 13.5, we calculated $\bar{X}_D = 4.55$. Thus, the numerators of the two test statistics were almost identical. However, the reason that the test statistic in Example 13.4 was so much smaller than that in Example 13.5 was because of the standard deviations of the sampling distributions. In Example 13.4, we calculated

$$s_p^2 = 235.1 \text{ and the denominator } \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = 4.85$$

Example 13.5 produced

$$s_D^2 = 52.16 \text{ and the denominator } \frac{s_D}{\sqrt{n_D}} = 1.615$$

As you see, the difference in the test statistics was caused not by the numerator but by the denominator. This raises another question: Why was the variation in the data of Example 13.4 so much greater than the variation in the data of Example 13.5? If you examine the data and statistics from Example 13.4, you will find that there was a great deal of variation *between* the cars. That is, some drivers drove in a way that extended the life of the tyres, while others drove faster and braked harder, resulting in shorter tyre lives. This high level of variation made the difference between the sample means appear to be small. As a result, we could not reject the null hypothesis.

Looking at the data from Example 13.5, we see that there is very little variation among the paired differences. Now the variation caused by different driving habits has been markedly decreased. The smaller variation causes the value of the test statistic to be larger. Consequently, we reject the null hypothesis.

- 2 Will the matched pairs experiment always produce a larger test statistic than the independent samples experiment? The answer is *not necessarily*. Suppose that in our example we found that most drivers drove in about the same way and that there was very little difference among drivers in the distances driven until tyre wear-out. In such circumstances, the matched pairs experiment would result in no significant decrease in variation when compared to independent samples. It is possible that the matched pairs experiment might be *less* likely to reject the null hypothesis than the independent samples experiment. The reason can be seen by calculating the degrees of freedom. In Example 13.4, the number of degrees of freedom was 38, whereas in Example 13.5 it was 19. Even though we had the same number of observations (20 in each sample), the matched pairs experiment had half the number of degrees of freedom as the equivalent independent samples experiment. For exactly the same value of the test statistic, a smaller number of degrees of freedom in a Student *t*-distributed test statistic yields a larger *p*-value. What this means is that if there is little reduction in variation to be achieved by the matched pairs experiment, the statistics practitioner should choose instead to conduct the experiment with independent samples.
- 3 As you have seen, in this book we deal with questions arising from experiments that have already been conducted. Thus, one of your tasks is to determine the appropriate test statistic. In the case of comparing two populations of numerical data, you must decide whether the samples are independent (in which case the parameter is $\mu_1 - \mu_2$) or matched pairs (in which case the parameter is μ_D) in order to select the correct test statistic. To help you do so, we suggest you ask and answer the following question: Does some natural relationship exist between *each pair* of observations that provides a logical reason to compare the first observation of sample 1 with the first observation of sample 2, the second observation of sample 1 with the second observation of sample 2, and so on? If so, the experiment was conducted by matched pairs. If not, it was conducted using independent samples.

13.2d Violation of the required condition

If the differences are very non-normal, we cannot use the *t*-test of μ_D . We can, however, employ a nonparametric technique – the Wilcoxon signed rank sum test for matched pairs.

13.2e Developing an understanding of statistical concepts

Two of the most important principles in statistics were applied in this section. The first is the concept of analysing sources of variation. In Examples 13.4 and 13.5, we showed that by reducing the variation among drivers we were able to detect a real difference between tyre brands. This was an application of the more general procedure of analysing data and attributing some fraction of the variation to several sources. In Example 13.5, the two sources of variation were the car drivers and the tyre brands. However, we were not interested in the variation among drivers because we weren't interested in determining whether drivers actually differ. Instead, we merely wanted to eliminate that source of variation, making it easier to determine if tyre brand represented a real source of variation, and thus that one tyre design is superior to another tyre design.

A technique called the *analysis of variance* does what its name suggests: it analyses sources of variation in an attempt to detect real differences. In most applications of this procedure, we will be interested in each source of variation and not simply in reducing one source. We refer to the process as *explaining* the variation. The concept of explaining the variation will be applied in Chapters 15–16.

The second principle demonstrated in this section is that statistics practitioners can design data-gathering procedures in such a way that we can analyse sources of variation. Before conducting the experiment in Example 13.5, the statistics practitioner suspected that there are large differences among drivers in the way they wear out tyres. Consequently he or she set up the experiment so that the effects of those differences were mostly eliminated. It is also possible to design experiments that allow for easy detection of real differences and minimise the costs of data gathering. Unfortunately, we will not present this topic. However, you should understand that the entire subject of the design of experiments is an important one, because managers often need to be able to analyse data to detect differences, and cost is almost always a factor.

Here is a summary of the test statistic and how we determine when to use the inferential techniques about the matched pairs experiment.

Test statistic for μ_D

$$t = \frac{\bar{X}_D - \mu_D}{s_D / \sqrt{n_D}} \quad \text{d.f.} = n_D - 1; \quad n_D = n_1 = n_2$$

IN SUMMARY

Factors that identify the *t*-test and estimator of μ_D

- 1** *Problem objective:* to compare two populations
- 2** *Data type:* numerical (quantitative)
- 3** *Descriptive measurement:* central location
- 4** *Experimental design:* matched pairs

EXERCISES

Learning the techniques

- 13.38 XR13-38** With $\alpha = 0.01$, test these hypotheses (assuming that the two populations are normal):

$$\begin{aligned} H_0: \mu_D &= 0 \\ H_A: \mu_D &\neq 0 \end{aligned}$$

given the following data generated from a matched pairs experiment:

Observation	Sample 1	Sample 2
1	25	32
2	11	14
3	17	16
4	7	14
5	29	36
6	21	22

- 13.39** Test the following hypotheses (assuming that X_D is normally distributed):

Hypotheses	α	n_D	\bar{X}_D	s_D
a $H_0: \mu_D = 0$ $H_A: \mu_D \neq 0$	0.05	15	2	4
b $H_0: \mu_D = 0$ $H_A: \mu_D < 0$	0.01	50	-8	20

- 13.40** Given the following computer output, test these hypotheses at the 10% significance level. Assume that X_D is normally distributed.

$$\begin{aligned} H_0: \mu_D &= 0 \\ H_A: \mu_D &< 0 \end{aligned}$$

Analysis variable X_D :

$$\begin{aligned} n_D &= 18; \bar{X}_D = -0.73615; t = -1.52; \\ p\text{-value (two-tail)} &= 0.1468 \end{aligned}$$

- 13.41** To test these hypotheses:

$$\begin{aligned} H_0: \mu_D &= 0 \\ H_A: \mu_D &> 0 \end{aligned}$$

we generated the following output results.

Analysis variable X_D :

$$\begin{aligned} n_D &= 22; \bar{X}_D = 63.57; t = 1.88; \\ p\text{-value (two-tail)} &= 0.0740 \end{aligned}$$

Interpret the results, and draw a conclusion. (Use $\alpha = 0.05$.) Assume that X_D is normally distributed.

- 13.42** The computer printout of the hypothesis test (assuming that X_D is normal)

$$\begin{aligned} H_0: \mu_D &= 0 \\ H_A: \mu_D &\neq 0 \end{aligned}$$

is shown below. What conclusion would you draw with $\alpha = 0.01$?

Analysis variable X_D :

$$\begin{aligned} n_D &= 60; \bar{X}_D = 12.702; t = 2.53; \\ p\text{-value (two-tail)} &= 0.0140 \end{aligned}$$

Applying the techniques

- 13.43 XR13-43** **Self-correcting exercise.** Refer to Exercise 11.29. Test whether the mean petrol prices in Queensland in 2019 were less than the mean petrol prices in 2017 at the 1% level of significance.

- 13.44 XR13-44** Many people use scanners to read documents and store them in a Word (or some other software) file. To help determine which brand of scanner to buy, a student conducts an experiment in which eight documents are scanned by each of the two scanners in which he is interested. He records the number of errors made by each. These data are listed here. Can he infer that Brand A (the more expensive scanner) is better than Brand B?

Pair	1	2	3	4	5	6	7	8
Brand A	17	29	18	14	21	25	22	29
Brand B	21	38	15	19	22	30	31	37

- 13.45 XR13-45** How effective are antilock brakes, which pump very rapidly rather than lock and thus avoid skids? As a test, a car buyer organised an experiment. He hit the brakes and, using a stopwatch, recorded the number of seconds it took to stop an ABS-equipped car and another identical car without ABS. The speeds when the brakes were applied and the number of seconds each took to stop on dry pavement are listed here. Can we infer that ABS is better?

Speeds	1	2	3	4	5	6	7	8
ABS	3.6	4.1	4.8	5.3	5.9	6.3	6.7	7.0
Non-ABS	3.4	4.0	5.1	5.5	6.4	6.5	6.9	7.3

- 13.46 XR13-46** In a preliminary study to determine whether the installation of a camera designed to catch cars

that go through red lights affects the number of violators, the number of red-light runners was recorded for each day of the week before and after the camera was installed. These data are listed here. Can we infer that the camera reduces the number of red-light runners?

Day	Sun	Mon	Tues	Wed	Thur	Fri	Sat
Before	7	21	27	18	20	24	16
After	8	18	24	19	16	19	16

Computer/manual applications

The following exercises require the use of a computer and software. The answers may be calculated manually using the sample statistics (calculated from the data) provided.

13.47 XR13-47 In a study to determine whether gender affects salary offers for graduating MBA students, 25 pairs of students were selected. Each pair consisted of a female and a male student who were matched according to their overall grade, subjects taken, age and previous work experience. The highest salary offered (in thousands of dollars) to each graduate was recorded. Part of the data is shown below.

- a Is there enough evidence at the 10% significance level to infer that gender is not a factor in salary offers?
- b Discuss why the experiment was organised in the way it was.
- c Is the required condition for the test in part (a) satisfied?

MBA pair	1	2	3	...	23	24	25
Female	71	55	68	...	47	57	46
Male	72	60	70	...	47	58	42

Sample statistics: $n_D = 25$, $\bar{X}_D = -0.72$, $s_D = 3.20$.

13.48 XR13-48 Samples of size 12 were drawn independently from two normal populations. These data are stored in columns 1 and 2 of the data file and part of the data is listed below. A matched pairs experiment was then conducted and 12 pairs of observations were drawn from the same populations. These data are stored in columns 3 and 4, and some of the data are shown below.

- a Using the data taken from independent samples, test to determine whether the means of the two populations differ. (Use $\alpha = 0.05$.)
- b Repeat part (a), using the matched pairs data.
- c Describe the differences between parts (a) and (b). Discuss why these differences occurred.

Independent samples

Sample 1	66	19	88	...	54	79	40
Sample 2	69	37	66	...	61	32	37

Matched pairs

Pair	1	2	3	...	10	11	12
Sample 1	55	45	52	...	60	67	53
Sample 2	48	37	43	...	53	59	37

Independent samples: $n_1 = 12$, $\bar{X}_1 = 59.83$, $s_1 = 20.391$; $n_2 = 12$, $\bar{X}_2 = 50.25$, $s_2 = 20.046$.

Matched pairs: $n_D = 12$, $\bar{X}_D = 25.5$, $s_D = 22.82$.

13.49 XR13-49 Repeat Exercise 13.48 using the data below.

Independent samples

Sample 1	199	261	...	249	218
Sample 2	286	211	...	116	203

Matched pairs

Pair	1	2	...	11	12
Sample 1	218	144	...	256	133
Sample 2	154	160	...	215	117

Independent samples: $n_1 = 12$, $\bar{X}_1 = 201.58$, $s_1 = 58.63$; $n_2 = 12$, $\bar{X}_2 = 176.17$, $s_2 = 62.17$.

Matched pairs: $n_D = 12$, $\bar{X}_D = 25.5$, $s_D = 22.82$.

13.50 XR13-50 Repeat Exercise 13.48 using the data below.

Independent samples

Sample 1	103	86	...	104	99
Sample 2	71	86	...	107	96

Matched pairs

Pair	1	2	...	11	12
Sample 1	91	120	...	87	102
Sample 2	88	75	...	107	98

Discuss what you have discovered from the solutions to Exercises 13.48–13.50.

Independent samples: $n_1 = 12$, $\bar{X}_1 = 101.5$, $s_1 = 10.724$; $n_2 = 12$, $\bar{X}_2 = 92.58$, $s_2 = 10.068$.

Matched pairs: $n_D = 12$, $\bar{X}_D = 9.08$, $s_D = 18.71$.

13.51 XR13-51 Refer to Exercise 11.32. Based on the same sample data, answer the following:

- a Can we infer at the 5% significance level that advertising in the local newspaper improves sales?
- b Check to ensure that the required condition(s) of the technique above is satisfied.

- c Would it be advantageous to perform this experiment with independent samples? Explain why or why not.

Sample statistics: $n_D = 40$, $\bar{X}_D = 29.625$, $s_D = 45.95$.

13.52 XR13-52 Research scientists at a pharmaceutical company have recently developed a new non-prescription sleeping pill. They decide to test its effectiveness by measuring the time it takes for people to fall asleep after taking the pill. Preliminary analysis indicates that the time to fall asleep varies considerably from one person to another. Consequently, they organise the experiment in the following way. A random sample of 50 volunteers who regularly suffer from insomnia is chosen. Each person is given one pill containing the newly developed drug and one placebo. (A placebo is a pill that contains absolutely no medication.) Participants are told to take one pill one night and the second pill one night a week later. (They do not know whether the pill they are taking is the placebo or the real thing, and the order of use is random.)

Each participant is fitted with a device that measures the time until sleep occurs. Some of the results are listed below, and all the data are stored in columns 1 and 2. Can we conclude that the new drug is effective? (Use a 5% significance level.)

Volunteer	1	2	3	...	48	49	50
Drug	38.4	9.8	12.0	...	21.0	15.7	10.2
Placebo	39.2	25.0	26.2	...	27.2	24.5	27.7

Sample statistics: $n_D = 50$, $\bar{X}_D = -3.47$, $s_D = 10.04$.

13.53 XR13-53 The fluctuations in the stock market induce some investors to sell and move their money into more stable investments. To determine the degree to which recent fluctuations affected ownership, a random sample of 170 people who owned some shares were surveyed. The values of the holdings at the end of last year and at the end of the year before were recorded. Can we infer that the value of the shareholdings has decreased?

Sample statistics: $n_D = 170$, $\bar{X}_D = -183.35$, $s_D = 1568.94$.

13.3 Testing the difference between two population proportions

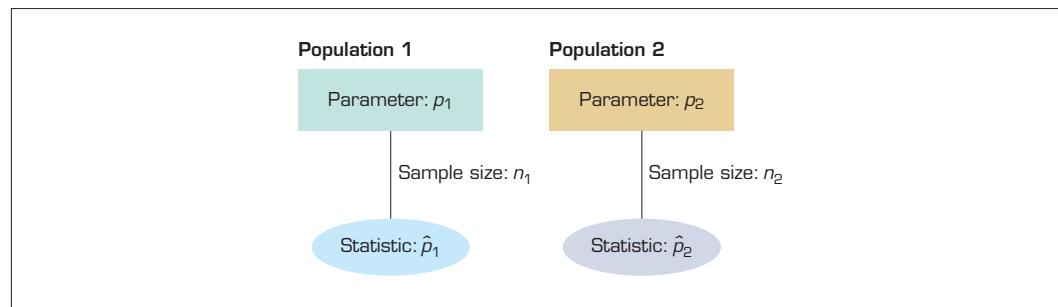
In this section, we present the procedures for drawing inferences about the difference between populations whose data are nominal. The number of applications of these techniques is almost limitless. For example, pharmaceutical companies test new drugs by comparing the new and old or the new versus a placebo. Marketing managers compare market shares before and after advertising campaigns. Operations managers compare defective rates between two machines. Political pollsters measure the difference in popularity before and after an election.

13.3a Parameter

When data are nominal, the only meaningful computation is to count the number of occurrences of each type of outcome and calculate proportions. Consequently, when the problem objective is to compare two populations and the data type is nominal, the parameter to be tested in this section is the difference between the two population proportions, $p_1 - p_2$. There are two different test statistics for this parameter; the choice of which one to use depends on the null hypothesis.

13.3b Sampling distribution

To draw inferences about $p_1 - p_2$, we take a random sample of size n_1 from population 1 and a random sample of size n_2 from population 2 (Figure 13.9 depicts the sampling process).

FIGURE 13.9 Sampling from two populations of nominal data

Sampling distribution of $\hat{p}_1 - \hat{p}_2$

- 1 The statistic $\hat{p}_1 - \hat{p}_2$ is approximately normally distributed provided the sample sizes are large enough so that $n_1 p_1$, $n_1 q_1$, $n_2 p_2$ and $n_2 q_2$ are all greater than or equal to 5. (Since p_1 , q_1 , p_2 and q_2 are unknown, we express the sample size requirement as $n_1 \hat{p}_1$, $n_1 \hat{q}_1$, $n_2 \hat{p}_2$ and $n_2 \hat{q}_2 \geq 5$.)
- 2 The mean of $\hat{p}_1 - \hat{p}_2$ is

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

- 3 The variance of $\hat{p}_1 - \hat{p}_2$ is

$$\text{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$$

where we have assumed that the samples are independent.

The standard deviation is

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

Thus, the variable

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

is approximately standard normally distributed.

13.3c Testing the difference between two proportions

We would like to use the z -statistic just described as our test statistic; however, the standard error of $\hat{p}_1 - \hat{p}_2$,

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

is unknown because both p_1 and p_2 are unknown. As a result, the standard error of $\hat{p}_1 - \hat{p}_2$ must be estimated from the sample data. There are two different estimators of this quantity, and the determination of which one to use depends on the null hypothesis. If the null hypothesis states that $p_1 - p_2 = 0$, the hypothesised equality of the two population proportions allows us to pool the data from the two samples to produce an estimate of the common value of the two proportions p_1 and p_2 .

The sample proportions are

$$\hat{p}_1 = \frac{X_1}{n_1} \text{ (the proportion of successes in sample 1)}$$

$$\hat{p}_2 = \frac{X_2}{n_2} \text{ (the proportion of successes in sample 2)}$$

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} \text{ (the proportion of successes in both samples combined)}$$

$$\hat{q} = 1 - \hat{p}$$

pooled proportion estimate

The weighted average of two sample proportions.

The quantity \hat{p} is called the **pooled proportion estimate**. Since under the null hypothesis we assume that the two population proportions are equal, we can produce a single estimate of that proportion to be used to estimate the standard deviation. As was the case when we calculated the *pooled variance* estimate in Chapter 11, it is better to combine the data from two samples if possible. Notice that under Case 2 (following), it is not possible to combine the sample data to estimate the standard deviation, because we assume there that the two population proportions differ.

Test statistic for $p_1 - p_2$

Case 1: $H_0: p_1 - p_2 = 0$

If the null hypothesis specifies that the difference between the two population proportions is zero ($H_0: p_1 - p_2 = 0$), the test statistic is

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

which is standard normally distributed, provided the sample sizes are sufficiently large such that $n_1\hat{p}_1 \geq 5$, $n_1\hat{q}_1 \geq 5$, $n_2\hat{p}_2 \geq 5$ and $n_2\hat{q}_2 \geq 5$.

Case 2: $H_0: p_1 - p_2 = D$, $D \neq 0$

If the null hypothesis states that the difference between the two population proportions is a non-zero value ($H_0: p_1 - p_2 = D$, where $D \neq 0$), the test statistic is

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - D}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}$$

which is standard normally distributed, provided the sample sizes are sufficiently large such that $n_1\hat{p}_1 \geq 5$, $n_1\hat{q}_1 \geq 5$, $n_2\hat{p}_2 \geq 5$ and $n_2\hat{q}_2 \geq 5$.

REAL-LIFE APPLICATIONS

Marketing: Test marketing

Marketing managers frequently make use of test marketing to assess consumer reaction to a change in a characteristic (such as price or packaging) of an existing product, or to assess consumers' preferences regarding a proposed new product. Test marketing involves experimenting with changes to the marketing mix in a small, limited test market, and assessing consumers' reaction in the test market before undertaking costly changes in production and distribution for the entire market.



Source: © Alamy/Caro

EXAMPLE 13.6

LO5

Test marketing of package designs: Part I

XM13-06 The General Products Company produces and sells a variety of household products. Because of stiff competition, one of its products, a bath soap, is not selling well. Hoping to improve sales, the General Products marketing manager decided to introduce more attractive packaging. The company's advertising agency developed two new designs. The first design features several bright colours to distinguish it from other brands. The second design is light green in colour with just the company's logo on it. As a test to determine which design is better, the marketing manager selected two supermarkets. In one supermarket, the soap was packaged in a box using the first design; in the second supermarket, the second design was used. The product scanner at each supermarket tracked every buyer of soap over a 1-week period. The supermarkets recorded the last four digits of the scanner code for each of the five brands of soap the supermarket sold. The code for the General Products brand of soap is 9077 (the other codes are 4255, 3745, 7118, and 8855). After the trial period, the scanner data were transferred to a computer file. Because the first design is more expensive, management has decided to use this design only if there is sufficient evidence to allow it to conclude that design is better. Should management switch to the brightly coloured design or the simple green one?

Solution

Identifying the technique

The problem objective is to compare two populations. The first is the population of soap sales in supermarket 1, and the second is the population of soap sales in supermarket 2. The data are nominal because the values are 'buy General Products soap' and 'buy other companies' soap.' These two factors tell us that the parameter to be tested is the difference between two population proportions $p_1 - p_2$ (where p_1 and p_2 are the proportions of soap sales that are a General Products brand in supermarkets 1 and 2 respectively). Because we want to know whether there is enough evidence to adopt the brightly coloured design, the alternative hypothesis is

$$H_A: p_1 - p_2 > 0$$

The null hypothesis must be

$$H_0: p_1 - p_2 = 0$$

which tells us that this is an application of Case 1. Thus, the test statistic is

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$



The complete test

Hypotheses:

$$H_0: p_1 - p_2 = 0$$

$$H_A: p_1 - p_2 > 0$$

(Left one-tail test)

Test statistic:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Level of significance:

$$\alpha \text{ (not given)}$$

Decision rule:

A 5% significance level seems to be appropriate. Thus, the rejection region is $z > z_\alpha = z_{0.05} = 1.645$ or $p\text{-value} < \alpha$.

Calculating manually

Value of the test statistic and the p -value:

To compute the test statistic manually requires the statistics practitioner to tally the number of successes in each sample, where success is represented by the code 9077. Reviewing all the sales reveals that

$$x_1 = 180, \quad n_1 = 904, \quad x_2 = 155, \quad n_2 = 1038$$

The sample proportions are

$$\hat{p}_1 = \frac{180}{904} = 0.1991$$

and

$$\hat{p}_2 = \frac{155}{1038} = 0.1493$$

The pooled proportion is

$$\hat{p} = \frac{180 + 155}{904 + 1038} = \frac{335}{1942} = 0.1725; \quad \hat{q} = 1 - \hat{p} = 0.8275$$

The value of the test statistic is

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(0.1991 - 0.1493)}{\sqrt{(0.1725)(0.8275)\left(\frac{1}{904} + \frac{1}{1038}\right)}} = 2.90$$

Therefore, $p\text{-value} = P(Z > 2.90) = 0.0019$

Conclusion: Since $z = 2.90 > 1.645$ or as $p\text{-value} = 0.0019 < 0.05 = \alpha$, reject the null hypothesis.

Interpreting the results

The value of the test statistic is $z = 2.90$; its p -value is 0.0019. There is enough evidence to infer that the brightly coloured design is more popular than the simple design. As a result, it is recommended that management switch to the first design.

Using the computer

Using Excel workbook

Data Analysis in Excel cannot perform this test. We use the **Test Statistics** workbook that is provided on the companion website (accessible through <https://login.cengagebrain.com>) to calculate the value of the test statistic and the p -value.





Excel output for Example 13.6

	A	B	C	D	E
1	z-Test of the Difference Between Two Proportions (Case 1)				
2					
3		Sample 1	Sample 2	z Stat	2.90
4	Sample proportion	0.1991	0.1493	P[Z<=z] one-tail	0.0019
5	Sample size	904	1038	z Critical one-tail	1.6449
6	Alpha	0.05		P[Z<=z] two-tail	0.0038
7				z Critical two-tail	1.9600

COMMANDS

Type the data in two columns or open the data file ([XM13-06](#)). Calculate the sample proportions for each sample. Open the **z-Test_2Proportions (Case 1)** worksheet in the **Test Statistics** workbook and enter the required sample statistics, the sample proportions and sample sizes, and the hypothesised difference.

Using XLSTAT

Alternatively, when raw data are available, we can use XLSTAT to perform this task.

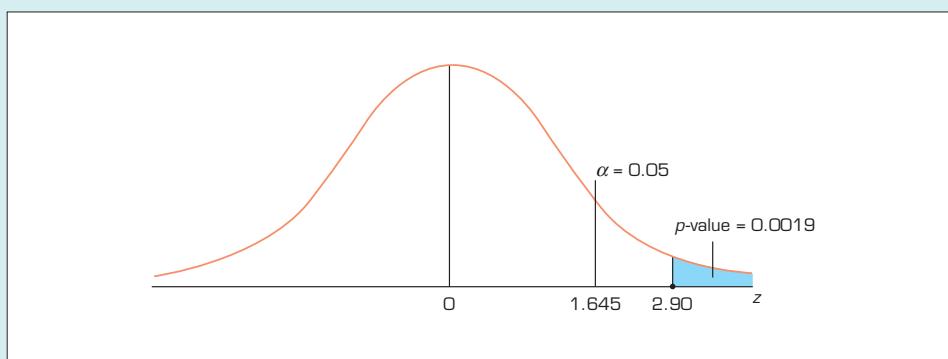
	A	B
1	Proportion 1: 0.1991	
2	Sample size 1: 904	
3	Proportion 2: 0.1493	
4	Sample size 2: 1038	
5	Hypothesized difference (D): 0	
6	Variance: pq(1/n1+1/n2)	
7	Significance level (%): 5	
8		
9	z-test for two proportions / Upper-tailed test:	
10	Difference	0.050
11	z (Observed value)	2.898
12	z (Critical value)	1.645
13	p-value (one-tailed)	0.0019
14	alpha	0.05

COMMANDS

- 1 Type the data in two columns or open the data file ([XM13-06](#)). Compute the frequencies for each sample.
- 2 Click **XLSTAT**, **Parametric tests**, and **Tests for two proportions**.
- 3 Click **Proportions** and input the frequencies and sample sizes. Click **z test**.
- 4 Click **Options** and choose the **Alternative hypothesis: Proportion 1 – Proportion 2 > D** and type the value of D (**0**). Under **Variance** click **pq(1/n1 + 1/n2)**. Click **OK**.

Figure 13.10 describes the sampling distribution of the test statistic.

FIGURE 13.10 Sampling distribution for Example 13.6



EXAMPLE 13.7

LO5

Test marketing of package designs: Part II

XM13-06 Suppose that in Example 13.6 the additional cost of the brightly coloured design requires that it outsell the simple design by more than 3%. Is there enough evidence to conclude that management should switch to the brightly coloured design? Use $\alpha = 0.05$.

Solution**Identifying the technique**

Because we want to determine if there is enough evidence for us to infer that p_1 is 3% more than p_2 , the alternative hypothesis is

$$H_A: p_1 - p_2 > 0.03$$

and the null hypothesis follows:

$$H_0: p_1 - p_2 = 0.03$$

Because the null hypothesis specifies a nonzero difference, we would apply the Case 2 test statistic.

It can be easily verified that $n_1\hat{p}_1$, $n_2\hat{p}_2$, $n_1\hat{q}_1$ and $n_2\hat{q}_2 \geq 5$. Thus, the test statistic

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}$$

is normally distributed.

Here is the complete test:

Hypotheses:	$H_0: p_1 - p_2 = 0.03$
	$H_A: p_1 - p_2 > 0.03$ (Right one-tail test)

Test statistic:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}$$

has a standard normal distribution, as $n_1\hat{p}_1 \geq 5$, $n_1\hat{q}_1 \geq 5$, $n_2\hat{p}_2 \geq 5$ and $n_2\hat{q}_2 \geq 5$.

Level of significance: $\alpha = 0.10$

Rejection region: $Z > z_{\alpha} = z_{0.05} = 1.645$

Decision rule: Reject H_0 if $Z > 1.645$.

Alternatively, reject H_0 if $p\text{-value} < \alpha = 0.05$.

Calculating manually

Value of the test statistic and the p -value:

From Example 13.6, the sample proportions are

$$\hat{p}_1 = 0.1991 \quad \text{and} \quad \hat{p}_2 = 0.1493$$

The value of the test statistic is

$$\begin{aligned} Z &= \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\left(\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2} \right)}} \\ &= \frac{(0.1991 - 0.1493) - 0.03}{\sqrt{\left(\frac{0.1991 \times 0.8009}{904} + \frac{0.1493 \times 0.8507}{1038} \right)}} = 1.15 \end{aligned}$$

Therefore, $p\text{-value} = P(Z > 1.15) = 0.1260$

Conclusion: Since $z = 1.15 < 1.645$ or as $p\text{-value} = 0.1260 > 0.05 = \alpha$, do not reject the null hypothesis.





Interpreting the results

There is not enough evidence to infer that the proportion of soap customers who buy the product with the brightly coloured design is more than 3% higher than the proportion of soap customers who buy the product with the simple design. In the absence of sufficient evidence, the analysis suggests that the product should be packaged using the simple design.

Using the computer

Using Excel workbook

Data Analysis in Excel cannot perform this test. We use the **Test Statistics** workbook to calculate the value of the test statistic and the *p*-value. Follow the same Commands as in Example 13.6.

Excel output for Example 13.7

	A	B	C	D	E
1	z-Test of the Difference Between Two Proportions (Case 2)				
2					
3		Sample 1	Sample 2	z Stat	1.15
4	Sample proportion	0.1991	0.1493	P(Z≤z) one-tail	0.1260
5	Sample size	904	1038	z Critical one-tail	1.6449
6	Hypothesized difference	0.03		P(Z≤z) two-tail	0.2520
7	Alpha	0.05		z Critical two-tail	1.9600

COMMANDS

Type the data in two columns or open the data file ([XM13-06](#)). Calculate the sample proportions for each sample. Open the **Test Statistics** workbook and select the **z-Test_2Proportions (Case 2)** tab. Copy or type the sample proportions, sample sizes, the hypothesised difference (**0.03**) in proportions.

Using XLSTAT

Alternatively, when raw data are available, we can use XLSTAT to perform this task.

	A	B
1	Proportion 1: 0.1991	
2	Sample size 1: 904	
3	Proportion 2: 0.1493	
4	Sample size 2: 1038	
5	Hypothesized difference (D): 0.03	
6	Variance: pq(1/n1+1/n2)	
7	Significance level (%): 5	
8		
9	z-test for two proportions / Upper-tailed test:	
10	Difference	0.050
11	z (Observed value)	1.152
12	z (Critical value)	1.645
13	p-value (one-tailed)	0.125
14	alpha	0.05

COMMANDS

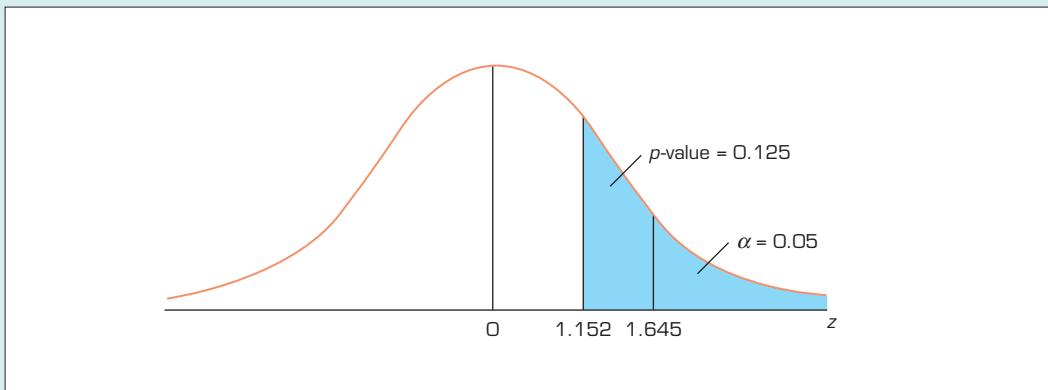
Follow the first three steps shown in Example 13.6. At step 4 click **Options** and choose the **Alternative hypothesis: Proportion 1 – Proportion 2 > D** and type the value of D (**0.03**). Under Variance click **pq(1/n1 + 1/n2)**. Click **OK**.



The value of the test statistic is 1.15 and its p -value is 0.125. As the p -value = 0.125 > 0.05 = α , do not reject H_0 : $p_1 - p_2 = 0.03$.

Figure 13.11 depicts the sampling distribution of this test statistic.

FIGURE 13.11 Sampling distribution for Example 13.7



The critical factors that tell us when to use the z -test of $p_1 - p_2$ are as follows.

IN SUMMARY

Factors that identify the z -test of $p_1 - p_2$

- 1 *Problem objective:* to compare two populations
- 2 *Data type:* nominal (categorical)

EXERCISES

Learning the techniques

The following exercises can be solved manually or by using Excel's **Test Statistics** workbook.

13.54 Test the following hypotheses:

- a $H_0: p_1 - p_2 = 0$
 $H_A: p_1 - p_2 \neq 0$
 $n_1 = 100$ $n_2 = 150$ $\alpha = 0.01$
 $X_1 = 50$ $X_2 = 90$
- b $H_0: p_1 - p_2 = 0.05$
 $H_A: p_1 - p_2 > 0.05$
 $n_1 = 500$ $n_2 = 400$ $\alpha = 0.05$
 $X_1 = 200$ $X_2 = 100$

13.55 Test the following hypotheses:

- $H_0: p_1 - p_2 = 0$
 $H_A: p_1 - p_2 > 0$
 $n_1 = 250$ $n_2 = 400$ $\alpha = 0.01$
 $\hat{p}_1 = 0.24$ $\hat{p}_2 = 0.17$

13.56 Given $n_1 = 40$, $\hat{p}_1 = 0.25$, $n_2 = 50$ and $\hat{p}_2 = 0.32$, test to determine whether there is enough evidence at the 5% significance level to show that $p_2 > p_1$. What is the p -value of the test?

13.57 A random sample of size $n_1 = 1000$ from population 1 produced $X_1 = 500$, and a random sample of size $n_2 = 1500$ from population 2 produced $X_2 = 500$. Can we conclude, with $\alpha = 0.10$, that p_1 exceeds p_2 by at least 0.10?

13.58 Random samples from two binomial populations yielded the following statistics:

$$\begin{array}{ll} \hat{p}_1 = 0.45 & n_1 = 100 \\ \hat{p}_2 = 0.40 & n_2 = 100 \end{array}$$

- a Calculate the p -value of a test to determine whether we can infer that the population proportions differ.
- b Repeat part (a), increasing the sample sizes to 400.

- c Describe what happens to the p -value when the sample sizes increase.

13.59 These statistics were calculated from two random samples:

$$\begin{array}{ll} \hat{p}_1 = 0.60 & n_1 = 225 \\ \hat{p}_2 = 0.55 & n_2 = 225 \end{array}$$

- a Calculate the p -value of a test to determine whether there is evidence to infer that the population proportions differ.
- b Repeat part (a), with $\hat{p}_1 = 0.95$ and $\hat{p}_2 = 0.90$.
- c Describe the effect on the p -value of increasing the sample proportions.
- d Repeat part (a) with $\hat{p}_1 = 0.10$ and $\hat{p}_2 = 0.05$.
- e Describe the effect on the p -value of decreasing the sample proportions.

Applying the techniques

13.60 Self-correcting exercise. In a public opinion survey, 60 out of a sample of 100 high-income voters and 40 out of a sample of 75 low-income voters supported the introduction of a new national security tax in Australia. Can we conclude at the 5% level of significance that there is a difference in the proportion of high- and low-income voters favouring a new national security tax?

13.61 A pharmaceutical company has produced a flu vaccine. In a test of its effectiveness, 1000 people were randomly selected; 500 were injected with the vaccine and the other 500 were untreated. The number of people in each group who contracted the flu during the next three months is summarised in the following table:

Condition	Number of people	
	Treated with vaccine	Untreated
Developed the flu	80	120
Did not develop the flu	420	380

Do these data provide sufficient evidence that the vaccine is effective in preventing the flu?
(Use $\alpha = 0.05$.)

13.62 In a random sample of 500 television sets from a large production line, there were 80 defective sets. In a random sample of 200 television sets from a second production line, there were 10 defective sets. Do these data provide sufficient evidence to establish that the proportion of defective sets from the first line exceeds the proportion of defective

sets from the second line by at least 3%?
(Use $\alpha = 0.05$.)

13.63 A survey to study the usefulness of the online zoom meetings at workplaces used a sample of 1320 managers, professionals and executives. This group included 528 users of Zoom meetings, 408 people who did not use zoom meetings, and 384 who said they would start using zoom meetings relatively soon. Asked if availability of zoom meetings would boost personal productivity, 50% of the non-users and intenders said 'no', as did 25% of the users. Can we conclude from these data, at the 1% level of significance, that users and non-users (including intenders) differ in their opinion of the usefulness of zoom meetings?

13.64 Surveys have been widely used by politicians around the world as a way of monitoring the opinions of the electorate. Six months ago, a survey was undertaken to determine the degree of support for a National Party politician. Of a sample of 1100 voters, 56% indicated that they would vote for this politician. This month, another survey of 800 voters revealed that 46% now support the politician.

- a At the 5% significance level, can we infer that the politician's popularity has decreased?
- b At the 5% significance level, can we infer that the politician's popularity has decreased by more than 5%?
- c Estimate the decrease in percentage support between now and six months ago.

13.65 Many stores sell extended warranties for products they sell. These are very lucrative for store owners. To learn more about who buys these warranties a random sample was drawn from a store's customers who recently purchased a product for which an extended warranty was available. Among other variables, each respondent reported whether they paid the regular price or a sale price and whether they purchased an extended warranty. The results are summarised in the table below. Can we conclude at the 10% significance level that those who paid the regular price are more likely to buy an extended warranty?

	Regular price	Sale price
Sample size	229	178
Number who bought extended warranty	47	25

- 13.66** One hundred normal-weight people and 100 obese people were observed at several Chinese-food buffets. For each diner researchers recorded whether the diner used chopsticks or knife and fork. The table shown here was created.

	Normal weight	Obese
Used chop sticks	26	7
Used knife and fork	74	93

Is there sufficient evidence at the 10% significance level to conclude that obese Chinese food eaters are less likely to use chopsticks?

(Source: Brian Wansink and Collin R. Payne, *The Cues and Correlates of Overeating at the Chinese Buffet*, Cornell University Food and Brand Lab working paper.)

Computer/manual applications

The following exercises require the use of a computer and software. The answers may be calculated manually using the sample statistics (calculated from the data) provided.

- 13.67 XR13-67** Refer to Exercise 11.42. Test at the 1% significance level to determine whether we can infer that the two population proportions of success differ.

Sample statistics: $n_1(0) = 301$, $n_1(1) = 699$; $n_2(0) = 156$, $n_2(1) = 444$.

- 13.68 XR13-68** Refer to Exercise 11.43.

- a Do these data allow us to infer at the 1% significance level that p_1 is greater than p_2 ?
- b Do these data allow us to infer at the 1% significance level that p_1 exceeds p_2 by more than 3%?

Sample statistics: $n_1(0) = 268$, $n_1(1) = 232$; $n_2(0) = 311$, $n_2(1) = 189$.

- 13.69 XR13-69** Refer to Exercise 11.44. Can the company conclude at the 5% significance level that smokers

have a higher incidence of heart disease than non-smokers?

Sample statistics: $n_1(0) = 37$, $n_1(1) = 19$; $n_2(0) = 119$, $n_2(1) = 25$.

- 13.70 XR13-70** Refer to Exercise 11.45. Can we conclude at the 5% significance level that men and women differ in their use of this oil-change service?

Sample statistics: $n_1(0) = 171$, $n_1(1) = 67$; $n_2(0) = 176$, $n_2(1) = 86$.

- 13.71 XR13-71** The operations manager of a maker of computer chips is in the process of selecting a new machine to replace several older ones. Although technological innovations have improved the production process, it is quite common for the machines to produce defective chips. The operations manager must choose between two machines. The cost of machine A is several thousand dollars greater than the cost of machine B. After an analysis of the costs, it was determined that machine A is warranted provided that the defect rate of machine B is more than 2% greater than that of machine A. To help decide, each machine is used to produce 200 chips. Each chip was examined, and whether it was defective (code = 2) or not (code = 1) was recorded. Should the operations manager select machine A? (Use $\alpha = 0.05$.)

Sample statistics: A: $n_1(1) = 189$, $n_1(2) = 11$; B: $n_2(1) = 178$, $n_2(2) = 22$.

- 13.72 XR13-72** Refer to Exercise 11.49. Can we infer at the 10% significance level that there has been a decrease in belief in the greenhouse effect?

Sample statistics:

Two years ago: $n_1(1) = 152$, $n_1(2) = 248$.

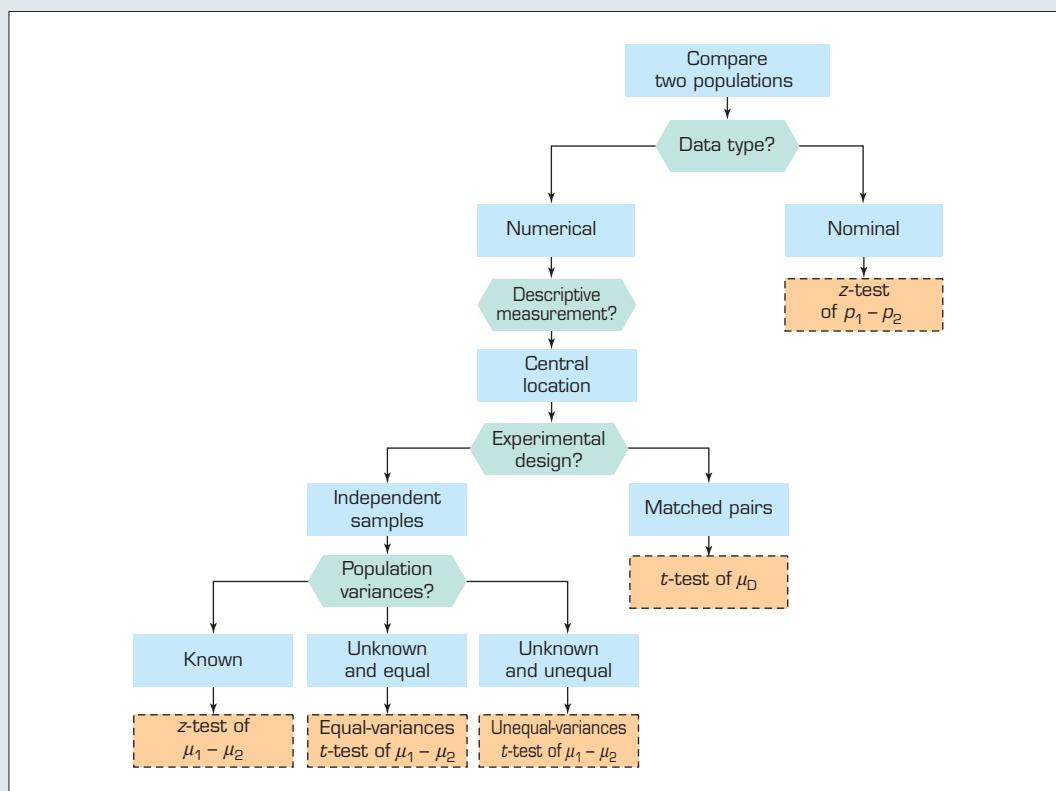
This year: $n_2(1) = 240$, $n_2(2) = 260$.

Study Tools

CHAPTER SUMMARY

In this chapter, we presented a variety of techniques that allow statistics practitioners to compare two populations. When the data are quantitative (numerical) and we are interested in measures of central location, we encountered two more factors that must be considered when choosing the appropriate technique. When the samples are independent, we can use either the equal-variances or unequal-variances formulas. When the samples are matched pairs, we have only one set of formulas. When the data are nominal, the parameter of interest is the difference between two proportions. For this parameter, we had two test statistics. Finally, we discussed observational and experimental data, important concepts in attempting to interpret statistical findings.

The techniques used to test for differences between two populations were described in this chapter. When the data type is numerical, the parameter we test is $\mu_1 - \mu_2$. The three test statistics for this parameter are listed in **Table 13.5**. When the two population variances are known, the two populations are normal or the sample sizes are large, and the two samples are independent, the z-statistic is used. When the two population variances are unknown, the two populations are normal and the two samples are independent, the t-statistic allows us to draw inferences. If the normally distributed data are generated from a *matched pairs experiment*, we estimate and test the mean difference μ_D using a t test, which is equivalent to $\mu_1 - \mu_2$. When the data type is nominal, the parameter of interest is $p_1 - p_2$. The two test statistics used for this parameter are also shown in **Table 13.5**.



The data files for Examples, Exercises and Cases are provided on the companion website (accessible through <https://login.cengagebrain.com/>).

SUMMARY OF FORMULAS

TABLE 13.5 Summary of test statistics for $\mu_1 - \mu_2$ and $p_1 - p_2$

Parameter	Test statistic	Required conditions
$\mu_1 - \mu_2$	$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	σ_1^2 and σ_2^2 are known; X_1 and X_2 are normally distributed or $n_1 \geq 30$ and $n_2 \geq 30$; samples are independent
$\mu_1 - \mu_2$	$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ $\text{d.f.} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$	σ_1^2 and σ_2^2 are unknown and not equal $\sigma_1^2 \neq \sigma_2^2$; X_1 and X_2 are normally distributed; samples are independent
$\mu_1 - \mu_2$	$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$ $\text{d.f.} = n_1 + n_2 - 2$	σ_1^2 and σ_2^2 are unknown and equal $\sigma_1^2 = \sigma_2^2$; X_1 and X_2 are normally distributed; samples are independent
$\mu_D = \mu_1 - \mu_2$	$t = \frac{\bar{X}_D - \mu_D}{s_D / \sqrt{n_D}}$ $\text{d.f.} = n_D - 1$	$X_D = X_1 - X_2$ is normally distributed; samples are not independent; samples are matched pairs
$p_1 - p_2$	$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}\hat{q} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$	The null hypothesis is $H_0: (p_1 - p_2) = 0$; $n_1\hat{p}_1$, $n_1\hat{q}_1$, $n_2\hat{p}_2$ and $n_2\hat{q}_2 \geq 5$
$p_1 - p_2$	$Z = \frac{(\hat{p}_1 - \hat{p}_2) - D}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}$	The null hypothesis is $H_0: (p_1 - p_2) = D$; where $D \neq 0$; $n_1\hat{p}_1$, $n_1\hat{q}_1$, $n_2\hat{p}_2$ and $n_2\hat{q}_2 \geq 5$

SUPPLEMENTARY EXERCISES

13.73 In a study that was highly publicised, doctors discovered that aspirin seems to help prevent heart attacks. The research project, which was scheduled to last for five years, involved 22 000 patients (all male). Half took an aspirin tablet three times per week, while the other half took a placebo on the same schedule. After three years, researchers determined that 104 of those who took aspirin and 189 of those who took the placebo had had heart attacks.

- a Calculate the p -value of the test to determine whether these results indicate that aspirin is effective in reducing the incidence of heart attacks.

- b At the 0.005 level of significance, do these results indicate that aspirin is effective in reducing the incidence of heart attacks?
- c Another study attempted to replicate the same research plan; however, it used only 5000 men (2500 took aspirin and 2500 took the placebo). Suppose that the respective proportions of men who suffered heart attacks were exactly the same as in the study in part (a). Do such results allow us to draw the same conclusion, with $\alpha = 0.005$? If the appropriate conclusions in the two studies are not the same, explain why not.

- 13.74** A restaurant located in an office building decides to adopt a new strategy for attracting customers. Every week it advertises in the city newspaper. In the 10 weeks immediately prior to the advertising campaign, the average weekly gross was \$10500, with a standard deviation of \$750. In the eight weeks after the campaign began, the average weekly gross was \$12000, with a standard deviation of \$1000.
- Assuming that the weekly grosses are normally distributed, can we conclude, with $\alpha = 0.10$, that the advertising campaign was successful?
 - Assume that the net profit is 20% of the gross. If the ads cost \$100 per week, can we conclude at the 10% significance level that the ads are worthwhile?

- 13.75 XR13-75** The owners of two city restaurants (whose customers are mostly office workers on coffee breaks) each claim to serve more coffee than the other. To test their claims, they counted the number of cups of coffee sold during five randomly selected days in Restaurant 1 and another five randomly selected days in Restaurant 2. The resulting data are represented in the following table and recorded. (After some analysis, it is determined that the number of cups of coffee is normally distributed.) Can we conclude that there is a difference between the average coffee sales of the two restaurants? (Use $\alpha = 0.10$.)

Number of cups of coffee sold	
Restaurant 1	Restaurant 2
670	410
420	500
515	440
690	640
825	650

- 13.76 XR13-76** In an effort to reduce absenteeism, an electronics company initiated a program under which employees could participate in a monthly lottery, provided that they had perfect attendance and punctuality during the month. A \$100 cash prize was awarded to the winner of each monthly lottery. Approximately 80 employees participated in the program. In order to assess the impact of the program, comparisons were made between sick leave expenditures for eight randomly selected

months before and six randomly selected months after the lottery was instituted. The expenditures are recorded and are shown in the table below. Can we conclude at the 5% significance level that the mean monthly sick leave expenditure is lower after the institution of the program?

Prior monthly costs (\$)	Post monthly costs (\$)
903	746
812	775
1012	596
855	767
826	469
814	670
755	
690	

- 13.77 XR13-77** Although the use of seat belts is known to save lives and reduce serious injuries, some people still don't wear them when they travel in cars. In an effort to increase the use of seat belts, a government agency sponsored a two-year study. Among its objectives was to determine if there was enough evidence to justify the following conclusions:

- Seat-belt usage increased between last year and this year.
- This year, seat belts were used more frequently by women than by men.

To test these beliefs, a random sample of female and male drivers were sampled in the two years and asked whether they always used their seat belts. The responses were recorded in the following way.

Column 1: last year's survey responses, female respondents: 1 = wear seat belt; 0 = do not wear seat belt

Column 2: last year's survey responses, male respondents: 1 = wear seat belt; 0 = do not wear seat belt

Column 3: this year's survey responses, female respondents: 1 = wear seat belt; 0 = do not wear seat belt

Column 4: this year's survey responses, male respondents: 1 = wear seat belt; 0 = do not wear seat belt

What conclusions can be drawn from the results? (Use a 5% significance level.)

Sample statistics:

Column 1: $n_1^{(0)} = 58, n_1^{(1)} = 146$;

Column 2: $n_2^{(0)} = 104, n_2^{(1)} = 163$

Column 3: $n_3^{(0)} = 38, n_3^{(1)} = 150$;

Column 4: $n_4^{(0)} = 72, n_4^{(1)} = 166$

13.78 XR13-78 An important component of the cost of living is the amount of money spent on housing. Housing costs include rent (for tenants), mortgage payments and property tax (for home owners), heating, electricity and water usage. An economist undertook a five-year study to determine how housing costs have changed. Five years ago, he took a random sample of 200 households and recorded the percentage of total income spent on housing. This year, he took another sample of 200 households. The data are stored in columns 1 (five years ago) and 2 (this year).

- a Conduct a test (with $\alpha = 0.10$) to determine whether the economist can infer that housing cost as a percentage of total income has increased over the last five years.
- b Use whatever statistical method you deem appropriate to check the required condition(s) of the test used in part (a).

Sample statistics: $n_1 = 200, \bar{X}_1 = 32.42, s_1 = 6.08$;
 $n_2 = 200, \bar{X}_2 = 33.72, s_2 = 6.75$.

13.79 XR13-79 In designing advertising campaigns to sell magazines, it is important to know how much time each of a number of demographic groups spends reading magazines. In a preliminary study, 40 people were randomly selected. Each was asked how much time per week he or she spent reading magazines. Additionally, each was categorised by gender and by income level (high or low). The data are stored in the following way: column 1 = time spent reading magazines per week in minutes for all respondents; column 2 = gender (1 = male and 2 = female); column 3 = income level (1 = low and 2 = high).

- a Is there sufficient evidence at the 5% significance level to conclude that men and women differ in the amount of time spent reading magazines?

- b Is there sufficient evidence at the 5% significance level to conclude that high-income individuals devote more time to reading magazines than low-income people?
- c Does it appear that the required conditions for the above tests are satisfied?

Sample statistics:

Male: $\bar{X} = 39.75, s = 28.35$;

Female: $\bar{X} = 49.00, s = 27.08$

Low income: $\bar{X} = 33.10, s = 16.69$;

High income: $\bar{X} = 56.84, s = 32.37$

13.80 XR13-80 Before deciding which of two types of stamping machine should be purchased, the plant manager of an automotive parts manufacturer wants to determine the number of units that each produces. The two machines differ in cost, reliability and productivity. The firm's accountant has calculated that machine A must produce 25 more non-defective units per hour than machine B to warrant buying machine A. To help decide, both machines were operated for 24 hours. The total number of units and the number of defective units produced by each machine per hour were recorded. These data are recorded in the following way: column 1 = total number of units produced by machine A; column 2 = number of defective units produced by machine A; column 3 = total number of units produced by machine B; column 4 = number of defective units produced by machine B. Determine which machine should be purchased. (Use a 5% significance level.)

Sample statistics (non-defective):

Machine A: $n_A = 24, \bar{X}_A = 230.125, s_A^2 = 79.51$;

Machine B: $n_B = 24, \bar{X}_B = 200.917, s_B^2 = 59.04$.

The following exercises require the use of a computer and software.

13.81 XR13-81 How important to your health are regular holidays? In a study, a random sample of men and women were asked how frequently they take holidays. The men and women were divided into two groups each. The members of group 1 had suffered a heart attack; the members of group 2 had not. The number of days of holiday last year was recorded for each person. Can we infer that

men and women who suffer heart attacks take fewer holidays than those who did not suffer heart attacks?

13.82 XR13-82 Most people exercise in order to lose weight. To determine better ways to lose weight, a random sample of male and female exercisers was divided into groups. The first group exercised vigorously twice a week. The second group exercised moderately four times per week. The weight loss for each individual was recorded. Can we infer that people who exercise moderately more frequently lose weight faster?

13.83 XR13-83 After observing the results of the test in Exercise 13.82, a statistics practitioner organised another experiment. People were matched according to gender, height and weight. One member of each matched pair then exercised vigorously twice a week and the other member exercised moderately four times per week. The weight losses are recorded. Can we infer that people who exercise moderately more frequently lose weight faster?

13.84 XR13-84 Many small retailers advertise in their neighbourhoods by sending out flyers. People deliver these to homes and are paid according to the number of flyers delivered. Each deliverer is given several streets whose homes become their responsibility. One of the ways retailers use to check the performance of deliverers is to randomly sample some of the homes and ask the homeowner whether he or she received the flyer. Recently, a group of university students started a new delivery service. They have promised better service at a competitive price. A retailer wanted to know whether the new company's delivery rate is better than that of the existing firm. She had both companies deliver her flyers. Random samples of homes were drawn and each was asked whether he or she received the flyer (2 = yes and 1 = no) and the responses were recorded. Can the retailer conclude that the new company is better? (Test with $\alpha = 0.10$.)

13.85 XR13-85 Is marriage good for your health? To answer this question, researchers at a WA university monitored 103 couples, each with one spouse who was slightly hypertensive (mild high blood pressure). Participants also completed a questionnaire about their marriages. Three years later, the blood pressure of the previously hypertensive spouse was measured. The reduction in blood pressure for those in happy marriages was stored in column 1. The reduction for those in unhappy marriages was stored in column 2. Can we infer that spouses in happy marriages have a greater reduction in blood pressure than those in unhappy marriages?

13.86 XR13-86 In designing advertising campaigns to sell magazines, it is important to know how much time each of a number of demographic groups spends reading magazines. In a preliminary study, 20 people were randomly chosen. Each was asked how much time per week he or she spent reading magazines; additionally, each was categorised by gender and by income level (high or low). The resulting data are listed in the table.

Respondent	Time spent reading magazines (min.)	Gender	Income
1	80	M	L
2	125	M	H
3	150	F	H
.	.	.	.
.	.	.	.
18	80	F	L
19	130	F	H
20	150	M	H

- a Is there sufficient evidence at the 5% significance level to allow us to conclude that men and women differ in their magazine-reading habits?
- b Is there sufficient evidence at the 5% significance level to allow us to conclude that those whose incomes are higher devote more time to reading magazines than do lower-income individuals?

Case Studies

CASE 13.1 Is there gender difference in spirits consumption?

According to Roy Morgan's *Alcohol Consumption Currency Report* published in December 2019, 26.7% of Australians (18 years and over) consumed spirits in an average four-week period. Further, the report states that, as of September 2019, 30.8% of men consumed spirits in an average four-week period, compared with 22.9% of women. More than one-third of those aged 18–24 consumed spirits in an average four-week period (36.7%), as did just over a quarter of those aged 25–49. The least likely group to drink spirits is those aged 50+ (23.0%). Roy Morgan CEO Michele Levine says that 'Looking at the types of alcohol Australians are drinking, spirits is one of the stronger categories over the last few years with consumption increasing from 25.9% in 2014 to 26.7% as of September 2019.' Assuming that the number of people surveyed in both years is 1750, investigate whether the percentage of Australian adults drinking spirits has significantly increased between 2014 and 2019 (use $\alpha = 0.05$)?

If the 2019 survey involved 900 men and 850 women, can we conclude at the 5% significance level that the percentage of men drinking spirits is at least 4% higher than that of women.

Source: <http://www.roymorgan.com/findings/8194-alcohol-consumption-currency-report-september-2019-201912012220>

CASE 13.2 Consumer confidence in New Zealand

C13-02 Consumer confidence is one of the major indicators that reflect the state of the economy. Many businesses, governments and policy analysts use consumer confidence ratings to revise various economic and business policies. The following are the Roy Morgan confidence ratings by New Zealanders for the five component questions during November 2015 and November 2019. Analyse the data.

Results for the weekly Roy Morgan consumer confidence rating

	2015	2019
Interviews (sample size)	999	998
Overall consumer confidence rating	122.7	120.7
Q1 Would you say you and your family are better-off financially or worse off than you were at this time last year?		
Better off	34	37
Worse off	25	20
Q2 This time next year, do you and your family expect to be better-off financially or worse off than you are now?		
Better off	47	41
Worse off	16	15
Q3 Thinking of economic conditions in New Zealand as a whole, in the next 12 months, do you expect we'll have good times financially or bad times?		
Good times	44	32
Bad times	29	28
Q4 Looking ahead, what would you say is more likely, that in New Zealand as a whole, we'll have continuous good times during the next five years or so, or we'll have bad times?		
Good times	45	34
Bad times	24	19
Q5 Generally, do you think now is a good time or a bad time for people to buy major household items?		
Good times	58	58
Bad times	20	17

Source: <http://www.roymorgan.com/morganpoll/consumer-confidence/anz-roymorgan-new-zealand-consumer-confidence>

CASE 13.3 New Zealand Government bond yields: Short term versus long term

C13-03 An investment manager of a company is considering his investment plans in relation to New Zealand Government bond yields. His colleague suggests that, in general, a medium-term government bond (say, a 5-year bond) would give a lower yield than a long-term bond such as a 10-year bond. The investment manager was not convinced by his colleague's assessment and wanted to investigate the matter himself. He collected sample monthly data on the New Zealand Government bond yields over the period from 2010 to 2020 from the Reserve Bank of New Zealand website, but is not sure what statistical test he needs to perform to test his colleague's claim. Can you help the investment manager?

CASE 13.4 The price of petrol in Australia: Is it similar across regions?

C13-04 The ACCC monitors fuel prices in all capital cities and more than 190 regional locations across Australia. Monthly average retail petrol prices for June 2019 were recorded for the six capital cities and for the 182 regional cities in the six states of Australia. Applying appropriate statistical techniques to the data, for each state, first investigate whether there is difference in the average price of petrol in the regional towns and their corresponding capital city. Present a comparison of average petrol prices between regional cities in (i) Queensland and South Australia, (ii) NSW and Tasmania and (iii) Victoria and WA.

Source: © Commonwealth of Australia 2019. CC BY 3.0AU <https://creativecommons.org/licenses/by-sa/3.0/au/>

CASE 13.5 Student surrogates in market research

Researchers in both the business world and the academic world often treat university students as representative of the adult population. This practice reduces sampling costs enormously, but its effectiveness is open to question. An experiment was performed to determine the suitability of using student surrogates in research. The study used three groups of people:

- 1 The first consisted of 59 adults (18 years of age or older) chosen so that they represented the adult population by age and occupation.
- 2 The second consisted of 42 students enrolled in an introductory marketing subject at a university. Many of the students were registered in a business program, but few were marketing majors.
- 3 The third consisted of 33 students enrolled in an advanced marketing subject, almost all of whom were marketing majors.

The experiment consisted of showing each group a sequence of three 30-second television advertisements dealing with financial institutions. Each respondent was asked to assess each commercial on the basis of believability and interest. The responses were recorded on a 10-point graphic rating scale in which a higher rating represented greater believability or interest. The sample means and standard deviations were then calculated, and these are shown in the tables below.

What conclusions can you draw regarding the suitability of using students as surrogates in marketing research?

Comparison of responses of introductory marketing students, advanced marketing students and adults: Believability of advertisement

Advertisement	Introductory marketing students		Advanced marketing students		Adults	
	\bar{X}	s	\bar{X}	s	\bar{X}	s
1	6.7	2.5	6.6	3.1	6.9	2.7
2	7.3	2.6	7.2	2.3	6.1	3.0
3	5.9	2.7	6.6	2.8	7.0	2.9

Comparison of responses of introductory marketing students, advanced marketing students and adults: Interest in advertisement

Advertisement	Introductory marketing students		Advanced marketing students		Adults	
	\bar{X}	s	\bar{X}	s	\bar{X}	s
1	4.5	3.2	4.3	2.8	5.9	2.4
2	6.0	2.7	6.1	2.5	4.5	2.9
3	4.0	2.6	4.3	2.8	5.8	3.1

CASE 13.6 Do expensive drugs save more lives?

Two prescription medications are commonly used to treat a heart attack. Streptokinase, the less expensive drug, has been available since 1959. The second, more expensive, drug is tPA, a genetically engineered product. Both streptokinase and tPA work by opening the arteries and dissolving blood clots, which are the cause of heart attacks. Several previous studies have failed to reveal any differences between the effects of the two drugs. Consequently, in many countries where health care is funded by governments, doctors are required to use the less expensive streptokinase. However, the maker of tPA, Genentech Inc., contended that in the earlier studies showing no difference between the two drugs, tPA was not used in the right way. Genentech decided to sponsor a more thorough experiment. The experiment was organised in 15 countries, including Australia, the United States and Canada, and involved a total of 41 000 patients. In this study, tPA was given to patients within 90 minutes, instead of three hours as in previous trials. Half of the sample of 41 000 patients was treated by a rapid injection of tPA with intravenous heparin, while the other half received streptokinase along with heparin. The number of deaths in each sample was recorded. A total of 1497 patients treated with streptokinase died, while 1292 patients who received tPA died. Is there enough evidence to support the findings of previous studies that there is no difference between the effects of the two drugs?

CASE 13.7 Comparing two designs of ergonomic desk: Part I

C13-07 The plant manager of a company that manufactures office equipment believes that worker productivity is a function of, among other things, the design of the job, which refers to the sequence of worker movements. Two designs are being considered for the production of a new type of ergonomic computer desk. To help decide which design should be used, an experiment was performed. Twenty-five randomly selected workers assembled the desk using design A, and 25 workers assembled the product using design B. The assembly times in minutes were recorded. The plant manager would like to know whether the assembly times of the two designs differ. A 5% significance level is judged to be appropriate. Assume that the two populations of assembly times are normally distributed.

Appendix 13.A

Excel instructions: Manipulating data

13.Aa Steps to unstack the data

In order to use Excel to test $\mu_1 - \mu_2$, the data must be unstacked. Suppose that the data are stacked in the following way: column A stores the observations, and column B stores the indexes.

If the data are scrambled (not in order), proceed to step 1 below. Otherwise, go to step 4.

- 1 Highlight columns A and B.
- 2 Click **Sort & Filter**. Select **Custom Sort...**
- 3 Specify **Column B** and **Smallest to largest**. Click **OK**.

The data will now be unscrambled – all the observations from the first sample will occupy the top rows of column A, and the observations from the second sample will occupy the bottom rows of column A. To unstack, issue the following commands. (The following commands assume that there are only two samples.)

- 4 Highlight the rows of column A that were taken from sample 2.
- 5 Click **Cut** on the Clipboard submenu.
- 6 Make cell C1 active.
- 7 Click **Paste** on the Clipboard submenu.
- 8 Delete column B.

Columns A and B will now store the unstacked data.

13.Ab Steps to stack the data

There are several statistical procedures to be conducted by Excel in later chapters that require the data to be stacked. If the data are presented to you in unstacked form, you will have to stack them. Suppose that there are two sets of observations now stored in columns A and B. To stack the observations of A on B, proceed as follows.

- 1 Highlight the cells in column B.
- 2 Click **Cut** on the Clipboard submenu.
- 3 Make the first empty cell in column A active. Click **Paste** on the Clipboard submenu.
- 4 Type the codes in column B.

Columns A and B will now contain the stacked data – all the observations in column A and the codes identifying the sample in column B.

Chi-squared tests

Learning objectives

This chapter introduces various popular statistical procedures that use the chi-squared distribution to conduct tests related to nominal data.

At the completion of this chapter, you should be able to:

- L01** conduct the chi-squared test of goodness-of-fit for a multinomial experiment
- L02** conduct the chi-squared test of independence on data arranged in a contingency table to determine whether two classifications of nominal data are statistically independent
- L03** perform the chi-squared test for normality
- L04** perform the chi-squared test for a Poisson distribution.

CHAPTER OUTLINE

- Introduction
- 14.1** Chi-squared goodness-of-fit test
- 14.2** Chi-squared test of a contingency table
- 14.3** Chi-squared test for normality
- 14.4** Summary of tests on nominal data

SPOTLIGHT ON STATISTICS

Has support for the death penalty for drug trafficking changed since 2005?

In April 2015, Australian citizens Andrew Chan and Myuran Sukumaran were executed in Indonesia following convictions for drug trafficking offences, almost ten years after the last Australian, Van Tuong Nguyen, was executed in Singapore for similar offences. The issue of the death penalty for drug trafficking around the world has been argued for many years. A few countries have abolished it, and others have kept the laws on their books but rarely use them. Where does the public stand on the issue, and has public support been constant or has it changed from year to year? One of the questions asked in the survey was 'If an Australian is convicted of trafficking drugs in another country and sentenced to death, in your opinion, should the penalty be carried out or not?' The responses are 'Support', 'Oppose' and 'Can't say'.

Based on the responses in 2005, 2010 and 2015, the summary data file **Ch14 XM14-00** was created by counting the total number of responses under each category for 2005, 2010 and 2015. Conduct a test to determine whether public support for the death penalty has varied from year to year. On pages 604–5 we answer this question.



Source: iStock.com/erhui1979

Introduction

We have looked at a variety of statistical techniques that are used when the data are nominal. In Chapter 3, we introduced bar and pie charts, both graphical techniques to describe a set of nominal data. Later in Chapter 3, we also showed how to describe the relationship between two sets of nominal data by producing a frequency table and a bar chart. However, these techniques simply describe the data, which may represent a sample or a population. In this chapter, we deal with similar problems, but the goal is to use statistical techniques to make inferences about populations from sample data. This chapter develops two statistical techniques that involve nominal data. The first is a *goodness-of-fit test* applied to data produced by a multinomial experiment, a generalisation of a binomial experiment. The second uses data arranged in a table (called a **contingency table**) to determine whether two classifications of a population of nominal data are statistically independent; this test can also be interpreted as a comparison of two or more populations. The sampling distribution of the test statistics in both tests is the chi-squared distribution introduced in Appendix 14.A

Following are two examples of situations in which chi-squared tests could be applied.

Example 1 Firms periodically estimate the proportion (or market share) of consumers who prefer their products, as well as the market shares of competitors. These market shares may change over time as a result of advertising campaigns or the introduction of new, improved products. To determine whether the actual current market shares are in accord with its beliefs, a firm might sample several consumers and, for each of k competing companies, calculate the proportion of consumers sampled who prefer that company's products. An experiment in which each consumer is classified as preferring one of the k companies is called a *multinomial experiment*. If only two companies were considered ($k = 2$), we would be dealing with the familiar binomial experiment. After calculating the proportion of consumers preferring each of the k companies, a goodness-of-fit test could be conducted to determine whether the sample proportions (or market shares) differ significantly from those hypothesised by the firm. The problem objective is to describe the population of consumers, and the data are nominal.

Example 2 For advertising and other purposes, it is important for a company to understand which segments of the market prefer which of its products. For example, it would be helpful for an automotive manufacturer to know if there is a relationship between the buyer preferences for its various models and the gender of the consumer. After conducting a survey to solicit consumers' preferences, the firm could classify each respondent according to two nominal variables: model preferred and gender. A test could then be conducted to determine whether consumers' preferences are independent of their gender. Rather than interpreting this test as a test of the independence of two nominal variables defined over a single population, we could view male and female consumers as representing two different populations. Then we could interpret the test as testing for differences in preferences between these two populations.

contingency table
(or cross-classification table) A table with the expected values calculated contingent on the assumption that the null hypothesis is true.

14.1 Chi-squared goodness-of-fit test

14.1a Multinomial experiment

This section presents another test designed to describe a single population of nominal data. The first such test was introduced in Section 12.6, in which we discussed the statistical procedure employed to test hypotheses about a population proportion. In that case, the nominal variable could assume one of only two possible values: *success* or *failure*. Our tests dealt with hypotheses about the proportion of successes in the entire population. Recall that

multinomial experiment

An extension of the binomial experiment, in which there are two or more possible outcomes per trial.

the experiment that produces the data is called a binomial experiment. In this section, we introduce the **multinomial experiment**, an extension of the binomial experiment, in which there are two *or more* possible outcomes per trial. For example, there are seven methods of payment by which a taxpayer can make payments to the Australian Taxation Office: cash, money order, BPAY, credit or debit card, electronic transfer, bank draft and direct-debit.

Multinomial experiment

A multinomial experiment is one that has the following properties:

- 1** The experiment consists of a fixed number (n) of trials.
- 2** The outcome of each trial can be classified into one of k categories, called cells.
- 3** The probability p_i that the outcome will fall into cell i remains constant for each trial. Moreover, $p_1 + p_2 + \dots + p_k = 1$.
- 4** Each trial of the experiment is independent of the other trials.

14.1b Observed frequencies

observed frequencies

The number of observations of each outcome in the experiment.

As you can see, when $k = 2$, the multinomial experiment is identical to the binomial experiment. Just as we count the number of successes (recall that we label the number of successes X) and failures in a binomial experiment, we count the number of outcomes falling into each of the k cells in a multinomial experiment. In this way, we obtain a set of **observed frequencies** o_1, o_2, \dots, o_k , where o_i is the observed frequency of outcomes falling into cell i , for $i = 1, 2, \dots, k$. Because the experiment consists of n trials and an outcome must fall into some cell,

$$o_1 + o_2 + \dots + o_k = n$$

Just as we used the number of successes \hat{X} (by calculating the sample proportion \hat{p} , which is equal to \hat{X}/n) to draw inferences about p , so do we use the observed frequencies to draw inferences about the cell probabilities. We will proceed in what by now has become a standard procedure.

14.1c Testing hypotheses

We will set up the hypotheses and develop the test statistic and its sampling distribution.

$$H_0: p_i = \hat{p}_i \text{ for all } i = 1, 2, \dots, k$$

$$H_A: \text{at least one } p_i \text{ is not equal to its specified value.}$$

where \hat{p}_i is the specified value in cell i , such that

$$\hat{p}_1 + \hat{p}_2 + \dots + \hat{p}_k = 1$$

14.1d Expected frequencies

expected frequency

The frequency of each outcome we expect to observe if the null hypothesis is true.

In general, the **expected frequency** (e_i) for each postposition is given by

$$e_i = np_i$$

This expression is derived from the formula for the expected value of a binomial random variable first seen in Section 7.6.

$$E(X) = np$$

14.1e Test statistic and decision rule

If the expected frequencies e_i and the observed frequencies f_i are quite different, we would conclude that the null hypothesis is false, and we would reject it. However, if the expected and observed frequencies are similar, we would not reject the null hypothesis. The test statistic we employ to assess whether the differences between the expected and observed frequencies are large has a chi-square (χ^2) distribution.

Chi-squared goodness-of-fit test statistic

$$\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

The sampling distribution of the test statistic is approximately chi-squared distributed with $k - 1$ degrees of freedom, provided that the sample size is large. We will discuss this condition later. (See Appendix 14.A for an introduction to chi-squared distribution, and the critical values for a chi-squared distribution are given in Table 5 of Appendix B.)

When the null hypothesis is true, the observed and expected frequencies should be similar, in which case the test statistic will be small. Thus, a small test statistic supports the null hypothesis. If the null hypothesis is untrue, some of the observed and expected frequencies will differ and the test statistic will be large. Consequently, we want to reject the null hypothesis in favour of the alternative hypothesis when the observed value of χ^2 is greater than $\chi^2_{\alpha, k-1}$.

Decision rule: Reject H_0 if $\chi^2 > \chi^2_{\alpha, k-1}$.

We will demonstrate the process with the following example.

EXAMPLE 14.1

LO1

Does postposition matter in winning a horse race?

XM14-01 A statistician occasionally engages in practical demonstrations of probability theory (i.e. he bets on horse races). His current method of assessing probabilities and applying wagering strategy has not been particularly successful. (He usually loses.) To help improve his cash flow, he decides to learn more about the outcomes of races. In particular, he would like to know whether some postpositions are more favourable than others, so that horses starting in these positions win more frequently than horses starting in other positions. If this is true, it is likely that the return on his investment will be positive often enough to overcome the more frequently occurring negative return. (He may be able to win money.) He records the postpositions of the winners of 150 randomly selected races. These data are summarised in the accompanying table. Using a significance level of 5%, can the statistician conclude that some postpositions win more frequently than others?

Results of 150 races

Postposition	Observed frequency f_i
1	22
2	33
3	21
4	30
5	23
6	21
Total	150



Solution

Identifying the technique

The population in question is the population of winning postpositions for all races. Although the data may appear at first glance to be numerical, they are in fact nominal. The numbers from one to six simply give unique names to each of the postpositions. Thus, calculating the mean winning postposition, for example, would be meaningless. The experiment described in this example matches the properties of a multinomial experiment. It follows that the parameters of interest are the probabilities (or proportions) p_1, p_2, \dots, p_6 that postpositions 1 to 6 are the winning positions. To determine if some postpositions win more frequently than others, we specify under the null hypothesis that all postpositions are equally likely to win. That means that the probabilities p_1, p_2, \dots, p_6 are equal, and so they are all equal to 1/6 (since their total must equal 1). Thus,

$$H_0: p_1 = 1/6, p_2 = 1/6, p_3 = 1/6, p_4 = 1/6, p_5 = 1/6, p_6 = 1/6$$

Because the point of the experiment is to determine whether at least one postposition wins more frequently than the others, we specify the alternative hypothesis as

$$H_A: \text{At least one } p_i \text{ is not equal to its specified value.}$$

Test statistic:

Calculating manually

If the null hypothesis is true, we would expect each postposition to win one-sixth of the races, and as $n = 150$, we expect that each postposition would win $150(1/6) = 25$ times.

The test statistic to assess whether differences between the expected and observed frequencies are large is the χ^2 -statistic

$$\chi^2 = \sum_{i=1}^6 \frac{(o_i - e_i)^2}{e_i}$$

where o_i is the observed frequency and $e_i = np_i$ is the expected frequency. As the sample size $n = 150$ is large, the test statistic χ^2 is approximately χ^2 distributed with $k - 1 = 5$ degrees of freedom.

The table below demonstrates the calculation of the χ^2 -statistic. Thus, the value of the test statistic is $\chi^2 = 5.36$. As usual, we need to judge the magnitude of the test statistic to determine our conclusion. This is the function of the rejection region.

Postposition i	Observed frequency f_i	Expected frequency e_i	$(o_i - e_i)$	$\frac{(o_i - e_i)^2}{e_i}$
1	22	25	-3	0.36
2	33	25	8	2.56
3	21	25	-4	0.64
4	30	25	5	1.00
5	23	25	-2	0.16
6	21	25	-4	0.64
Total	150	150		$\chi^2 = 5.36$

Decision rule:

Reject H_0 if $\chi^2 > \chi^2_{\alpha, k-1} = \chi^2_{0.05, 5} = 11.0705$.

Alternatively, reject H_0 if the p -value $< \alpha = 0.05$.

Value of the test statistic: We have already calculated the test statistic and found it to be $\chi^2 = 5.36$.

The p -value of the test is, $p\text{-value} = P(\chi^2 > 5.36)$. Unfortunately, Table 5 in Appendix B does not allow us to perform this calculation (except for approximation by interpolation). The p -value must be produced by computer. From the output below, $p\text{-value} = 0.3735$.



Conclusion:

As $\chi^2 = 5.36 < 11.0705$, do not reject H_0 .

Alternatively, as $p\text{-value} = 0.3735 > 0.05$, do not reject H_0 .

The complete test follows:

Hypotheses:

$H_0: p_i = 1/6, i = 1, 2, \dots, 6$

$H_A: \text{At least one } p_i \neq 1/6$

Test statistic:

$$\chi^2 = \sum_{i=1}^6 \frac{(o_i - e_i)^2}{e_i} \sim \chi^2_{5} \quad (k-1=5)$$

where o_i is the observed frequencies; and $e_i = np_i = 150(1/6) = 25$ the expected frequencies.

Level of significance:

$$\alpha = 0.05$$

Decision rule:

Reject H_0 if $\chi^2 > \chi^2_{\alpha, k-1} = \chi^2_{0.05, 5} = 11.07$.

Alternatively, reject H_0 if $p\text{-value} < \alpha = 0.05$.

Value of the test statistic:

$$\begin{aligned} \chi^2 &= \frac{(22-25)^2}{25} + \frac{(33-25)^2}{25} + \dots + \frac{(21-25)^2}{25} \\ &= 0.36 + 2.56 + \dots + 0.64 \\ &= 5.36 \end{aligned}$$

Conclusion:

As $\chi^2 = 5.36 < 11.07$, do not reject H_0 , or, alternatively (using the computer), since $p\text{-value} = P(\chi^2 > 5.36) = 0.3735 > \alpha = 0.05$, do not reject H_0 .

There is not enough evidence to infer that some postpositions win more frequently than others.

Interpreting the results

We conclude that there is not enough statistical evidence to infer that some postpositions win more frequently than others. What this means is that the statistics practitioner cannot use the postposition to help determine a winner. He would be advised to seek some other way of deciding on which horse to bet; or, yet better advice would be to quit betting on the horses.

Using the computer

Using Excel function

The output for Example 14.1 from the commands listed below is the p -value of the test, which is 0.3735.

Excel output for Example 14.1

	A	B
11	p-value=	0.373541673

COMMANDS

- Type the observed and expected values in two adjacent columns (**observed values 22, 33, 21, 30, 23, 21 in cells B2:B7; expected values 25, 25, 25, 25, 25, 25 in cells C2:C7**) or open the data file (**XM14-01**). (If you wish, you can type the cell probabilities specified in the null hypothesis and let Excel convert these into expected values by multiplying by the sample size.)
- To calculate the p -value of the chi-squared test, activate an empty cell and type

=CHITEST([Actual_range], [Expected_range])

where the ranges are the cells containing the actual observations and the expected values. Do not include the cells containing the names of the columns. For example, type **=CHITEST(B2:B7, C2:C7)**, which will calculate the p -value, 0.3735.

If we have the raw data representing the nominal responses, we must first determine the frequency of each category (the observed values) using the COUNTIF function described in Example 3.1 on page 49. To do so proceed as follows:

- Activate any empty cell.
- Click **FORMULAS, fx, All** from the categories dropdown menu and **COUNTIF**.
- Specify the **Range** of the data (do not include the cell containing the name of the variable) and the category you wish to count (**Criteria**). Click **OK**.

Using XLSTAT

	B	C
9	Chi-square test:	
10	Chi-square (Observed value)	5.360
11	Chi-square (Critical value)	11.070
12	DF	5
13	p-value	0.374
14	alpha	0.05

COMMANDS

- Type the observed and expected frequencies in two adjacent columns (**B1:B7, C1:C7**), including column labels or open the data file (**XM14-01**).
- Click **XLSTAT**, **Parametric tests**, and **Multinomial goodness of fit test**.
- Specify the range of the (observed) **Frequencies (B1:B7)** and the range of the **Expected frequencies (C1:C7)**. Specify **Data format: Frequencies** and check the **Chi-square test**. Click **OK**.

Although it is common to test for no differences in the cell probabilities p_1, p_2, \dots, p_k (as we did in Example 14.1), we are not restricted to this formulation. We may hypothesise a different value for each p_i , as long as the sum of the probabilities is 1. Moreover, the test described in this section is also called a goodness-of-fit test because it can be used to test how well the data fit a hypothesised distribution. In this application, we would use the hypothesised distribution to calculate a set of probabilities and employ the chi-squared test of a multinomial experiment to test the belief.

EXAMPLE 14.2

LO1

Testing market shares

XM14-02 Two Australian soft drink companies A and B have recently conducted aggressive advertising campaigns in order to maintain and possibly increase their respective shares of the soft drink market. These two companies enjoy a dominant position in the market. Before the advertising campaigns began, the market share of company A was 40%, while company B had 35%. Other competitors accounted for the remaining 25%. To determine whether these market shares changed after the advertising campaigns, a marketing analyst solicited the preferences of a random sample of 400 soft drink consumers. Of the 400 consumers, 204 indicated a preference for company A's drink, 144 preferred company B's drink, and the remaining 32 preferred another competitor's soft drink. Can the analyst infer at the 5% significance level that soft drink consumers' preferences have changed from the levels they were at before the launch of the advertising campaigns?

Solution

Identifying the technique

The objective of the problem is to describe the population of soft drink consumers. The data are nominal because each respondent will choose one of three possible answers – company A, company B or other. In this problem we are interested in the proportions of three categories. We recognise this experiment as a multinomial experiment, and we identify the technique as the chi-squared test of a multinomial experiment.

Because we want to know if the market shares have changed, we specify those pre-campaign market shares in the null hypothesis:

$$H_0: p_1 = 0.40, p_2 = 0.35, p_3 = 0.25$$

The alternative hypothesis attempts to answer our question: Have the proportions changed? Thus,

H_A : At least one p_i is not equal to its specified value.

The complete test follows:

Hypotheses: $H_0: p_1 = 0.40, p_2 = 0.35, p_3 = 0.25$

H_A : At least one p_i is not equal to its specified value.

Test statistic:

$$\chi^2 = \sum_{i=1}^3 \frac{(o_i - e_i)^2}{e_i} \sim \chi^2_{k-1} \quad (k-1=2)$$

Level of significance: $\alpha = 0.05$

Decision rule: Reject H_0 if $\chi^2 > \chi^2_{\alpha,k-1} = \chi^2_{0.05,2} = 5.99147$.

Alternatively, reject H_0 if $p\text{-value} < \alpha = 0.05$.

Calculating manually

Value of the test statistic: The expected values are as follows:

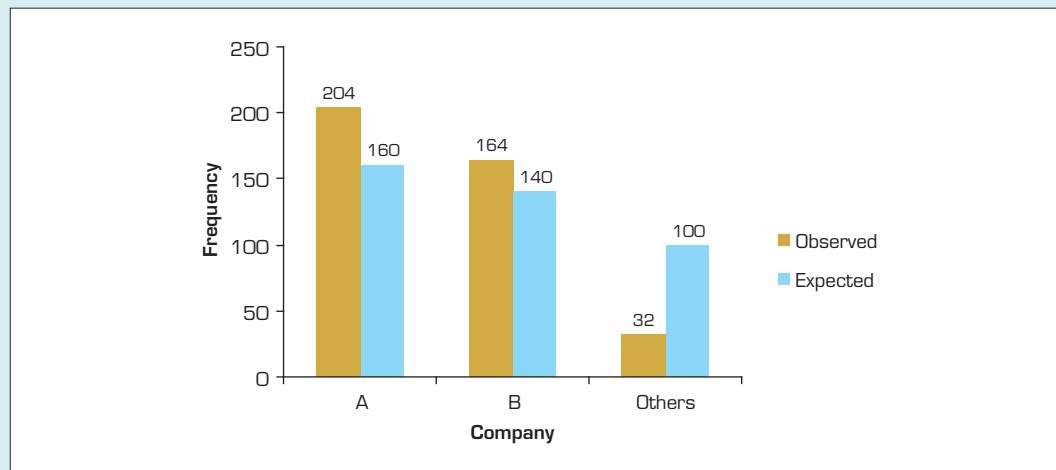
$$e_1 = np_1 = 400(0.40) = 160$$

$$e_2 = np_2 = 400(0.35) = 140$$

$$e_3 = np_3 = 400(0.25) = 100$$

Figure 14.1 is a bar chart that compares the observed (post-campaign) and expected frequencies (pre-campaign). If the pre-campaign and post-campaign frequencies are quite different, we would expect the null hypothesis to be rejected. As can be seen from **Figure 14.1** there are differences between the frequencies. We will now formally test this observation.

FIGURE 14.1 Bar chart of observed and expected market shares of soft drinks



The relevant frequencies and calculations are produced in the following table.

Company	Observed frequency (o_i)	Expected frequency (e_i)	$\frac{(o_i - e_i)^2}{e_i}$
A	204	160	12.10
B	164	140	4.11
Others	32	100	46.24
Total	400	400	$\chi^2 = 62.45$

$$\chi^2 = \sum_{i=1}^3 \frac{(o_i - e_i)^2}{e_i} = \frac{(204 - 160)^2}{160} + \frac{(164 - 140)^2}{140} + \frac{(32 - 100)^2}{100} \\ = 12.10 + 4.11 + 46.24 = 62.45$$

Conclusion: As $\chi^2 = 62.45 > 5.99$, reject the null hypothesis or, alternatively (using the computer), since $p\text{-value} = P(\chi^2 > 62.45) = 0.00 < \alpha = 0.05$, reject the null hypothesis.

Interpreting the results

There is sufficient evidence to infer that the proportions have changed since the advertising campaigns. If the sampling was conducted properly, we can be quite confident in our conclusion. This technique has only one required condition, which is satisfied. (See the rule of five described below.) It is probably a worthwhile exercise to determine the nature and causes of the changes. The results of this analysis will determine the design and timing of other advertising campaigns.

Using the computer

The commands are the same as in Example 14.1. The Excel output gives the p -value of the test, which is 0.000.

Excel output for Example 14.2

	A	B
11	p-value=	0.0000

14.1f Rule of five (I)

The test statistic used to compare the relative sizes of observed and expected frequencies is

$$\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

We previously stated that this test statistic has an approximate chi-squared distribution. In fact, the actual distribution of this test statistic is discrete, but it can be approximated conveniently by using a continuous chi-squared distribution when the sample size n is large, just as we approximated the discrete binomial distribution by using the normal distribution. This approximation may be poor, however, if the expected cell frequencies are small. For the (discrete) distribution of the test statistic to be adequately approximated by the (continuous) chi-squared distribution, the conventional (and conservative) rule – known as the **rule of five** – is to require that the expected frequency for each cell be at least 5.¹ Where necessary, cells should be combined in order to satisfy this condition. The choice of cells to be combined should be made in such a way that meaningful categories result from the combination.

Consider the following modification of Example 14.2. Suppose that three companies (A, B and C) have recently conducted aggressive advertising campaigns. The market shares prior to the campaigns were $p_1 = 0.45$ for company A, $p_2 = 0.40$ for company B, $p_3 = 0.13$ for company C, and $p_4 = 0.02$ for other competitors. In a test to see if market shares changed after the advertising campaigns, the null hypothesis would now be

$$H_0: p_1 = 0.45, p_2 = 0.40, p_3 = 0.13, p_4 = 0.02$$

Hence, if the preferences of a sample of 200 customers were solicited, the expected frequencies would be

$$e_1 = 90, e_2 = 80, e_3 = 26, e_4 = 4$$

¹ To be on the safe side, this rule of thumb is somewhat conservative. A discussion of alternatives to the rule of five can be found in Conover, W.J. 1971, *Practical nonparametric statistics*, John Wiley, New York, p. 152, and in Siegel, S. 1956, *Nonparametric statistics for the behavioral sciences*, McGraw-Hill, New York, p. 178.

As the expected cell frequency e_4 is less than five, the rule of five requires that it be combined with one of the other expected frequencies (say, e_3) to obtain a combined cell frequency of (in this case) 30. Although e_4 could have been combined with e_1 or e_2 , we have chosen to combine it with e_3 so that we still have a separate category representing each of the two dominant companies (A and B). After this combination is made, the null hypothesis reads

$$H_0: p_1 = 0.45, p_2 = 0.40, p_3 = 0.15$$

where p_3 now represents the market share of all competitors of companies A and B. Therefore, the appropriate number of degrees of freedom for the chi-squared test statistic would be $k - 1 = 3 - 1 = 2$, where k is the number of cells after some have been combined to satisfy the rule of five.

Let's summarise the factors that allow us to recognise when to use the statistical procedure described in this section.

IN SUMMARY

Factors that identify the chi-squared goodness-of-fit test

- 1** *Problem objective:* to describe a population
- 2** *Data type:* nominal (categorical)
- 3** *Number of categories:* two or more

EXERCISES

Learning the techniques

Exercises 14.1–14.6 are 'what-if' analyses designed to determine what happens to the test statistic of the goodness-of-fit test when elements of the statistical inference change. These problems can be solved manually or using Excel's CHITEST.

- 14.1 XR14-01** Consider a multinomial experiment involving $n = 300$ trials and $k = 5$ cells. The observed frequencies resulting from the experiment are shown in the accompanying table, and the hypotheses to be tested are as follows:
- $$H_0: p_1 = 0.1, p_2 = 0.2, p_3 = 0.3, p_4 = 0.2, p_5 = 0.2$$
- $$H_A: \text{At least one } p_i \text{ is not equal to its specified value.}$$

Test the hypothesis at the 1% significance level.

Cell	1	2	3	4	5
Frequency	24	64	84	72	56

- 14.2 XR14-02** Repeat Exercise 14.1, with $n = 150$ and the following frequencies:

Cell	1	2	3	4	5
Frequency	12	32	42	36	28

- 14.3 XR14-03** Repeat Exercise 14.1, with $n = 75$ and the following frequencies:

Cell	1	2	3	4	5
Frequency	6	16	21	18	14

- 14.4** Review the results of Exercises 14.1–14.3. What is the effect of decreasing the sample size?

- 14.5 XR14-05** Consider a multinomial experiment involving $n = 150$ trials and $k = 4$ cells. The observed frequencies resulting from the experiment are shown in the accompanying table, and the null hypothesis to be tested is as follows:

$$H_0: p_1 = 0.3, p_2 = 0.3, p_3 = 0.2, p_4 = 0.2$$

Cell	1	2	3	4
Frequency	38	50	38	24

- a** State the alternative hypothesis.
b Test the hypothesis, using $\alpha = 0.05$.

- 14.6 XR14-06** For Exercise 14.5, retest the hypotheses, assuming that the experiment involved twice as many trials ($n = 300$) and that the observed

frequencies were twice as high as before, as shown in the accompanying table. (Use $\alpha = 0.05$.)

Cell	1	2	3	4
Frequency	76	100	76	48

- 14.7** Review the results of Exercises 14.5 and 14.6. What is the effect of increasing the sample size?

Applying the techniques

Exercises 14.8–14.18 require the use of a computer and software. Use a 5% significance level unless specified otherwise. Some of the answers to Exercises 14.8–14.13 may be calculated manually using the sample statistics provided.

- 14.8 XR14-08** A multinomial experiment was conducted with $k = 4$. Each outcome is stored as an integer from 1 to 4 and the results of a survey were recorded. Test the following hypotheses.

$$H_0: p_1 = 0.15, p_2 = 0.40, p_3 = 0.35, p_4 = 0.10$$

$$H_A: \text{At least one } p_i \text{ is not equal to its specified value.}$$

Sample statistics: $n(1) = 41, n(2) = 107, n(3) = 66, n(4) = 19$

- 14.9 XR14-09 Self-correcting exercise.** To determine whether a single die is balanced or fair, it was rolled 600 times. Is there sufficient evidence to conclude, at the 5% level of significance, that the die is not fair?

Sample statistics: $n(1) = 114, n(2) = 92, n(3) = 84, n(4) = 101, n(5) = 107, n(6) = 102$

- 14.10 XR14-10** For Exercise 14.9, suppose that the die were rolled 1200 times and the observed frequencies were twice as high as before. Is there now sufficient evidence to conclude, at the 5% level of significance, that the die is not fair?

Sample statistics: $n(1) = 228, n(2) = 184, n(3) = 168, n(4) = 202, n(5) = 214, n(6) = 204$

- 14.11 XR14-11** Grades assigned by an economics lecturer for his postgraduate coursework students at a university in New South Wales have historically followed a symmetrical distribution: 5% HDs, 25% Ds, 40% Cs, 25% Ps and 5% Fs. This year, a sample of 150 grades was drawn and the grades (1 = A, 2 = B, 3 = C, 4 = D, and 5 = F) were recorded. Can you conclude, at the 1% level of

significance, that this year's grades are distributed differently from grades in the past?

Sample statistics: $n(1) = 11, n(2) = 32, n(3) = 62, n(4) = 29, n(5) = 16$

- 14.12** A firm in a major Australian city has been accused of engaging in prejudicial hiring practices. According to the most recent census, the percentages of Anglo-Saxons (1), Asians (2) and others (3) in the community where the firm is located are 70%, 12% and 18%, respectively. A random sample of 200 employees of the firm revealed that 165 were Anglo-Saxon, 14 were Asians and 21 were others. What would you conclude, at the 5% level of significance?

Sample statistics: $n(1) = 165, n(2) = 14, n(3) = 21$

- 14.13 XR14-13** Pat Statsdud is about to do a multiple-choice exam with 25 questions but, as usual, knows absolutely nothing. He plans to guess one of the five choices for each question. Pat has been given one of the lecturer's previous exams with the correct answers marked. The correct choices were recorded where 1 = (a), 2 = (b), 3 = (c), 4 = (d), and 5 = (e). Help Pat determine whether this lecturer does not randomly distribute the correct answer over the five choices. If this is true, how does it affect Pat's strategy?

Sample statistics: $n(1) = 8, n(2) = 4, n(3) = 3, n(4) = 8, n(5) = 2$

Computer applications

The following exercises require the use of a computer and software.

- 14.14 XR14-14** Finance managers are interested in the speed with which customers who make purchases on credit pay their bills. In addition to calculating the average number of days that unpaid bills (called accounts receivable) remain outstanding, they often prepare an ageing schedule. An ageing schedule classifies outstanding accounts receivable according to the time that has elapsed since billing, and records the proportion of accounts receivable belonging to each classification. A large firm has determined its ageing schedule for the past five years. These results are shown in the accompanying table. During the past few months, however, the economy has taken a downturn.

The company would like to know if the recession has affected the ageing schedule. A random sample of 250 accounts receivable was drawn and each account was classified as follows:

- 1 = 0–14 days outstanding
- 2 = 15–29 days outstanding
- 3 = 30–59 days outstanding
- 4 = 60 or more days outstanding.

Number of days outstanding	Proportion of accounts receivable (past five years)
0–14	0.72
15–29	0.15
30–59	0.10
60 and over	0.03

- 14.15 XR14-15** The results of a multinomial experiment with $k = 5$ are recorded. Each outcome is identified by the numbers 1 to 5 stored in column 1. Test to determine if there is enough evidence to infer that the proportions of outcomes differ. (Use $\alpha = 0.05$.)

Sample statistics: $n(1) = 28$, $n(2) = 17$, $n(3) = 19$, $n(4) = 17$, $n(5) = 19$

- 14.16 XR14-16** In an election held last year that was contested by three parties, party A captured 31% of the vote, party B garnered 51%, and party C received the remaining votes. A recent survey of 1200 voters asked each to identify the party that they would vote for in the next election. These results are recorded as 1 = party A, 2 = party B, and 3 = party C. Can we infer at the 10% significance

level that voter support has changed since the election?

Sample statistics: $n(1) = 408$, $n(2) = 571$, $n(3) = 221$

- 14.17 XR14-17** In a number of pharmaceutical studies, volunteers who take placebos (but are told they have taken a cold remedy) report the following side effects:

Headache (1)	5%
Drowsiness (2)	7%
Stomach upset (3)	4%
No side effect (4)	84%

A random sample of 250 people who were given a placebo, but who thought they had taken an anti-inflammatory drug, reported whether they had experienced each of the side effects. These data are recorded using the codes in parentheses. Can we infer at the 5% significance level that the reported side effects of the placebo for an anti-inflammatory drug differ from those of a cold remedy?

- 14.18 XR14-18** Registration records in an Australian state reveal that 15% of cars are small (1), 25% are medium (2), 40% are large (3), and the rest are an assortment of other styles and models (4). A random sample of accidents involving cars registered in the state was drawn. The type of car was recorded using the codes in parentheses. Can we infer that certain sizes of cars are involved in a higher than expected percentage of accidents?

14.2 Chi-squared test of a contingency table

In Chapter 3, we developed the *cross-classification table* as a first step in graphing the relationship between two nominal variables. Our goal was to determine whether the two variables were related. In this section we extend the technique to statistical inference. We introduce another chi-squared test, this one designed to satisfy two different problem objectives. The **chi-squared test of a contingency table** is used to determine whether there is enough evidence to infer that two nominal variables are related, and to infer that differences exist between two or more populations of nominal variables. Completing both objectives entails classifying items according to two different criteria. To see how this is done, consider the following example.

chi-squared test of a contingency table

Test used to determine whether there is enough evidence to infer that two nominal variables are related, and to infer that differences exist between two or more populations of nominal variables.

14.2a Test statistic

The test statistic is the same as the one employed to test proportions in the multinomial experiment. That is, the test statistic is

$$\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

where k is the number of cells in the contingency table. If you examine the null hypotheses described in the chi-squared test of a multinomial experiment and in the one specified above, you will discover a major difference. In the chi-squared test of a multinomial experiment, the null hypothesis specifies values for the probabilities p_i . The null hypothesis for the chi-squared test of a contingency table only states that the two variables are independent. However, we need the probabilities in order to calculate the expected values (e_i), which in turn permit us to calculate the value of the test statistic. (The entries in the contingency table are the observed values, o_i .) The question immediately arises: From where do we get the probabilities? The answer is that they will come from the data after we assume that the null hypothesis is true.

Expected frequencies for a contingency table

The expected frequency of the cell in column j and row i is

$$e_{ij} = \frac{(\text{column } j \text{ total}) \cdot (\text{row } i \text{ total})}{\text{sample size}}$$

We illustrate this test in detail using the following example.

EXAMPLE 14.3

LO2

Political affiliation versus economic options

XM14-03 and **XM14-03T** One of the issues that came up in a recent national election (and is likely to arise in many future elections) is how to improve Australian economic growth. Specifically, should governments cut public spending and/or introduce tax reforms, job creation and increase education funding? Politicians need to know which sectors of the electorate support these options to improve economic growth. Suppose that a random sample of 1000 voters was asked which option they support and their political affiliations: Labor, Liberal–National Coalition or Others (which included a variety of political persuasions). The responses were summarised in a table called a *contingency* or *cross-classification table* and shown below. Do these results allow us to conclude that political affiliation affects support for the economic options?

Economic options	Political affiliation			Total
	Labor (A1)	Coalition (A2)	Others (A3)	
Cut public spending (B1)	101	282	61	444
Introduce tax reforms (B2)	38	67	25	130
Job creation (B3)	131	88	31	250
Increase education funding (B4)	61	90	25	176
Total	331	527	142	1000



Solution

One way to solve the problem is to consider that there are two variables represented by the contingency table: economic options and political affiliation. Both are nominal. The values of economic options are 'cut public spending' (B1), 'introduce tax reforms' (B2), 'job creation' (B3) and 'increase education funding' (B4). The values of political affiliation are 'Labor' (A1), 'Coalition' (A2) and 'Others' (A3). The problem objective is to analyse the relationship between the two variables. Specifically, we want to know whether one variable affects the other.

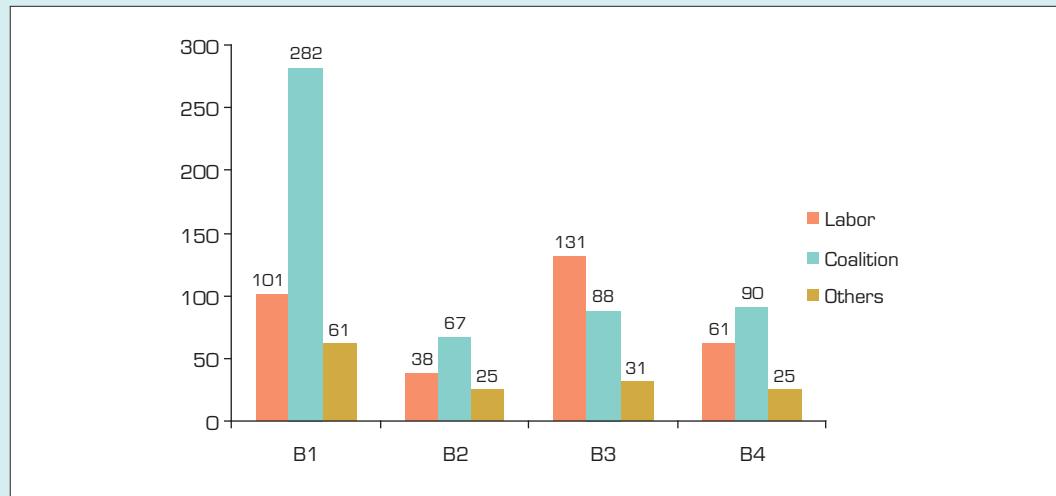
Another way of addressing the problem is to determine whether differences exist among the Labor, Coalition and Others affiliates. In other words, treat each political group as a separate population. Each population has four possible values, represented by the four economic options. (We can also answer the question by treating the economic options as populations and the political affiliations as the values of the random variable.) Here the problem objective is to compare three populations.

As you will discover shortly, both objectives lead to the same test. Consequently, we can address both objectives at the same time.

Graphical technique

Figure 14.2 depicts the graphical technique introduced in Chapter 3 to show the relationship (if any) between the two nominal variables. The bar chart displays the data from the sample. Based on the displayed sample information, it appears that there is a relationship between the two nominal variables. However, to draw inferences about the population of voters, we need to apply an inferential technique. We will now formally test this observation using the chi-squared test of a multinomial experiment.

FIGURE 14.2 Bar chart of economic options and party affiliation



The null hypothesis will specify that there is no relationship between the two variables. We state this in the following way:

H_0 : The two variables are independent.

The alternative hypothesis specifies that one variable affects the other, and is expressed as

H_A : The two variables are dependent.

If the null hypothesis is true, political affiliation and economic options are independent of one another. This means that whether someone is a Labor, Coalition or Others affiliate does not affect his or her economic choice. Consequently, there is no difference among Labor, Coalition and Others affiliates in their support for the four economic options. If the alternative hypothesis is true, political affiliation does affect which economic option is preferred. Thus, there are differences among the three political groups.



Calculating the expected frequencies

If we consider each political affiliation to be a separate population, each column of the contingency table represents a multinomial experiment with four cells. If the null hypothesis is true, the three multinomial experiments should produce similar proportions in each cell. We can estimate the marginal row probabilities by calculating the total in each row and dividing by the sample size. Thus,

$$\begin{aligned} P(\text{Cut public spending}) = P(B1) &\approx \frac{101+282+61}{1000} = \frac{444}{1000} \\ P(\text{Introduce tax reforms}) = P(B2) &\approx \frac{38+67+25}{1000} = \frac{130}{1000} \\ P(\text{Job creation}) = P(B3) &= \frac{131+88+31}{1000} = \frac{250}{1000} \\ P(\text{Increase education funding}) = P(B4) &= \frac{61+90+25}{1000} = \frac{176}{1000} \end{aligned}$$

The notation \approx indicates that these probabilities are only approximations (because they are estimated on the basis of sample data).

In a similar manner the marginal column probabilities are estimated by dividing the column sums by the total sample size.

$$\begin{aligned} P(\text{Labor}) = P(A1) &\approx \frac{331}{1000} \\ P(\text{Coalition}) = P(A2) &\approx \frac{527}{1000} \\ P(\text{Other}) = P(A3) &\approx \frac{142}{1000} \end{aligned}$$

Having estimated the marginal column and row probabilities, we can proceed to estimate the cell probabilities. Recall from Chapter 6 that, if events X and Y are independent, $P(X \text{ and } Y) = P(X) \cdot P(Y)$. Thus, assuming the null hypothesis to be true, we may apply this multiplication rule for independent events to obtain the joint probability that a voter falls into the first cell:

$$P(A1 \text{ and } B1) = P(A1) \cdot P(B1) \approx \left(\frac{331}{1000} \right) \left(\frac{444}{1000} \right)$$

By multiplying this joint probability by the sample size, we obtain the expected number of voters who fall into the first cell:

$$e_{11} = n \cdot P(A1 \text{ and } B1) \approx 1000 \left(\frac{331}{1000} \right) \left(\frac{444}{1000} \right) = \frac{(331)(444)}{1000} = 146.96$$

Observe that e_{11} was calculated by multiplying the total of column one by the total of row one and dividing by n . The other expected frequencies are estimated in a similar manner, using the following general formula. The expected frequency of the cell in column j and row i is

$$e_{ij} = \frac{(\text{column } j \text{ total})(\text{row } i \text{ total})}{\text{sample size}}$$

Expected value of the economic options of the three political affiliations

Labor affiliates	Economic option	Expected value
	Cut public spending	$e_{11} = \frac{(331 \times 444)}{1000} = 146.96$
	Introduce tax reforms	$e_{21} = \frac{(331 \times 130)}{1000} = 43.03$

Labor affiliates	Economic option	Expected value
	Job creation	$e_{31} = \frac{(331 \times 250)}{1000} = 82.75$
	Education funding	$e_{41} = \frac{(331 \times 176)}{1000} = 58.26$
Coalition affiliates	Economic option	Expected value
	Cut public spending	$e_{12} = \frac{(527 \times 444)}{1000} = 233.99$
	Introduce tax reforms	$e_{22} = \frac{(527 \times 130)}{1000} = 68.51$
	Job creation	$e_{32} = \frac{(527 \times 250)}{1000} = 131.75$
	Education funding	$e_{42} = \frac{(527 \times 176)}{1000} = 92.75$
Others affiliates	Economic option	Expected value
	Cut public spending	$e_{13} = \frac{(142 \times 444)}{1000} = 63.05$
	Introduce tax reforms	$e_{23} = \frac{(142 \times 130)}{1000} = 18.46$
	Job creation	$e_{33} = \frac{(142 \times 250)}{1000} = 35.50$
	Education funding	$e_{43} = \frac{(142 \times 176)}{1000} = 24.99$

The observed and expected cell frequencies are shown in the table below. As in the case of the chi-squared test of the multinomial experiment, the expected cell frequencies should satisfy the rule of five.

Contingency table for Example 14.3

Economic options	Political affiliation						Total	
	Labor (A1)		Coalition (A2)		Others (A3)			
	Observed	Expected	Observed	Expected	Observed	Expected		
Cut public spending (B1)	101	146.96	282	233.99	61	63.05	444	
Introduce tax reforms (B2)	38	43.03	67	68.51	25	18.46	130	
Job creation (B3)	131	82.75	88	131.75	31	35.50	250	
Increase education funding (B4)	61	58.26	90	92.75	25	24.99	176	
Total	331		527		142		1000	

The complete test is as follows:

Hypotheses:

H_0 : The two variables are independent (party affiliation is independent of economic options).

H_A : The two variables are dependent.

Test statistic:

$$\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i} \sim \chi^2_{(r-1)(c-1)} \quad \text{with } k = rc, r = 4, c = 3$$

Significance level: $\alpha = 0.05$

Decision rule:

To determine the rejection region, we need to know the number of degrees of freedom associated with this χ^2 -statistic. The number of degrees of freedom for a contingency table with r rows and c columns is
 $d.f. = (r - 1)(c - 1)$

For this example, the number of degrees of freedom is

$$d.f. = (r - 1)(c - 1) = (4 - 1)(3 - 1) = 6$$

Therefore, the decision rule is

$$\text{Reject } H_0 \text{ if } \chi^2 > \chi^2_{\alpha,(r-1)(c-1)} = \chi^2_{0.05,6} = 12.6.$$

$$\text{Alternatively, reject } H_0 \text{ if } p\text{-value} = P(\chi^2 > \chi^2_{\text{calculated}}) < \alpha = 0.05.$$

Value of the test statistic:**Calculating manually**

$$\begin{aligned} \chi^2 &= \sum_{i=1}^{12} \frac{(o_i - e_i)^2}{e_i} \\ &= \frac{(101 - 146.96)^2}{146.96} + \frac{(38 - 43.03)^2}{43.03} + \frac{(131 - 82.75)^2}{82.75} + \frac{(61 - 58.26)^2}{58.26} \\ &\quad + \frac{(282 - 233.99)^2}{233.99} + \frac{(67 - 68.51)^2}{68.51} + \frac{(88 - 131.75)^2}{131.75} + \frac{(90 - 92.75)^2}{92.75} \\ &\quad + \frac{(61 - 63.05)^2}{63.05} + \frac{(25 - 18.46)^2}{18.46} + \frac{(31 - 35.50)^2}{35.50} + \frac{(25 - 24.99)^2}{24.99} \\ &= 70.675 \end{aligned}$$

Notice that we continue to use a single subscript in the formula of the test statistic when we should use two subscripts, one for the row and one for the column. We feel that it is clear that for each cell, we need to calculate the squared difference between the observed and expected frequencies divided by the expected frequency. We don't believe that the satisfaction of using the mathematically correct notation would overcome the unnecessary complication.

Conclusion: Because $\chi^2 = 70.675 > 12.6$, we reject the null hypothesis.

Interpreting the results

We conclude that there is evidence of a relationship between political affiliation and support for the economic options. It follows that the three political affiliations differ in the support for the four economic options. We can see from the data that Coalition affiliates generally favour cutting public spending, whereas Labor affiliates prefer job creation.

Using the computer**Using Excel – Do it yourself Excel**

For this statistical technique, you will have to create the spreadsheet yourself. We call it 'Do It Yourself Excel'. We demonstrate how to create a spreadsheet to solve any exercise by providing instructions for Example 14.3. You can produce the χ^2 -statistic from either a contingency table that is already tabulated or raw data. The data file **XM14-03** contains raw data that use the following codes:

Column 1 (Economic options)	Column 2 (Party affiliation)
1 = Cut public spending	1 = Labor
2 = Introduce tax reforms	2 = Coalition
3 = Job creation	3 = Others
4 = Increase education funding	

Here are the Excel output and instructions.

Excel output for Example 14.3

	A	B	C	D	E
1	Observed frequency - Contingency Table (Pivot table)				
2		Party			
3	Option	1	2	3	Grand Total
4	1	101	282	61	444
5	2	38	67	25	130
6	3	131	88	31	250
7	4	61	90	25	176
8	Grand Total	331	527	142	1000

	G	H	I	J	K
1	Expected frequency				
2		Party			
3	Option	1	2	3	Grand Total
4	1	147	234	63	444
5	2	43	69	18	130
6	3	83	132	36	250
7	4	58	93	25	176
8	Grand Total	331	527	142	1000
9					
10	Chi-squared statistic				
11	Option	1	2	3	
12	1	14.38	9.85	0.07	
13	2	0.59	0.03	2.32	
14	3	28.13	14.53	0.57	
15	4	0.13	0.08	0.00	
16					
17	Chi-squared test statistic=		70.67		
18	df=		6		
19	p-value=		0.0000		
20	Alpha=		0.05		
21	Chi-squared critical=		12.59		

COMMANDS

- Type or open the data file (**XM14-03**). Column A represents the codes for one nominal variable, and column B represents the codes for the other nominal variable. The codes must be positive integers.
- Use the **PivotTable** commands in Example 3.7 (page 70) to create the **cross-classification table** or **contingency table (frequencies)** with observed frequencies (count) and copy the values in a new worksheet **Pivot table** (cells **A2:E8**, including the titles and totals).
- Calculate the **expected values**. For example, to calculate the expected value for the first cell type in cell H4: **=B\$8*\$E4/\$E\$8**. Drag down the column and then across the rows to compute all the expected values.
- To compute the **chi-squared statistic**, first calculate the values in each cell. Start in H12 and type **=((B4-H4)^2)/H4**. Drag down the column and then across the rows.
- Calculate the **chi-squared statistic**. In cell I17, type: **=SUM(H12:J15)**.
- Compute the **p-value** using the **CHIDIST** function in which we specify the value of the chi-squared statistic (I17) and the degrees of freedom (6). In cell I19, type **=CHIDIST(I17,6)**.

Using XLSTAT

XLSTAT output for Example 14.3

	B	C	D	E	F
19	Test of independence between the rows and the columns (Option / Party):				
20	Chi-square (Observed value)	70.675			
21	Chi-square (Critical value)	12.592			
22	DF	6			
23	p-value	< 0.0001			
24	alpha	0.05			



COMMANDS

- 1 Type the data in two columns or open the data file (XM14-03).
- 2 Click **XLSTAT, Correlation/Association tests**, and **Tests on contingency tables (Chi-square...)**.
- 3 Specify the range of the **Row variables (A1:A1001)** and the range of the **Column Variables (B1:B100199)**. Select the **Data format Qualitative variables**.
- 4 Click **Options** and check **Chi-square test**.
- 5 Click **Outputs** and check **Contingency table**. Click **OK**.

14.2b Degrees of freedom for contingency table

To locate the critical value, we have used the number of degrees of freedom for a contingency table with r rows and c columns as

$$\text{d.f.} = (r - 1)(c - 1)$$

We will briefly indicate why this is so. In general, the number of degrees of freedom for a chi-squared distribution is $(k - 1 - m)$, where k is the number of cells and m is the number of independent population parameters that must be estimated from the sample data before the expected frequencies can be determined. No parameters were estimated in the test involving a multinomial experiment, so m was equal to zero and the number of degrees of freedom was $k - 1$. For a contingency table with r rows and c columns, however, there are $k = rc$ cells, and the sample data are used to estimate $r - 1$ row probabilities and $c - 1$ column probabilities. Once these are estimated, the one remaining row probability and the one remaining column probability are automatically determined, since both the row probabilities and the column probabilities must sum to one. Therefore, the number of degrees of freedom for a contingency table with r rows and c columns is

$$\text{d.f.} = rc - 1 - [(r - 1) + (c - 1)] = (r - 1)(c - 1)$$

If you didn't understand our perfectly clear analysis of how the degrees of freedom are determined, don't worry. The important thing for you to know is that *the number of degrees of freedom for the chi-square test of a contingency table is $(r - 1)(c - 1)$* .

14.2c Data formats

In Example 14.3, the data were stored in two columns, one column containing the values of one nominal variable and the second column storing the values of the second nominal variable. The data can be stored in another way. In Example 14.3, we could have recorded the data in three columns, one column for each party affiliation. The columns would contain the codes representing the economic option. Alternatively, we could have stored the data in four columns, one column for each economic option. The columns would contain the codes for the party affiliation. In either case, we have to count the number of each value and construct the cross-tabulation table using the counts.

EXAMPLE 14.4

L01 L03

Quality of workmanship among three daily shifts

XM14-04 The operations manager of a company that manufactures shirts wants to determine whether there are differences in the quality of workmanship among the three daily shifts. She randomly selects 600 recently made shirts and carefully inspects them. Each shirt is classified as either perfect or flawed, and the shift that produced



it is also recorded. The accompanying table summarises the number of shirts that fell into each cell. Do these data provide sufficient evidence at the 5% significance level to infer that there are differences in quality among the three shifts?

Contingency table classifying shirts

Shirt condition	Shift		
	1	2	3
Perfect	240	191	139
Flawed	10	9	11

Solution

Identifying the technique

The problem objective is to compare three populations (the shirts produced by the three shifts). The data are nominal because each shirt will be classified as either *perfect* or *flawed*. This problem-objective–data-type combination indicates that the statistical procedure to be employed is the chi-squared test of a contingency table. The null and alternative hypotheses are as follows:

Hypotheses:

H_0 : The two variables are independent (i.e. shirt condition is independent of shift).

H_A : The two variables are dependent.

Test statistic:

$$\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i} \sim \chi^2_{(r-1)(c-1)}, \text{ where } r = 2 \text{ and } c = 3$$

$$\text{d.f.} = (r-1)(c-1) = (2-1)(3-1) = 2$$

Level of significance:

$$\alpha = 0.05$$

Decision rule:

$$\text{Reject } H_0 \text{ if } \chi^2 > \chi^2_{\alpha,(r-1)(c-1)} = \chi^2_{0.05,2} = 5.99.$$

Value of the test statistic:

Calculating manually

We calculated the row and column totals and used them to determine the expected values. For example, the expected number of perfect shirts produced in shift 1 is

$$e_{11} = \frac{(250 \times 570)}{600} = 237.5$$

The remaining expected values are calculated similarly. The original table and expected values are shown in the table below.

Shirt condition	Shift						Total	
	1		2		3			
	f_i	e_i	f_i	e_i	f_i	e_i		
Perfect	240	237.5	191	190.0	139	142.5	570	
Flawed	10	12.5	9	10.0	11	7.5	30	
Total	250		200		150		600	

The value of the test statistic is

$$\begin{aligned} \chi^2 &= \sum_{i=1}^6 \frac{(o_i - e_i)^2}{e_i} \\ &= \frac{(240 - 237.5)^2}{237.5} + \frac{(10 - 12.5)^2}{12.5} + \frac{(191 - 190.0)^2}{190.0} + \frac{(9 - 10.0)^2}{10.0} + \frac{(139 - 142.5)^2}{142.5} + \frac{(11 - 7.5)^2}{7.5} \\ &= 2.36 \end{aligned}$$

Conclusion: As $c^2 = 2.36 < 5.99$ or as $p\text{-value (output)} = 0.3087 > 0.05$, do not reject the null hypothesis.



Interpreting the results

There is not enough evidence to allow us to conclude that differences in quality exist between the three shifts. This means that the shirt condition is independent of the shift in which it was produced. Had we discovered differences, we would attempt to determine why the shifts differ in quality. Possible explanations include workers on one shift using a different operating procedure or incompetent supervisors. In this case, as no significant differences were detected, any improvements in quality must be accomplished over all three shifts. These could include, for example, acquiring newer machines, training the employees or instituting statistical process control.

Using the computer

Using Excel – Do it yourself Excel

For this statistical technique, as only the observed frequencies (contingency table) are given, you will have to create the spreadsheet yourself as demonstrated in Example 14.3. You can produce the chi-squared statistic from the contingency table that is already tabulated by following similar Excel commands as in Example 14.3. Please note that the number of rows and columns are different and the cell references should be changed accordingly. Below we give the output.

Excel output for Example 14.4

	A	B	C	D	E
1	Observed frequency (Contingency table)				
2		Shift 1	Shift 2	Shift 3	Grand total
3	Perfect	240	191	139	570
4	Flawed	10	9	11	30
5	Grand total	250	200	150	600
6					
7	Expected frequency				
8		Shift			
9	Quality	Shift 1	Shift 2	Shift 3	Grand total
10	Perfect	238	190	143	570
11	Flawed	13	10	8	30
12	Grand total	250	200	150	600
13					
14	Chi-squared statistic				
15	Option	1	2	3	
16	1	0.03	0.01	0.09	
17	2	0.50	0.10	1.63	
18					
19	Chi-squared test statistic=	2.351			
20	df=		2		
21	p-value=		0.3087		
22	Alpha=		0.05		
23	Chi-squared critical=	5.99			

The Excel commands (Do it Yourself) are the same as in Example 14.3 (note that as r and c are different in this example, the cell references in each calculation will have to be changed accordingly).

Using XLSTAT

As raw data are not given, XLSTAT cannot be used to complete the test.

14.2d Rule of five (II)

In the previous section, we pointed out that the expected values should be at least 5 to ensure that the chi-squared distribution provides an adequate approximation of the sampling distribution. *In a contingency table where one or more cells have expected values of less than five, we need to combine rows or columns to satisfy the rule of five.* To illustrate, suppose that we want to test for dependence in the following contingency table.

	C1	C2	C3	Total
R1	10	14	4	28
R2	12	16	7	35
R3	8	8	4	20
Total	30	38	15	83

The expected values are as follows:

	C1	C2	C3	Total
R1	10.1	12.8	5.1	28
R2	12.7	16.0	6.3	35
R3	7.2	9.2	3.6	20
Total	30	38	15	83

The expected value of the cell in row 3, column 3 is less than 5. To eliminate the problem, we can add column 3 to one of columns 1 and 2 or add row 3 to either row 1 or row 2. The combining of rows or columns should be done so that the combination forms a logical unit, if possible. For example, if the columns represent the age groups, young (under 40), middle-aged (40–65) and senior (over 65), it is logical to combine columns 2 and 3. The observed and expected values are combined to produce the following table (expected values in parentheses).

	C1		C2,3		Total
	f_i	e_i	f_i	e_i	
R1	10	10.1	18	17.9	28
R2	12	12.7	23	(22.3)	35
R3	8	7.2	12	12.8	20
Total	30		53		83

The degrees of freedom must be changed as well. The number of degrees of freedom of the original contingency table is $(3 - 1) \times (3 - 1) = 4$. The number of degrees of freedom of the combined table is $(3 - 1) \times (2 - 1) = 2$. The rest of the procedure is unchanged.

Here is a summary of the factors that tell us when to apply the chi-squared test of a contingency table. Note that there are two problem objectives satisfied by this statistical procedure.

IN SUMMARY

Factors that identify the chi-squared test of a contingency table

- 1 *Problem objectives*: to analyse the relationship between two variables; to compare two or more populations
- 2 *Data type*: nominal (categorical)

We will now present the solution for the opening example to this chapter.

SPOTLIGHT ON STATISTICS

Has support for the death penalty for drug trafficking changed since 2005? Solution

Identifying the technique

The problem objective is to compare public opinion in three different years.

The variable is nominal because its values are 'Support', 'Oppose' and 'Can't say' respectively. The appropriate technique is the chi-squared test of a contingency table.

The hypotheses are:

H_0 : The two variables (opinion and year) are independent.

H_A : The two variables are dependent.

In this application, the two variables are year (2005, 2010 and 2015) and the answer to the question (opinion) posed by the survey about carrying out the death penalty for drug trafficking (Support, Oppose and Can't say). To produce the statistical result, we will need to count the number in support, the number opposed and the number who responded 'Can't say' in each of the three years. The following table was determined by summarising the responses for each year.

Opinion	Year			
	2005	2010	2015	
Support	570	600	715	
Oppose	360	528	591	
Can't say	70	72	69	

Using the computer

We use the same Excel commands as in Example 14.3 (see pages 599–600) using the contingency table. As only the observed frequencies (contingency table) are given, you will have to create the spreadsheet yourself as demonstrated in Example 14.3. You can produce the chi-squared statistic from the contingency table that is already tabulated by following Excel commands similar to those used in Example 14.3. Please note that as the numbers of rows and columns are different, the cell references should be changed accordingly. The resulting output is given below.

	A	B	C	D	E
1	Observed frequencies				
2		2005	2010	2015	Grand total
3	Support	570	600	715	1885
4	Oppose	360	528	591	1479
5	Can't say	70	72	69	211
6	Grand Total	1000	1200	1375	3575
7					
8	Expected frequency				
9		Party			
10		2005	2010	2015	Grand total
11	Support	527	633	725	1885
12	Oppose	414	496	569	1479
13	Can't say	59	71	81	211
14	Grand Total	1000	1200	1375	3575
15					
16	Chi-squared statistic				
17		2005	2010	2015	
18	Support	3.46	1.69	0.14	
19	Oppose	6.97	2.01	0.86	
20	Can't say	2.04	0.02	1.82	
21					
22	Chi-squared test statistic=	19.02			
23	df=		4		
24	p-value=	0.0008			
25	Alpha=	0.05			
26	Chi-squared critical=	9.49			



Source: iStock.com/erhui1979

▶ Interpreting the results

The p -value is 0.0008. There is not enough evidence to infer at the 5% level of significance that the two variables (opinion and year) are independent. Thus, we can conclude that support for the death penalty for drug trafficking does vary from year to year.

EXERCISES

Learning the techniques

- 14.19 XR14-19** Conduct a test to determine whether the two classifications L and M are independent, using the data in the following contingency table and $\alpha = 0.05$.

	M_1	M_2
L_1	28	68
L_2	56	36

- 14.20 XR14-20** Repeat Exercise 14.19 using the following table.

	M_1	M_2
L_1	14	34
L_2	28	18

- 14.21 XR14-21** Repeat Exercise 14.19 using the following table.

	M_1	M_2
L_1	7	17
L_2	14	9

- 14.22** Review the results of Exercises 14.19–14.21. What is the effect of decreasing the sample size?

- 14.23 XR14-23** Conduct a test to determine whether the two classifications R and C are independent, using the data in the accompanying contingency table and $\alpha = 0.10$.

R	C		
	C_1	C_2	C_3
R_1	40	32	48
R_2	30	48	52

Applying the techniques

- 14.24 XR14-24 Self-correcting exercise.** The trustee of a company's superannuation scheme has solicited the opinions of a sample of the company's

employees towards a proposed revision of the scheme. A breakdown of the responses is shown in the following table. Is there evidence that the responses differ among the three groups of employees? (Test at the 5% level of significance.)

Responses	Blue-collar workers	White-collar workers	Managers
For	67	32	11
Against	63	18	9

- 14.25 XR14-25** A market survey was conducted for the purpose of forming a demographic profile of individuals who would like to own an iWatch. This profile will help to establish the target market for the iWatch, which in turn will be used in developing an advertising strategy. The portion of data collected that relates to the consumers' gender is summarised in the table. Is there sufficient evidence to conclude that the desire to own an iWatch is related to the consumer's gender? (Test using $\alpha = 0.05$.)

Response	Men	Women
Want iWatch	32	20
Don't want iWatch	118	130

- 14.26 XR14-26** A survey analysing the relationship between newspapers read and occupational class was undertaken. A sample of newspaper readers was asked to name the newspaper they read and to state whether they are a blue-collar worker, a white-collar worker or a professional. The following data were generated. Can we infer that occupational class and newspaper read are related?

	Blue-collar worker	White-collar worker	Professional
Newspaper A (N1)	27	29	33
Newspaper B (N2)	18	43	51
Newspaper C (N3)	38	15	24
Newspaper D (N4)	37	21	18

14.27 A well-known consulting firm wants to test how it can influence the proportion of questionnaires returned for its surveys. In the belief that the inclusion of an inducement to respond may be influential, the firm sends out 1000 questionnaires: 200 promise to send respondents a summary of the survey results; 300 indicate that 20 respondents (selected by a lottery) will be awarded gifts; and 500 are accompanied by no inducements. Of these, 80 questionnaires promising a summary, 100 questionnaires offering gifts, and 120 questionnaires offering no inducements are returned. What can you conclude from these results?

14.28 XR14-28 The marketing analyst for a cola company believes that the cola preferences of Americans and Canadians are quite similar, and she wishes to determine if there are differences in cola preferences between this combined group and Australians. She obtains the accompanying table of data. Using $\alpha = 0.025$, test for a difference in cola preferences.

Nationality	Cola preference				Total
	A	B	C	D	
American/Canadian	98	18	28	56	200
Australian	7	10	14	19	50
Total	105	28	42	75	250

Computer applications

14.29 XR14-29 A newspaper publisher, trying to pinpoint his market's characteristics, wondered whether the approach people take to reading a newspaper is related to the reader's educational level. A survey asked adult readers to report which section of the paper they read first and their highest educational level. These data were recorded (column 1 = first section read, where 1 = front page, 2 = sports, 3 = editorial, 4 = other; column 2 = educational level, where 1 = did not complete high school, 2 = high school graduate, 3 = university or TAFE college graduate, 4 = postgraduate degree). What do these data tell the publisher about how educational level affects the way adults read the newspaper?

14.30 XR14-30 An anti-smoking group recently had a large advertisement published in newspapers throughout Australia. In the hope that it would have meaningful impact, several statistical facts and medical details were included. The anti-smoking group is concerned, however, that smokers might have read

less of the advertisement than non-smokers. This concern is based on the belief that a reader tends to spend more time reading articles that agree with his or her predisposition. The anti-smoking group conducted a survey asking those who saw the advertisement if they read the headline only (1), some detail (2) or most of the advertisement (3). The questionnaire also asked respondents to identify themselves as either a heavy smoker – more than two packs per day (1), a moderate smoker – between one and two packs per day (2), a light smoker – less than one pack per day (3), or a non-smoker (4). The results are stored as column 1: type of smoker; column 2: survey responses. Do the data indicate that the anti-smoking group has reason to be concerned? (Use $\alpha = 0.05$.)

Sample statistics:

The numbers of combinations of smoker categories and survey responses:

$$n(1,1) = 33; n(1,2) = 24; n(1,3) = 19;$$

$$n(2,1) = 23; n(2,2) = 17; n(2,3) = 26;$$

$$n(3,1) = 16; n(3,2) = 27; n(3,3) = 46;$$

$$n(4,1) = 14; n(4,2) = 38; n(4,3) = 57.$$

14.31 XR14-31 An investor who can correctly forecast the direction and size of changes in foreign currency exchange rates is able to reap huge profits in the international currency markets.

A knowledgeable reader of the *Australian Financial Review* (in particular, of the currency futures market quotations) can determine the direction of change in various exchange rates predicted by all investors, viewed collectively.

Predictions from 216 investors, together with the subsequent actual directions of change, are recorded. (Column 1: predicted change where 1 = positive and 2 = negative; column 2: actual change where 1 = positive and 2 = negative).

- a** Test the hypothesis (with $\alpha = 0.10$) that a relationship exists between the predicted and the actual directions of change.
- b** To what extent would you make use of these predictions in formulating your forecasts of future exchange rate changes?

Sample statistics:

The numbers of combinations of predicted changes and actual changes:

$$n(1,1) = 65; n(1,2) = 64; n(2,1) = 39; n(2,2) = 48.$$

14.32 XR14-32 During the past decade many cigarette smokers have attempted to quit. Unfortunately, nicotine is highly addictive. Smokers employ a large number of different methods to help them quit, including nicotine patches, hypnosis and various forms of therapy. A researcher for the Addiction Research Council wanted to determine why some people quit for good, while others relapse. He surveyed 1000 people who planned to quit smoking and recorded their educational level and whether, one year later, they continued to smoke. Educational level was recorded in the following way:

- 1 = Did not finish high school
- 2 = High school graduate
- 3 = University or TAFE college graduate
- 4 = Completed a postgraduate degree

A continuing smoker was recorded as 1; a quitter was recorded as 2. Can we infer that the amount of education is a factor in determining whether a smoker will quit?

Sample statistics:

The numbers of combinations of smoker categories and education level:

$$\begin{aligned}n(1,1) &= 34; n(1,2) = 251; n(1,3) = 159; n(1,4) = 16; \\n(2,1) &= 23; n(2,2) = 212; n(2,3) = 248; n(2,4) = 57.\end{aligned}$$

14.33 XR14-33 The relationship between pharmaceutical companies and medical researchers is under scrutiny because of a possible conflict of interest. The issue that started the controversy was a 1995 case-control study that suggested that the use of calcium-channel blockers to treat hypertension led to an increased risk of heart disease. This led to an intense debate in both technical journals and the press. Researchers writing in the *New England Journal of Medicine* ('Conflict of Interest in the Debate over Calcium Channel Antagonists', 8 January 1998, p. 101) looked at the 70 reports on the subject that appeared during 1996–97, classifying them as either favourable, neutral or critical towards the drugs. The researchers then contacted the authors of the reports and questioned them about financial ties to pharmaceutical companies. The results were recorded in the following way:

Column 1: Results of the scientific study:

- 1 = favourable; 2 = neutral; 3 = critical

Column 2: Ties to pharmaceutical companies:

- 1 = financial ties; 2 = no ties

Do these data allow us to infer that research findings on calcium-channel blockers are affected by whether the research is funded by a pharmaceutical company?

Sample statistics:

The numbers of combination of financial ties and results:

$$\begin{aligned}n(1,1) &= 29; n(1,2) = 10; n(1,3) = 9; n(2,1) = 1; n(2,2) = 7; \\n(2,3) &= 14.\end{aligned}$$

14.34 XR14-34 After a thorough analysis of the market, a publisher of business and economics statistics books has divided the market into three general approaches to teaching applied statistics: (1) use of a computer and statistical software with no manual calculations; (2) traditional teaching of concepts and solution of problems manually; (3) mathematical approach with emphasis on derivations and proofs.

The publisher wanted to know whether this market breakdown could be segmented on the basis of the educational background of the instructor. As a result, she organised a survey in which 195 lecturers of business and economics statistics were asked to report their approach to teaching and in which of the following disciplines they have their highest degree:

- 1 = Business
- 2 = Economics
- 3 = Mathematics or engineering
- 4 = Other

The data are coded and recorded.

- a Can the publisher infer that differences exist in type of degree among the three teaching approaches? If so, how can the publisher use this information?
- b Suppose that you work in the marketing department of a textbook publisher. Prepare a report for the publisher that describes your analysis.

Sample statistics:

The numbers of combinations of approaches and degrees:

$$\begin{aligned}n(1,1) &= 51; n(1,2) = 8; n(1,3) = 5; n(1,4) = 11; \\n(2,1) &= 24; n(2,2) = 14; n(2,3) = 12; n(2,4) = 8; \\n(3,1) &= 26; n(3,2) = 9; n(3,3) = 19; n(3,4) = 8.\end{aligned}$$

14.35 XR14-35 A statistics practitioner took random samples from Canada, Australia, New Zealand and the United Kingdom, and classified each person as

either obese (2) or not (1). Can we conclude from these data that there are differences in obesity rates between the four Commonwealth nations? (Source: Adapted from Statistical Abstract of the United States 2012, Table 1342.)

Sample statistics:

The numbers of combinations of obese classifications and countries:

$$\begin{aligned}n(1,1) &= 152; n(1,2) = 151; n(1,3) = 147; n(1,4) = 151; \\n(2,1) &= 48; n(2,2) = 49; n(2,3) = 53; n(2,4) = 49.\end{aligned}$$

14.36 XR14-36 To measure the extent of cigarette smoking around the world, random samples of adults in Denmark, Finland, Norway, and Sweden were drawn. Each was asked whether he or she smoked (2 = Yes, 1 = No). Can we conclude that there are differences in smoking between the four Scandinavian countries? (Source: Adapted from Statistical Abstract of the United States 2012, Table 1343.)

Sample statistics:

The numbers of combinations of smoking classification and countries:

$$\begin{aligned}n(1,1) &= 420; n(1,2) = 398; n(1,3) = 395; n(1,4) = 430; \\n(2,1) &= 80; n(2,2) = 102; n(2,3) = 105; n(2,4) = 70.\end{aligned}$$

14.37 XR14-37 Refer to Exercise 14.36. The survey was also performed in Canada, Australia, New Zealand and the United Kingdom. Is there enough evidence to infer that there are differences in adult cigarette smoking between the four Commonwealth countries? (Source: Adapted from Statistical Abstract of the United States 2012, Table 1343.)

Sample statistics:

The numbers of combinations of smoking classification and countries:

$$\begin{aligned}n(1,1) &= 165; n(1,2) = 167; n(1,3) = 164; n(1,4) = 156; \\n(2,1) &= 35; n(2,2) = 33; n(2,3) = 36; n(2,4) = 44.\end{aligned}$$

14.38 XR14-38 A survey asked a random sample of federal government and private sector workers to judge their well-being. The responses are 1 = thriving, 2 = struggling, 3 = suffering. Is there enough evidence to conclude that government and private sector workers differ in their well-being?

Sample statistics:

The numbers of combinations of well-being responses and sectors:

$$\begin{aligned}n(1,1) &= 63; n(1,2) = 69; n(2,1) = 50; n(2,2) = 85; \\n(3,1) &= 24; n(3,2) = 49.\end{aligned}$$

14.3 Chi-squared test for normality

The goodness-of-fit test for a multinomial population was introduced in Section 14.2. The chi-squared test described there can be used to test the hypothesis that a population has a particular probability distribution. Because use of the normal distribution is so prevalent – particularly in the assumptions adopted for many statistical techniques – it would be useful to be able to test whether or not a sample of data has been drawn from a normal population. This section describes one such test.

14.3a Test for a normal distribution

The **chi-squared goodness-of-fit test** for a normal distribution proceeds in essentially the same way as the chi-squared test for a multinomial population. But the multinomial test presented in Section 14.2 dealt with a single population of nominal data, whereas a normal distribution has numerical data. Therefore, we must begin by subdividing the range of the normal distribution into a set of intervals, or categories, in order to obtain nominal data. We will discuss the reasons for our choices of intervals later.

chi-squared goodness-of-fit test

This test is commonly used to test whether a data set comes from a normal population under a multinomial framework.

EXAMPLE 14.5

LO3

Is lifetime of batteries normally distributed?

XM14-05 A battery manufacturer wishes to determine whether the lifetimes of its batteries are normally distributed. Such information would be helpful in establishing the guarantee that should be offered. The lifetimes of a sample of 200 batteries are measured, and the resulting data are grouped into a frequency distribution, as shown in the following table. The mean and the standard deviation of the sample of lifetimes are calculated to be $\bar{x} = 164$ and $s = 10$. Can the manufacturer infer that the lifetimes of batteries are normally distributed?

Lifetime (in hours)	Number of batteries
140 up to 150	15
150 up to 160	54
160 up to 170	78
170 up to 180	42
180 up to 190	11
Total	200

Solution**Identifying the technique**

Notice that, by grouping the lifetimes into categories, we have obtained nominal data. Moreover, once we find the probability that an observation will fall into each of these categories, we have a multinomial experiment and can proceed as before. That is, the chi-squared test statistic can be used to compare the observed frequencies of lifetimes falling into the various categories with the expected frequencies, which are calculated under the assumption that the data are normally distributed. We therefore have the following hypotheses:

Hypotheses:

H_0 : The data are normally distributed.

H_A : The data are not normally distributed.

Test statistic:

$$\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i} \sim \chi^2_{k-3}$$

Recall from the previous section that the number of degrees of freedom for a chi-squared distribution is $(k - 1 - m)$, where k is the number of categories and m is the number of population parameters that must be estimated from the sample data. As we are estimating μ and σ from the sample data, $m = 2$ and d.f. = $k - 3$. The number of categories in the frequency distribution is 5. To accommodate all possible values in the hypothesised normal population, however, the first and last categories should be redefined as open-ended categories (less than 150 and 180 or more, respectively). We therefore have $k = 5$ categories and $k - 3 = 2$ degrees of freedom.

Level of significance: $\alpha = 0.10$

Decision rule: Reject H_0 if $\chi^2 > \chi^2_{\alpha, k-3} = \chi^2_{0.10, 2} = 4.61$ or if $p\text{-value} < \alpha = 0.10$.

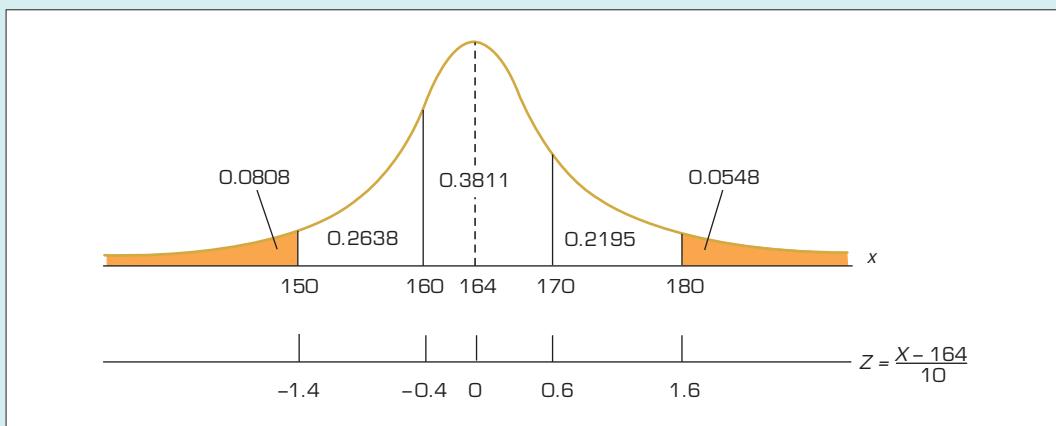
Value of the test statistic:

Before calculating the value of the test statistic, we must first determine the expected frequency for each category. In a sample of 200 observations, the expected frequency for any category is simply 200 times the probability that an observation will fall into that category. Assuming that the null hypothesis is correct, this probability can be found by using the standard normal table. The hypothesised normal distribution is shown in

Figure 14.3, together with the probabilities that a lifetime X will fall into the various categories. For example, using a standard normal table, the probability that a lifetime falls between 160 and 170 hours is

$$\begin{aligned} P(160 \leq X \leq 170) &= P\left(\frac{160 - 164}{10} \leq \frac{X - 164}{10} \leq \frac{170 - 164}{10}\right) \\ &= P(-0.4 \leq Z \leq 0.6) = 0.3811 \end{aligned}$$



**FIGURE 14.3** Hypothesised normal distribution of battery lifetimes

The expected frequency for this category is therefore $(200)(0.3811) = 76.22$. Similarly, the expected number of lifetimes falling below 150 hours is $(200)(0.0808) = 16.16$. The remaining expected frequencies are calculated in a similar manner and recorded in the following table, which also contains the calculation of the chi-squared test statistic.

Calculation of χ^2 for 200 battery lifetimes

Lifetime (in hours)	Observed frequency f_i	Expected frequency e_i	$(f_i - e_i)^2$	$\frac{(f_i - e_i)^2}{e_i}$
Less than 150	15	16.16	1.35	0.083
150 up to 160	54	52.76	1.54	0.029
160 up to 170	78	76.22	3.17	0.042
170 up to 180	42	43.90	3.61	0.082
180 or more	11	10.96	0.00	0.000
Total	200	200.00		$\chi^2 = 0.236$

Conclusion: As the calculated value of $\chi^2 = 0.236$ does not exceed the critical value of 4.61, we cannot reject the null hypothesis.

Interpreting the results

The test results reveal that there is some support for the null hypothesis at the 10% level of significance that the lifetimes of the manufacturer's batteries are normally distributed.

14.3b Choosing class intervals

In practice, you can use any intervals you like. We chose the intervals we did to facilitate the calculation of the normal probabilities. The number of intervals was chosen to comply with the rule of five, which requires that all expected values be at least equal to 5. Because the number of degrees of freedom is $k - 3$, the minimum number of intervals is $k = 4$.

14.3c Using the computer

We programmed Excel to calculate the value of the test statistic so that the expected values are at least 5 (where possible) and the minimum number of intervals is four. Hence, if the number of observations is more than 220, the intervals and probabilities are:

Interval	Probability
$Z \leq -2$	0.0228
$-2 < Z \leq -1$	0.1359
$-1 < Z \leq 0$	0.3413
$0 < Z \leq 1$	0.3413
$1 < Z \leq 2$	0.1359
$Z > 2$	0.0228

If the sample size is less than or equal to 220 and greater than 80, the intervals are:

Interval	Probability
$Z \leq -1.5$	0.0668
$-1.5 < Z \leq -0.5$	0.2417
$-0.5 < Z \leq 0.5$	0.3829
$0.5 < Z \leq 1.5$	0.2417
$Z > 1.5$	0.0668

If the sample size is less than or equal to 80, we employ the minimum number of intervals, four. [When the sample size is less than 32, at least one expected value will be less than five.] The intervals are:

Interval	Probability
$Z \leq -1$	0.1587
$-1 < Z \leq 0$	0.3413
$0 < Z \leq 1$	0.3413
$Z > 1$	0.1587

EXAMPLE 14.6

LO7

Distance travelled by a fleet of taxis

Consider Example 10.2, in which the required condition is that the distance travelled by a fleet of taxis is normally distributed. Verify this required condition.

Solution

As we have raw data, we first standardise the data using the sample mean and standard deviation and then count the number of observations (observed frequencies) falling in the intervals $Z \leq -1$, $-1 < Z \leq 0$, $0 < Z \leq 1$ and $Z > 1$. Using standard normal probabilities for each interval, we can obtain the expected frequencies. Follow the steps used in Example 14.5. The output is shown below.



Excel output for normality test in Example 10.2

	D	E	F	G	H
1	Mean=	7.7			
2	Std dev=	2.92672922			
3					
4		Observed	Probability	Expected	$(f_i - e_i) / e_i$
5	Z <= -1.0	6	0.1587	6.5	0.0392
6	-1 < Z <= 0	12	0.3413	14.0	0.2844
7	0 < Z <= 1.0	18	0.3413	14.0	1.1460
8	Z > 1.0	5	0.1587	6.5	0.3481
9	Total	41		41	1.8178
10					
11	Chi-squared test statistic=			1.8178	
12	df=			1	
13	p-value=			0.1776	
14	Alpha=			0.05	
15	Chi-squared critical=			3.8415	

As the p -value of the normality test is 0.1776, which is greater than $\alpha = 0.05$, we do not reject the null hypothesis of normality. That is, the distance travelled by a fleet of taxis (in Example 10.2) is normally distributed.

14.3d Interpreting the results of a chi-squared test for normality

In applications where a t -test is employed, we require the underlying population to be normally distributed. However, even if we found evidence of non-normality, this would not necessarily invalidate the usual t -test we conduct to make inferences about the mean. As we pointed out in earlier chapters, the t -test of a mean is a robust procedure, which means that only if the variable is extremely non-normal and the sample size is small, is the conclusion of the technique suspect. The problem here is that if the sample size is large and the variable is only slightly non-normal, the chi-squared test for normality will, in many cases, conclude that the variable is not normally distributed. However, if the variable is even quite non-normal and the sample size is large, the t -test will still be valid. Although there are situations where we need to know whether a variable is non-normal, we continue to advocate that the way to decide if the normality requirement for almost all statistical techniques applied to numerical data is satisfied is to draw histograms and look for shapes that are far from bell shaped (e.g. highly skewed or bimodal).

We conclude this section by making three points. First, if the value of either the mean or the standard deviation of the population is hypothesised, rather than estimated from the sample data, the number of degrees of freedom for the chi-squared test statistic must be adjusted accordingly. The number of degrees of freedom is $(k - 2)$ if only one of the parameters is estimated, and it is $(k - 1)$ if neither parameter is estimated. Second, when applied to small samples of data, the chi-squared test usually fails to reject the hypothesis of normality when the data have a symmetrical distribution with a single mode – even though the distribution may be non-normal. It is therefore advisable to work with sample sizes greater than 100 whenever possible. An alternative test (called the *Lilliefors test*) of the null hypothesis of normality, with unspecified mean and standard deviation, can also be used. This is a nonparametric test and is not covered in this book. Although it is somewhat more powerful than the chi-squared test, the Lilliefors test requires that you work with the individual sample observations. If you only have access to grouped data – as might be the case if the data have been obtained from a secondary source – you must resort to the chi-squared test for normality. The final point is that the procedure described in this section can also

be used to test the fit of other distributions, such as the binomial and Poisson distributions. We have singled out the normal distribution for attention because of its importance.

EXERCISES

Learning the techniques

- 14.39** Suppose that a random sample of 100 observations was drawn from a population. After calculating the mean and the standard deviation, each observation was standardised and the number of observations in each of the following intervals was counted. Can we infer at the 5% significance level that the data were not drawn from a normal population?

Interval	Frequency
$Z \leq -1.5$	10
$-1.5 < Z \leq -0.5$	18
$-0.5 < Z \leq 0.5$	48
$0.5 < Z \leq 1.5$	16
$Z > 1.5$	8

- 14.40** A random sample of 50 observations yielded the following frequencies for the standardised intervals.

Interval	Frequency
$Z \leq -1$	6
$-1 < Z \leq 0$	27
$0 < Z \leq 1$	14
$Z > 1$	3

Can we infer that the data are not normal? (Use $\alpha = 0.10$.)

- 14.41 XR14-41** Test the hypothesis that the following sample of data is drawn from a normal population. The sample mean and the standard deviation are 34 and 12, respectively. (Use $\alpha = 0.01$.)

Class	Frequency
10 up to 20	15
20 up to 30	24
30 up to 40	30
40 up to 50	18
50 up to 60	13

- 14.42** Determine the rejection regions for tests of normality having the following null hypotheses (assuming that there are seven categories and $\alpha = 0.05$).

- a The data are normally distributed.
- b The data are normally distributed, with a mean of 25.
- c The data are normally distributed, with a mean of 25 and a standard deviation of 8.

- 14.43 XR14-43** Test the hypothesis that the following sample of data is drawn from a normal population. (Use $\alpha = 0.10$.)

Class	Frequency
-20 up to -10	8
-10 up to 0	21
0 up to 10	43
10 up to 20	48
20 up to 30	25
30 up to 40	15

Applying the techniques

- 14.44 XR14-44 Self-correcting exercise.** A common measure of a firm's liquidity is its *current ratio*, defined as its current assets divided by its current liabilities. A relatively high current ratio provides some evidence that a firm can meet its short-term financial obligations. The current ratios for a sample of 200 firms are recorded in the following table. Is there evidence at the 10% level of significance that this sample was drawn from a normal population?

Current ratio	Frequency
0 up to 1.0	20
1.0 up to 1.5	33
1.5 up to 2.0	47
2.0 up to 2.5	40
2.5 up to 3.0	31
3.0 up to 4.0	29

- 14.45 XR14-45** The instructors for an introductory accounting course attempt to construct the final examination so that the marks are normally distributed, with a mean of 65. From the sample of marks appearing in the accompanying frequency

distribution table, can you conclude that they have achieved their objective? (Use $\alpha = 0.05$.)

Mark	Frequency
30 up to 40	4
40 up to 50	17
50 up to 60	29
60 up to 70	49
70 up to 80	33
80 up to 90	18

Computer applications

The following exercises require the use of a computer and software.

- 14.46** In Exercise 11.31, you estimated the mean matched pairs difference. The estimation depends on the requirement that the differences are normally distributed. Test with a 10% significance level to determine whether the requirement is violated.

14.47 Exercise 12.16 requires that the income of blue-collar workers is normally distributed. Conduct a test with $\alpha = 0.05$ to determine whether the required condition is unsatisfied.

14.48 Exercise 13.31 required you to conduct a *t*-test of the difference between two means. The amount of time wasted in both successful and unsuccessful firms is required to be normally distributed. Is the required condition violated? Test with $\alpha = 0.05$.

14.49 Exercise 13.35 asked you to conduct a *t*-test of the difference between two means (reaction times). Test to determine whether there is enough evidence to infer that the reaction times are not normally distributed. A 5% significance level is judged to be suitable.

14.4 Summary of tests on nominal data

At this point in the textbook, we have described four tests used when the data are nominal:

- 1 *z*-test of p (Section 12.6)
- 2 *z*-test of $p_1 - p_2$ (Section 13.3)
- 3 Chi-squared test of goodness of fit (a multinomial experiment) (Section 14.1)
- 4 Chi-squared test of a contingency table (Section 14.2)

In the process of presenting these techniques, it was necessary to concentrate on one technique at a time and to focus on the kinds of problems each addresses. However, this approach tends to conflict somewhat with our promised goal of emphasising the ‘when’ of statistical inference. In this section, we summarise the statistical tests on nominal data to ensure that you are capable of selecting the correct method.

14.4a Identifying the statistical technique

There are two critical factors in identifying the technique used when the data are nominal. The first, of course, is the problem objective. The second is the number of categories that the nominal variable can assume. **Table 14.1** provides a guide to help select the correct technique.

TABLE 14.1 Statistical techniques for nominal data

Problem objective	Number of categories	Statistical technique
Describe a single population	2	<i>z</i> -test of p or the chi-squared test of multinomial experiment
Describe a single population	2 or more	Chi-squared test of a multinomial experiment
Compare two populations	2	<i>z</i> -test of $p_1 - p_2$ or the chi-squared test of a contingency table
Compare two populations	2 or more	Chi-squared test of a contingency table
Compare two or more populations	2 or more	Chi-squared test of a contingency table
Analyse the relationship between two variables	2 or more	Chi-squared test of a contingency table

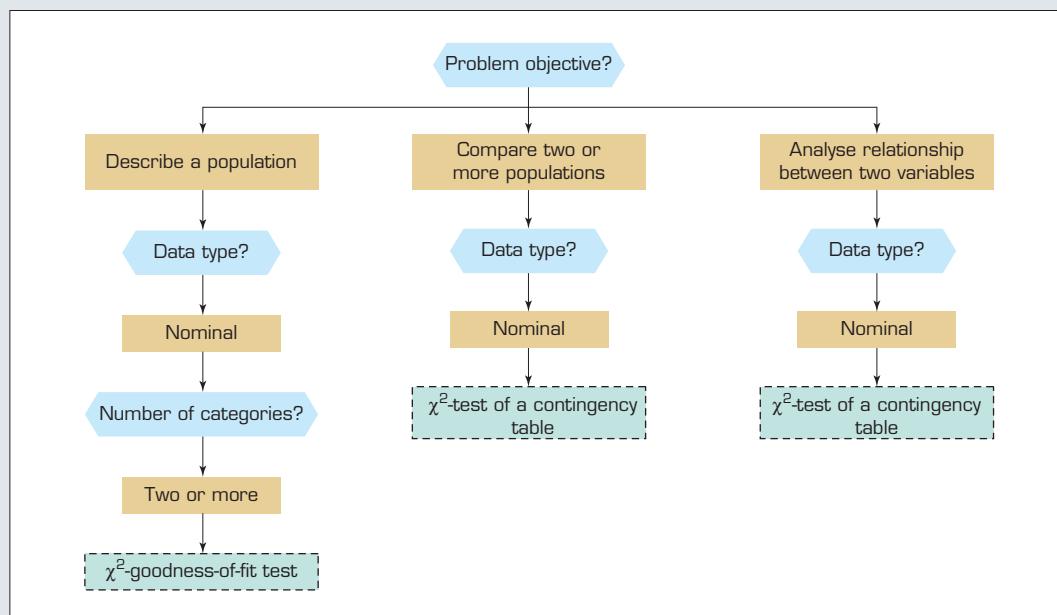
14.4b Developing an understanding of statistical concepts

Table 14.1 summarises how we deal with nominal data. We determine the frequency of each category and use these frequencies to calculate test statistics. We can then calculate proportions to calculate z -statistics or use the frequencies to calculate χ^2 -statistics. Because squaring a standard normal random variable produces a chi-squared variable, we can employ either statistic to test for difference. As a consequence, when you encounter nominal data in the problems described in this book (and other introductory applied statistics books), the most logical starting point in selecting the appropriate technique will be either a z -statistic or a χ^2 -statistic. However, you should know that there are other statistical procedures that can be applied to nominal data, techniques that are not included in this book.

Study Tools

CHAPTER SUMMARY

This chapter described three statistical techniques. The first is the *chi-squared goodness-of-fit test* (also called *test of a multinomial experiment*), which is applied when the problem objective is to describe a single population of nominal data with two or more categories. The second is the *chi-squared test of a contingency table*. This test has two objectives: to analyse the relationship between two nominal variables and to compare two or more populations of nominal data. The last procedure is designed to use the goodness-of-fit test to determine whether a sample was drawn from a normal population. In the Appendix, a goodness-of-fit test is provided to determine whether a sample was drawn from a Poisson population.



The data files for Examples, Exercises and Cases are provided on the companion website (accessible through <https://login.cengagebrain.com/>).

SYMBOLS

Symbol	Pronounced	Represents
o_i	o -sub- i	Observed frequency of the i th category
e_i	e -sub- i	Expected frequency of the i th category
χ^2	<i>Chi-squared</i>	Test statistic

SUMMARY OF FORMULAS

Expected frequency	$e_i = np_i$
Chi-squared test statistic (for all procedures)	$\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$
Expected frequencies for a contingency table	$e_{ij} = \frac{(\text{column } j \text{ total})(\text{row } i \text{ total})}{\text{sample size}}$

SUPPLEMENTARY EXERCISES

14.50 XR14-50 An organisation dedicated to ensuring fairness in television game shows is investigating *Wheel of Fortune*. In this show, three contestants are required to solve puzzles by selecting letters. Each contestant selects a letter and continues selecting until he or she chooses a letter that is not in the hidden word, phrase or name. The order of contestants is random. However, contestant 1 gets to start game 1, contestant 2 starts game 2, and so on. The contestant who wins the most money is declared the winner, and he or she is given an opportunity to win a grand prize. Usually, more than three games are played per show, and as a result it appears that contestant 1 has an advantage; contestant 1 will start two games, whereas contestant 3 will usually start only one game. To see whether this is the case, a random sample of 30 shows was taken and the starting position of the winning contestant for each show was recorded. These are shown in the following table.

Starting position	Number of winners
1	14
2	10
3	6

Do the tabulated results allow us to conclude with $\alpha = 0.10$ that the game is unfair?

14.51 XR14-51 Consider a multinomial experiment involving $n = 200$ trials and $k = 4$ cells. The observed frequencies resulting from the experiment are shown in the accompanying table, and the null hypothesis to be tested is

$$H_0: p_1 = 0.4, p_2 = 0.3, p_3 = 0.2, p_4 = 0.1$$

Cell	1	2	3	4
Frequency	96	54	28	22

- a State the alternative hypothesis.
- b Test the hypotheses, using $\alpha = 0.05$.

14.52 XR14-52 It has been estimated that employee absenteeism costs Australian companies billions of dollars per year. As a first step in addressing the rising cost of absenteeism, the personnel department of a large corporation recorded the week days on which individuals in a sample of 362 absentees were away over the past

several months. Do these data suggest that absenteeism is higher on some days of the week than on others? (Use $\alpha = 0.05$.)

Day of the week	Mon	Tues	Wed	Thurs	Fri
Number absent	87	62	71	68	74

14.53 XR14-53 Suppose that the personnel department in Exercise 14.52 continued its investigation by categorising absentees according to the shift on which they worked, as shown in the following table. Is there evidence of a relationship between the days on which employees were absent and the shift on which the employees worked? (Use $\alpha = 0.10$.)

Shift	Mon	Tues	Wed	Thurs	Fri
Day	52	28	37	31	33
Evening	35	34	34	37	41

14.54 XR14-54 A relationship has long been suspected between smoking and susceptibility to having a stroke. Strong statistical evidence on this issue has been lacking; however, recently a research paper was published based on the participation of 2000 residents who were studied over a 26-year period. Because of the large number of individuals observed, researchers were able to maintain reasonable sample sizes even after segmenting the observed group to control for factors other than smoking that might influence the individual susceptibility to a stroke, such as gender and blood-pressure level. The results for men with low-blood pressure levels are shown in the following table. What would you conclude at the 5% level of significance?

Individual	Stroke	No stroke
Smoker	37	183
Non-smoker	21	274

14.55 XR14-55 The use of credit cards as a source of consumer credit has become increasingly prevalent over the last few decades. A recent study attempted to profile holders of VISA and MasterCard. From the data shown in the following table, would you conclude that there is a relationship between gender and the credit card that is held? (Test using $\alpha = 0.05$.)

Gender	Credit card held		
	VISA only	MC only	Both cards
Male	128	66	137
Female	295	165	287

14.56 XR14-56 The study mentioned in Exercise 14.55 also investigated the payment habits of credit card users. From the data given below, would you conclude that there is a relationship between the credit card that is held and the amount that is paid monthly? (Test using $\alpha = 0.05$.)

Amount paid monthly	Credit card held		
	VISA only	MC only	Both cards
In full	204	99	148
Min. amount due	55	37	81
More than min. but not in full	148	85	174

14.57 XR14-57 An industrial relations expert on academic and non-academic staff has been studying the relationship between gender reporting structures in the workplace and the level of employee job satisfaction. The results of a recent survey in an institution are shown below. Using $\alpha = 0.10$, conduct a test to determine whether the level of job satisfaction depends on the boss–employee gender relationship.

Level of satisfaction	Supervisor/employee				Total
	Female/female	Female/male	Male/male	Male/female	
Satisfied	20	25	50	75	170
Neutral	40	50	50	40	180
Dissatisfied	30	45	10	15	100
Total	90	120	110	130	450

14.58 XR14-58 A discount electrical goods shop sells televisions, refrigerators and washing machines. The store's manager wishes to determine whether there is a relationship between the method of payment and the item purchased. The relevant data on last month's purchases are shown in the following table. Using $\alpha = 0.10$, test to see if the data in the table suggest such a relationship.

Payment method	TV	Fridges	Washing machines
Cash	11	4	7
Credit card	52	19	12
Instalment	27	32	11

14.59 XR14-59 The proportion of a company's earnings paid out to its shareholders in the form of dividends is called the company's dividend payout ratio. A frequency distribution of dividend payout ratios (expressed as percentages) for a sample of 125 Melbourne companies is shown in the following table. Ten of these companies paid no dividend.

Dividend payout ratio (%)	Frequency
0 up to 10	13
10 up to 20	7
20 up to 30	10
30 up to 40	23
40 up to 50	28
50 up to 60	21
60 up to 70	14
70 up to 80	5
80 up to 90	3
90 up to 100	1

- a Construct a histogram for these data.
- b Using $\alpha = 0.10$, test the hypothesis that dividend payout ratios are normally distributed.
- c Repeat part (b), considering only companies that paid a dividend.

14.60 XR14-60 The management of a large pension fund in Melbourne is interested in studying the distribution of monthly rates of return on large, well-diversified portfolios of common stock. Intensive gathering of data has yielded 90 monthly returns on various large portfolios. These returns are summarised in the following frequency distribution. Is there sufficient evidence to allow management to conclude that the monthly rates of return are not normally distributed? (Use $\alpha = 0.05$.)

Return (%)	Frequency
-4 up to -3	4
-3 up to -2	4
-2 up to -1	10
-1 up to 0	13
0 up to 1	16
1 up to 2	15
2 up to 3	13
3 up to 4	8
4 up to 5	5
5 up to 6	2

Computer applications

The following exercises require the use of a computer and software.

- 14.61 XR14-61** Given the high cost of medical care, research that highlights how to avoid illness is welcome. Previously performed research has revealed that stress negatively affects the immune system. Two scientists at a medical centre in Victoria asked 114 healthy adults about their social circles; they were asked to list every group they had contact with at least once every two weeks – family, co-workers, neighbours, friends, religious and community groups. Participants also reported negative life events over the past year, such as death of a relative or friend, divorce or job-related problems. In addition, whether each person contracted a cold over the next 12 weeks was recorded (1 = cold, 2 = no cold). The participants were divided into four groups:

- Group 1: Highly social and highly stressed
- Group 2: Not highly social and highly stressed
- Group 3: Highly social and not highly stressed
- Group 4: Not highly social and not highly stressed

Can we infer that differences exist between the four groups in terms of contracting a cold?

- 14.62 XR14-62** How does dieting affect the brain? This question was addressed by researchers in Australia. The experiment used 40 middle-aged women in Adelaide; half were on a diet and half were not. The mental arithmetic part of the experiment required the participants to add two three-digit numbers. The amount of time taken to solve the 48 problems was recorded. The participants were given another test that required them to repeat a string of five letters they had been told 10 seconds earlier. They were asked to repeat the test with five words told to them 10 seconds earlier. The data were recorded in the following way:

- Column 1: Identification number
- Column 2: 1 = dieting, 2 = not dieting
- Column 3: Time to solve 48 problems (seconds)
- Column 4: Repeat string of 5 letters (1 = no, 2 = yes)

Column 5: Repeat string of 5 words (1 = no, 2 = yes)

Is there sufficient evidence to infer that dieting adversely affects the brain?

- 14.63 XR14-63** Mutual funds are a popular way of investing in the stock market. A financial analyst wanted to determine the effect income had on ownership of mutual funds and whether the relationship had changed from four years earlier. She took a random sample of adults 25 years of age and older and asked each person whether he or she owned mutual funds (No = 1 and Yes = 2) and to report the annual household income. The categories are

- 1 Less than \$25000
- 2 \$25000 to \$34999
- 3 \$35000 to \$49999
- 4 \$50000 to \$74999
- 5 \$75000 to \$100000
- 6 More than \$100000

Can we infer from the data that household income and ownership of mutual funds are related?

- 14.64 XR14-64** A survey of the business school graduates undertaken by a university placement office asked, among other questions, their gender and in which area each person was employed. The respondents reported (among other questions) gender (1 = Female, 2 = Male) and area of employment (1 = Accounting, 2 = Finance, 3 = General management, 4 = Marketing/sales, 5 = Other). Can we infer from the data that female and male graduates differ in their areas of employment?

- 14.65 XR14-65** Are you more likely to smoke if your parents smoke? To shed light on the issue, a sample of 20- to 40-year-old people was asked whether they smoked and whether their parents smoked. The results are stored the following way:

- Smoke: 1 = Do not smoke, 2 = Smoke
- Parent: 1 = Neither parent smoked, 2 = Father smoked, 3 = Mother smoked, 4 = Both parents smoked

Test to determine whether there is enough evidence to infer that parents' smoking and their children smoking are related.

Case Studies

CASE 14.1 Gold lotto

C14-01 Gold lotto is a national lottery that operates as follows. Players select eight different numbers (six primary and two supplementary numbers) from 1 to 45. Once a week, the corporation that runs the lottery selects eight numbers (six primary and two supplementary numbers) at random from 1 to 45. Winners are determined by how many numbers on their tickets agree with the numbers drawn. In selecting their numbers, players often look at past patterns as a way to help predict future drawings. A regular feature that appears in the newspaper identifies the number of times each number has occurred since draw 413 (Saturday 6 July 1985). The data recorded in the following table appeared in the *Saturday Gold Lotto* website after the completion of draw 4047 (2 May 2020). What would you recommend to anyone who believes that past patterns of lottery numbers are useful in predicting future drawings?

Drawing frequency of lotto numbers since draw 413 (as at 2 May 2020)

Lotto number	Number of times drawn	Lotto number	Number of times drawn	Lotto number	Number of times drawn
1	361	16	324	31	318
2	312	17	301	32	320
.
14	301	29	320	44	290
15	333	30	301	45	309

<https://www.thelott.com/saturday-gold-lotto/results>

CASE 14.2 Exit polls

C14-02 After the polls close on election day, television networks compete to be the first to predict which candidate will win. Their predictions are based on counts taken in exit polls at a number of polling booths in marginal seats. Exit polls are conducted by asking random samples of voters who have just exited from the polling booth which candidate they voted for. In addition, pollsters ask a variety of other questions that provide information to politicians, journalists and others. The responses to questions relating to gender, age, education and the party for which his/her vote was cast (Labor or Coalition) in a recent federal election were collected and recorded. Based on the data, analyse the relationship between the party vote cast and the variables gender, education and income.

CASE 14.3 How well is the Australian Government managing the coronavirus pandemic?

C14-03 The coronavirus pandemic has affected more than 300 countries around the world within 3 months of its initial detection. Governments of different countries have introduced various measures to control the spread of the virus. The Lowy Institute in Australia, which conducted a COVIDpoll among Australians on 14 May 2020, found that 'almost all Australians say Australia has handled COVID-19 well'. The survey which asked about Australians' view on this issue considered 6 countries and the results are summarised in the table below. Analyse the relationship between the level of satisfaction and the country.

Country	Level of satisfaction				
	Very well	Fairly well	Fairly badly	Very badly	Don't know
Australia	43	50	6	1	0
Singapore	23	56	15	3	3
China	6	25	25	44	0
United Kingdom	3	27	49	21	0
Italy	2	13	44	40	1
United States	2	8	27	63	0

Source: <https://www.lowyinstitute.org/publications/covidpoll-lowy-institute-polling-australian-attitudes-coronavirus-pandemic#sec42571>.

Appendix 14.A

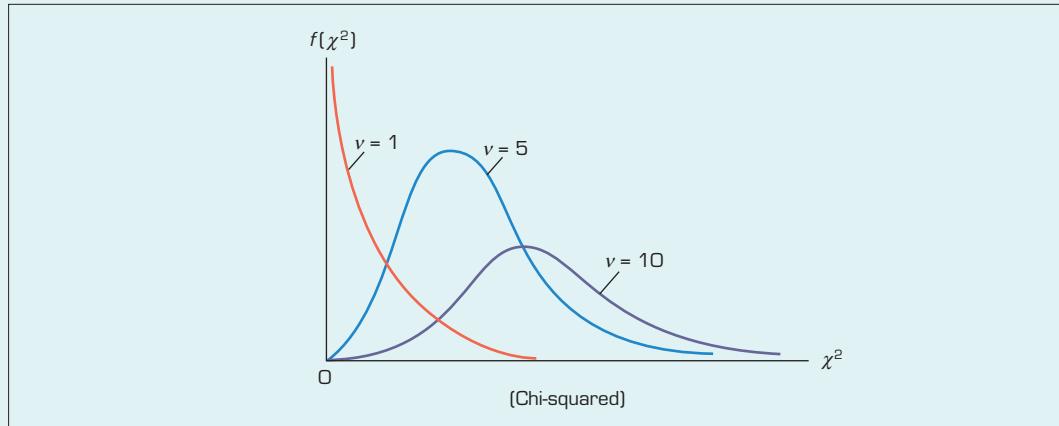
Chi-squared distribution

chi-squared distribution

A continuous nonsymmetric distribution used in statistical inference.

The **chi-squared distribution** is positively skewed and ranges between 0 and ∞ . Like that of the t -distribution, its shape depends on its number of degrees of freedom. **Figure A14.1** depicts several chi-squared distributions that have different degrees of freedom ($v = 1, 5$ and 10).

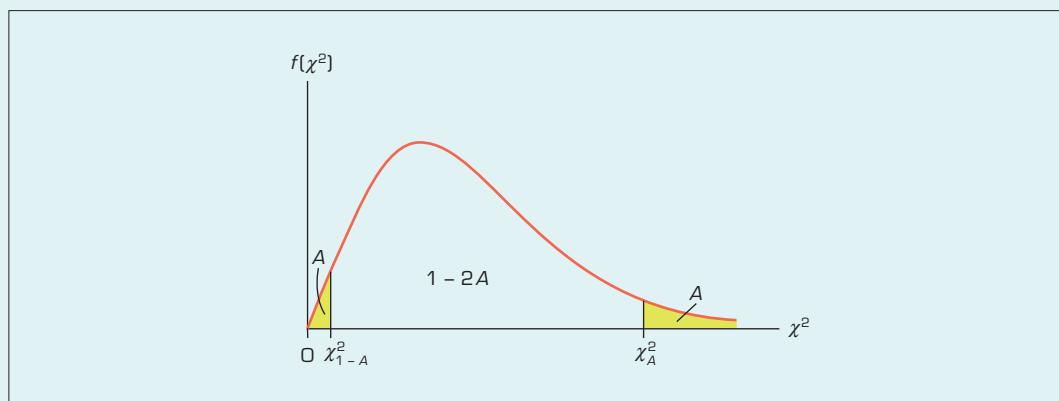
FIGURE A14.1 Chi-squared distributions



14.Aa Determining the chi-squared values manually

The value of χ^2 such that the area to its right under the chi-squared curve is equal to A , is denoted χ_A^2 . Because the χ^2 variable assumes only positive values, we need a notation that is different from the one used for z and t , to define the point for which the area to its left is equal to A . We therefore define χ_{1-A}^2 as the point for which the area to its right is $(1 - A)$ and the area to its left is A . **Figure A14.2** describes this notation. Table 5 in Appendix B provides the values of $\chi_{A,v}^2$ and $\chi_{1-A,v}^2$ for various values of A and various degrees of freedom v . **Table A14.1** is a partial reproduction of this table. To illustrate, $\chi_{0.05,10}^2 = 18.3070$ and $\chi_{0.95,10}^2 = 3.94030$.

FIGURE A14.2 χ_A^2 and χ_{1-A}^2



For values of degrees of freedom greater than 100, the chi-squared distribution can be approximated by a normal distribution with $\mu = v$ and $\sigma = \sqrt{2v}$.

TABLE A14.1 Reproduction of part of Table 5 in Appendix B: Critical values of χ^2

Degrees of freedom	$\chi^2_{0.995}$	$\chi^2_{0.990}$	$\chi^2_{0.975}$	$\chi^2_{0.950}$	$\chi^2_{0.050}$	$\chi^2_{0.025}$	$\chi^2_{0.010}$	$\chi^2_{0.005}$
1	0.0000393	0.0001571	0.0009821	0.0039321	3.84146	5.02389	6.63490	7.87944
2	0.0100251	0.0201007	0.0506356	0.102587	5.99147	7.37776	9.21034	10.5966
3	0.0717212	0.114832	0.215795	0.351846	7.81473	9.34840	11.3449	12.8381
4	0.206990	0.297110	0.484419	0.710721	9.48773	11.1433	13.2767	14.8602
5	0.411740	0.554300	0.831211	1.145476	11.0705	12.8325	15.0863	16.7496
6	0.675727	0.872085	1.237347	1.63539	12.5916	14.4494	16.8119	18.5476
7	0.989265	1.239043	1.68987	2.16735	14.0671	16.0128	18.4753	20.2777
8	1.344419	1.646482	2.17973	2.73264	15.5073	17.5346	20.0902	21.9550
9	1.734926	2.087912	2.70039	3.32511	16.9190	19.0228	21.6660	23.5893
10	2.15585	2.55821	3.24697	3.94030	18.3070	20.4831	23.2093	25.1882
11	2.60321	3.05347	3.81575	4.57481	19.6751	21.9200	24.7250	26.7569
12	3.07382	3.57056	4.40379	5.22603	21.0261	23.3367	26.2170	28.2995

14.Ab Determining the chi-squared values using the computer

To calculate the probability to the right of any chi-squared value, proceed as follows:

COMMANDS

- 1 Click **FORMULAS, fx, All** from the categories dropdown menu and select the **CHIDIST** function. Click **OK**.
- 2 Type the value of x (X) and the degrees of freedom (Deg _freedom).

Alternatively, type into any cell

= CHIDIST([X],[Degrees of freedom]).

For example, CHIDIST(7.81473,3) = 0.05.

To determine a value of a **chi-squared random variable**, follow these instructions:

chi-squared random variable

A random variable that is chi-squared distributed.

COMMANDS

- 1 Click **FORMULAS, fx, All** from the categories dropdown menu and select the **CHIINV** function. Click **OK**.
- 2 Type the cumulative probability (Probability) and the degrees of freedom (Deg _freedom).

Alternatively, type into any cell

= CHIINV([Probability],[Degrees of freedom]).

For example, CHIINV(0.025,5) = 12.8325.

Simple linear regression and correlation

Learning objectives

In this chapter we present an analysis of the relationship between two variables. The data types considered are numerical and ordinal.

At the completion of this chapter, you should be able to:

- L01** identify the dependent and the independent variables
- L02** use the least squares method to derive estimators of simple linear regression model parameters
- L03** understand the required conditions to perform statistical inferences about a linear regression model
- L04** test the significance of the regression model parameters
- L05** calculate measures used to assess the performance of a regression model
- L06** use the regression equation for prediction
- L07** calculate the prediction interval of the dependent variable
- L08** calculate the coefficient of correlation between two variables and assess the strength of the relationship
- L09** detect violations of required conditions using diagnostic checks on the regression model results.

CHAPTER OUTLINE

- Introduction
- 15.1 Model**
- 15.2 Estimating the coefficients**
- 15.3 Error variable: Required conditions**
- 15.4 Assessing the model**
- 15.5 Using the regression equation**
- 15.6 Testing the coefficient of correlation**
- 15.7 Regression diagnostics – I**

SPOTLIGHT ON STATISTICS

Would increasing tax on alcoholic beverages reduce consumption?

For many reasons, consumption of excessive alcoholic beverages and its effects continues to attract a lot of media attention. Governments, the health profession and social workers are searching for ways to reduce excessive alcohol consumption to reduce the level of harm to society as a whole, and the cost to the health system. The Australian Federal Government's National Alcohol Strategy 2019–2028 was released as a national framework to prevent and minimise alcohol-related harms among individuals, families and communities. One suggestion coming from many circles is to increase the price of alcoholic beverages via taxation,



Source: iStock.com/Ilyabolotov

► which is expected to reduce consumption. You are asked by a social worker to investigate the relationship between the prices and consumption of beer and wine so that a tax policy on these alcoholic beverages can be developed. Data for per adult consumption and the relative price of beer and wine for Australia for the years 1988–2017 are stored in **CH15\XM15-00**. Analyse the relationship between consumption and relative price of beer and wine. We answer this question on pages 655–6.

Introduction

This chapter is the first in a series of three in which the problem objective is to analyse the relationship between numerical variables. **Regression analysis** is used to predict the value of one variable on the basis of other variables. This technique may be the most commonly used statistical procedure because, as you can easily appreciate, almost all companies and government institutions forecast variables, such as energy demand, interest rates, inflation rates, prices of raw materials and labour costs.

The technique involves developing a mathematical equation or model that describes the relationship between the variable to be forecast, which is called the *dependent variable*, and the related variables, which are called *independent variables*. The dependent variable is denoted y , while the related independent variables are denoted x_1, x_2, \dots, x_k , where k is the number of independent variables.

If we are only interested in determining whether a relationship exists, we employ correlation analysis. We have already introduced this technique. In Chapter 4, we presented the graphical method to describe the association between two numerical variables – the *scatter diagram*. We introduced the regression models, coefficient of correlation and covariance in Chapter 5.

Because regression analysis involves a number of new techniques and concepts, we divide the presentation into three chapters. In this chapter, we present techniques that allow us to determine the relationship between only two variables. In Chapter 16, we expand our discussion to more than two variables.

Here are three illustrations of regression analysis.

Illustration 1 The product manager in charge of an Australian beer company would like to predict the demand for beer (y) during the next year. In order to use regression analysis, the manager lists the following variables as likely to affect sales:

- Price of the company beer (x_1)
- Number of Australian adults above 15 years of age (the target market) (x_2)
- Price of competitors' beers (x_3)
- Price of wine (x_4)
- Price of spirits (x_5)
- Effectiveness of advertising (as measured by advertising exposure) (x_6)
- Annual sales in previous years (x_7)
- Consumers' income (x_8)

Illustration 2 A gold speculator is considering a major purchase of gold bullion. He would like to forecast the price of gold (y) two years from now (his planning horizon), using regression analysis. In preparation, he produces the following list of independent variables:

- Interest rates (x_1)
- Inflation rate (x_2)
- Price of oil (x_3)
- Demand for gold jewellery (x_4)
- Demand for industrial and commercial gold (x_5)
- Price of gold (x_6)

regression analysis

A technique that estimates the relationship between variables and aids forecasting.

Illustration 3 A real estate agent wants to more accurately predict the selling price of houses (y). She believes that the following variables affect the price of a house:

- Size of the house (number of squares) (x_1)
- Number of bedrooms (x_2)
- Frontage of the block (x_3)
- Size of the block (number of square metres) (x_4)
- Condition of the house (x_5)
- Age of the house (x_6)
- Location (x_7)
- House has a pool (x_8)

In each of these illustrations, the primary motive for using regression analysis is forecasting. Nonetheless, analysing the relationship between variables can also be quite useful in managerial decision making. For instance, in the first application, the product manager may want to know how price is related to product demand, so that a decision about a prospective change in pricing can be made.

Regardless of why regression analysis is performed, the next step in the technique is to develop a mathematical equation or model that accurately describes the nature of the relationship that exists between the dependent variable and the independent variables. This stage – which is only a small part of the total process – is described in the next section. In the ensuing sections of this chapter (and in Chapter 16), we will spend considerable time assessing and testing how well the model fits the actual data. Only when we are satisfied with the model do we use it to estimate and forecast.

15.1 Model

The job of developing a mathematical equation can be quite complex, because we need to have some idea about the nature of the relationship between each of the independent variables and the dependent variable. For example, the gold speculator mentioned in illustration 2 needs to know how interest rates affect the price of gold. If he proposes a linear relationship, that may imply that as interest rates rise (or fall), the price of gold will rise or fall. A quadratic relationship may suggest that the price of gold will increase over a certain range of interest rates but will decrease over a different range. Perhaps certain combinations of values of interest rates and other independent variables influence the price in one way, while other combinations change it in other ways. The number of different mathematical models that could be proposed is almost infinite.

15.1a Deterministic and probabilistic models

In business subjects, you might have seen the following equations:

$$\text{Profit} = \text{Revenue} - \text{Costs}$$

$$\text{Total cost} = \text{Fixed cost} + (\text{Unit variable cost} \times \text{Number of units produced})$$

deterministic model

An equation in which the value of the dependent variable is completely determined by the values of the independent variable(s).

probabilistic model

A model that contains a random term.

The above are examples of **deterministic models**, so named because – except for small measurement errors – such equations allow us to determine the value of the dependent variable (on the left side of the equation) from the value of the independent variables. In many practical applications of interest to us, deterministic models are unrealistic. For example, is it reasonable to believe that we can determine the selling price of a house solely on the basis of its size? Unquestionably, the size of a house affects its price, but many other variables (some of which may not be measurable) also influence price. What must be included in most practical models is a method to represent the randomness that is part of a real-life process. Such a model is called a **probabilistic model**.

To create a probabilistic model, we start with a deterministic model that approximates the relationship we want to model. We then add a random term that measures the error of the

deterministic component. Suppose that in illustration 3 described above, the real estate agent knows that the cost of building a new house (including the builder's profit) is about \$6000 per square and that most lots (land) sell for about \$210000. The approximate selling price would be

$$y = 210000 + 6000x$$

where y = selling price and x = size of the house in squares. A house of 25 squares would therefore be estimated to sell for

$$y = 210000 + 6000 \times 25 = 360000$$

We know, however, that the selling price of a 25-square house is not likely to be exactly \$360000. Prices may actually range from \$340000 to \$385000. In other words, the deterministic model is not really suitable. To represent this situation properly, we should use the *probabilistic model* (also known as a *stochastic model*):

$$y = 210000 + 6000x + \varepsilon$$

where ε (the Greek letter *epsilon*) represents the random term, also called the *error variable* – the difference between the actual selling price and the estimated price based on the size of the house. The error thus accounts for all the variables, measurable and immeasurable, that are not part of the model. The value of ε will vary from one sale to the next, even if x remains constant. That is, houses of exactly the same size will sell for different prices because of differences in location, selling season, decorations and other variables.

15.1b Simple linear regression model

In the three chapters devoted to regression analysis, we will present only probabilistic models. Additionally, to simplify the presentation, all models will be linear. In this chapter, we restrict the number of independent variables to one. The model to be used in this chapter is called the first **simple linear regression model**,¹ or the *first-order linear model*.

simple linear regression model

Also called the first-order linear model, this is a linear regression equation with only one independent variable.

Simple linear regression model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where

y = dependent variable

x = independent variable

β_0 = y -intercept

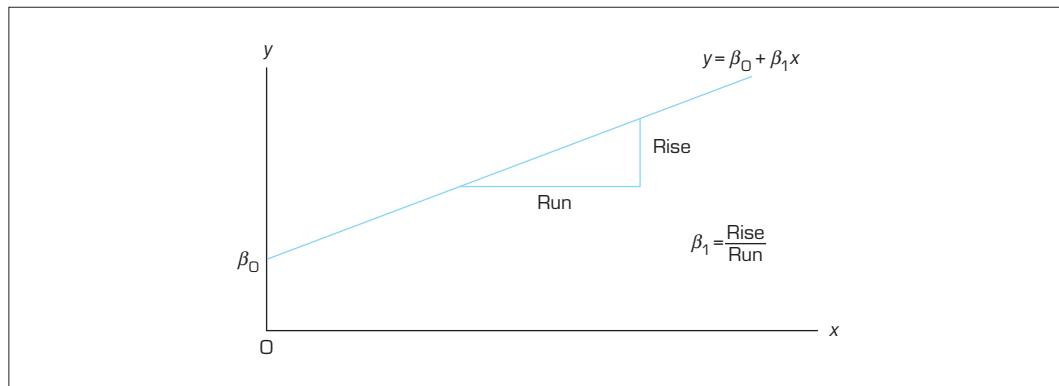
β_1 = slope of the line (defined as the ratio rise/run, or the change in y /change in x)

ε = error variable

¹ We use the term *linear* in two ways. The 'linear' in linear regression refers to the form of the model in which the terms form a linear combination of the coefficients β_0 and β_1 . Thus, for example, the model $y = \beta_0 + \beta_1 x^2 + \varepsilon$ is a linear combination whereas $y = \beta_0 + \beta_1^2 x + \varepsilon$ is not. The simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$ describes a straight-line, or linear, relationship between the dependent variable and one independent variable. In this book, we use the linear regression technique only. Hence, when we use the word *linear* we will be referring to the straight-line relationship between the variables.

Figure 15.1 depicts the deterministic component of the model.

FIGURE 15.1 Simple linear model: Deterministic component



The problem objective addressed by the model is to analyse the relationship between two variables, x and y , both of which must be numerical. To define the relationship between x and y , we need to know the value of the coefficients of the linear model β_0 and β_1 . However, these coefficients are population parameters, which are almost always unknown. In the next section, we discuss how these parameters are estimated.

EXERCISES

- 15.1** Graph each of the following straight lines.

Identify the y -intercept and the slope.

- a $y = 2 + 3x$
- b $y = 5 - 2x$
- c $y = -2 + 4x$
- d $y = x$
- e $y = 4$

- 15.2** For each of the following data sets, plot the points on a graph. Draw a straight line through the data. Determine the y -intercept and the slope of the line you drew.

a	<table border="1"> <thead> <tr> <th>x</th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th></tr> </thead> <tbody> <tr> <td>y</td><td>4</td><td>8</td><td>12</td><td>16</td><td>20</td></tr> </tbody> </table>	x	1	2	3	4	5	y	4	8	12	16	20
x	1	2	3	4	5								
y	4	8	12	16	20								

b	<table border="1"> <thead> <tr> <th>x</th><th>1</th><th>3</th><th>5</th><th>4</th><th>7</th></tr> </thead> <tbody> <tr> <td>y</td><td>5</td><td>7</td><td>10</td><td>9</td><td>16</td></tr> </tbody> </table>	x	1	3	5	4	7	y	5	7	10	9	16
x	1	3	5	4	7								
y	5	7	10	9	16								
c	<table border="1"> <thead> <tr> <th>x</th> <th>7</th> <th>9</th> <th>2</th> <th>3</th> <th>6</th> </tr> </thead> <tbody> <tr> <td>y</td> <td>4</td> <td>1</td> <td>6</td> <td>10</td> <td>5</td> </tr> </tbody> </table>	x	7	9	2	3	6	y	4	1	6	10	5
x	7	9	2	3	6								
y	4	1	6	10	5								

- 15.3** Graph the following observations of x and y .

Draw a straight line through the data. Determine the y -intercept and the slope of the line you drew.

	<table border="1"> <thead> <tr> <th>x</th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th></tr> </thead> <tbody> <tr> <td>y</td><td>4</td><td>6</td><td>7</td><td>7</td><td>9</td><td>11</td></tr> </tbody> </table>	x	1	2	3	4	5	6	y	4	6	7	7	9	11
x	1	2	3	4	5	6									
y	4	6	7	7	9	11									

15.2 Estimating the coefficients

We estimate the parameters β_0 and β_1 in a way similar to the methods used to estimate all the other parameters discussed in this book. We draw a random sample from the populations of interest (parameters) and calculate the sample statistics we need. Because β_0 and β_1 represent the coefficients of a straight line, their estimators are based on drawing a straight line through the sample data. Let y_i be the observed value of the annual bonus (\$'000) and x_i be the number of years of service. Then the simple linear regression model can be used and the equation of the line of best possible fit written as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

where \hat{y}_i is the fitted value of y_i , and $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimators of the parameters β_0 and β_1 .

15.2a Least squares regression line

A least squares regression line can be obtained by minimising SSE, the *sum of squared differences* (e_i^2) between the fitted value \hat{y}_i and the corresponding observed value y_i ,

$$\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2$$

where $e_i = y_i - \hat{y}_i$. The technique that produces the best possible fitted line is called the *least squares method*. The line itself is called the least squares line, the fitted line or the regression line. The ‘hats’ on the coefficients remind us that they are estimators of the parameters β_0 and β_1 .

By using calculus, we can produce formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$.

Least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

$$s_{xy} = \frac{1}{n-1} \left[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right]$$

$$s_x^2 = \frac{1}{n-1} \left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]$$

$$\bar{y} = \frac{\sum y_i}{n} \quad \text{and} \quad \bar{x} = \frac{\sum x_i}{n}$$

The shortcut formulas for s_{xy} and s_x^2 are given below:

Alternative formulas for s_{xy} and s_x^2

$$s_{xy} = \frac{1}{n-1} \left[\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} \right] = \frac{1}{n-1} \left[\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n} \right]$$

$$s_x^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - n \bar{x}^2 \right] = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \right]$$

As you can see, to estimate the regression coefficients manually, we need to determine the following summations:

$$\sum x \quad \sum y \quad \sum x^2 \quad \sum xy$$

Although the calculations are straightforward, we would rarely compute the regression line manually because the work is time consuming. However, we illustrate the manual calculations for a very small sample.

EXAMPLE 15.1**Annual bonus and years of experience: Part 1**

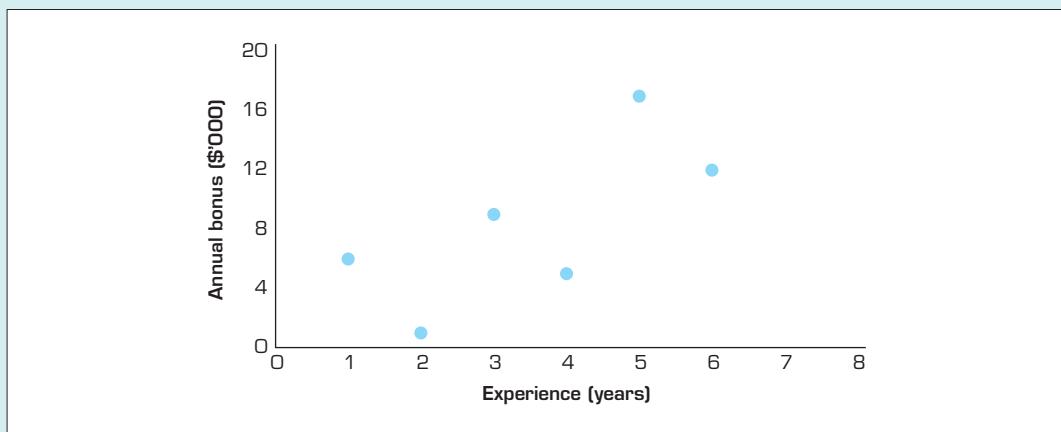
XM15-01 The annual bonuses (\$'000) of six employees with different years of experience were recorded as follows. We wish to determine the linear relationship between annual bonus and years of experience.

Years of experience, x	1	2	3	4	5	6
Annual bonus, y	6	1	9	5	17	12

Solution

As a first step we graph the data, as shown in **Figure 15.2**. Recall (from Chapter 4) that this graph is called a scatter diagram. The scatter diagram can be used to see visually whether a linear or non-linear relationship exists between the two variables. Evidently, in this case, a linear model is justified. Also, it is evident that the data support the expected belief that the two variables, years of experience and the annual bonus, are positively related. That is, annual bonus increases with increasing years of experience. In the next step, our task is to estimate a linear relationship between the annual bonus (dependent variable) and years of experience (the independent variable).

FIGURE 15.2 Scatter diagram for Example 15.1



The best fitted line based on the least squares method is given by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

where

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

To compute these two coefficients, we first find the required sums and means:

$$\sum x_i = 21$$

$$\sum y_i = 50$$

$$\sum x_i^2 = 91$$

$$\sum x_i y_i = 212$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{21}{6} = 3.5$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{50}{6} = 8.333$$



Using these summations in our shortcut formulas, we find s_{xy} and s_x^2 :

$$s_{xy} = \frac{1}{n-1} \left[\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n} \right] = \frac{1}{6-1} \left[212 - \frac{(21)(50)}{6} \right] = 7.4$$

and

$$s_x^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \right] = \frac{1}{6-1} \left[91 - \frac{(21)^2}{6} \right] = 3.5$$

The sample slope coefficient is calculated next:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{7.4}{3.5} = 2.114$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 8.333 - (2.114)(3.5) = 0.934$$

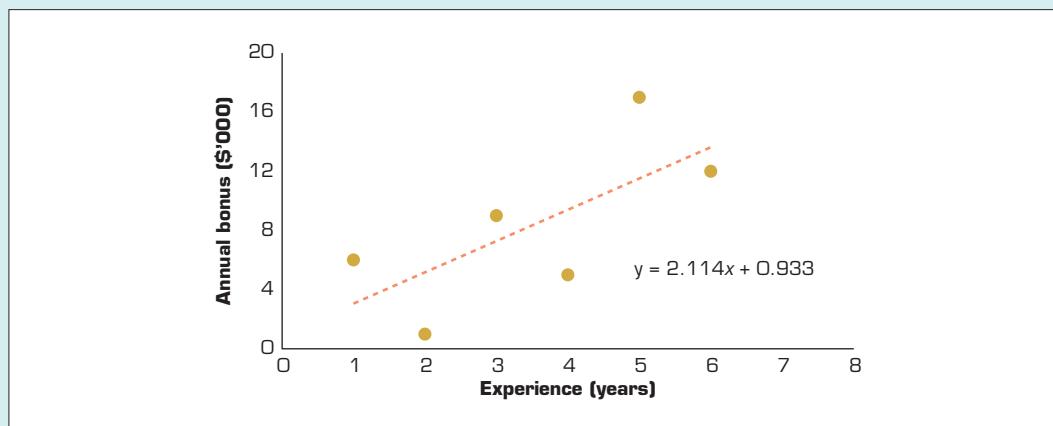
Thus, the least squares regression line can be written as

$$\hat{y} = 0.934 + 2.114x.$$

Interpretation: In the estimated regression line, $\hat{\beta}_0 = 0.934$ is the y -intercept and $\hat{\beta}_1 = 2.114$ is the slope of the regression line. This means that for every additional year of experience, annual income would increase, on average, by \$2.114 ('000), which is equivalent to \$2114.

Figure 15.3 depicts the least squares (or regression) line. As you can see, the line fits the data reasonably well.

FIGURE 15.3 Scatter diagram with regression line



15.2b Fitness of the regression line

We can measure how well the regression line fits the data by calculating the value of the minimised sum of squared differences. The differences between the points and the line are called errors or **residuals** and denoted e_i . That is,

$$e_i = y_i - \hat{y}_i$$

The residuals are the observed values of the error variable. Consequently, the minimised sum of squared differences is called the **sum of squares error**, denoted SSE.

residual

The difference between the predicted value of the dependent variable and its actual value.

sum of squares error (SSE)

The sum of squared residuals in a regression model.

Sum of squares for error

$$\text{SSE} = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

EXAMPLE 15.2

LO3

Annual bonus and years of experience: Part II

Calculate the SSE for Example 15.1.

Solution

The calculation of the residuals is shown in **Figure 15.4**. Notice that we calculate \hat{y}_i by substituting x_i into the formula for the regression line. The residuals e_i are the differences between the observed values y_i and the calculated or fitted values \hat{y}_i . Table 15.1 describes the calculation of SSE.

FIGURE 15.4 Calculation of the residuals in Example 15.1

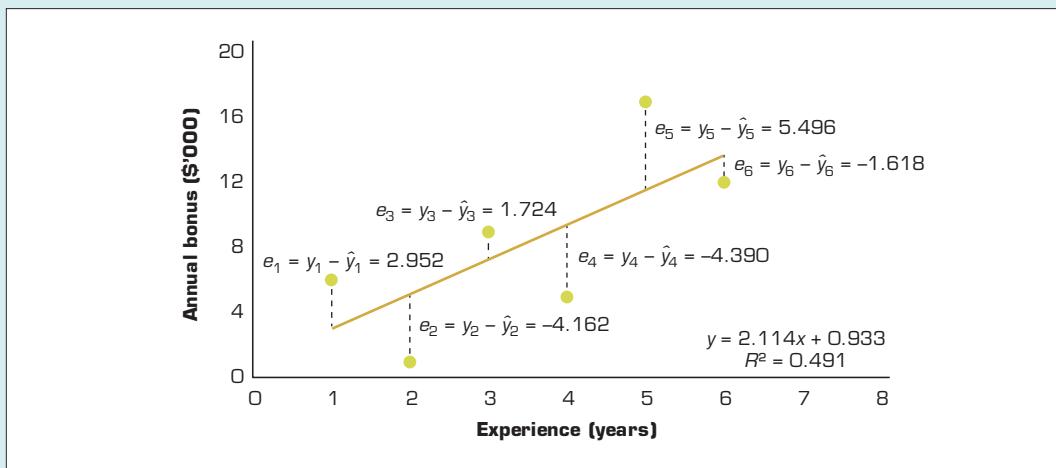


TABLE 15.1 Calculation of SSE for Example 15.1

i	x_i	y_i	$\hat{y}_i = 0.934 + 2.114x$	Residual $e_i = y_i - \hat{y}_i$	Residual squared $e_i^2 = (y_i - \hat{y}_i)^2$
1	1	6	3.048	2.952	8.714
2	2	1	5.162	-4.162	17.322
3	3	9	7.276	1.724	2.972
4	4	5	9.390	-4.390	19.272
5	5	17	11.504	5.496	30.206
6	6	12	13.618	-1.618	2.618
Sum					$\sum e_i^2 = 81.104$

Thus, $\text{SSE} = 81.104$. No other straight line will produce a sum of squared errors smaller than 81.104. In that sense, the regression line fits the data best. The sum of squares for error is an important statistic because it is the basis for other statistics that assess how well the linear model fits the data. We will introduce these statistics later in Section 15.4 of this chapter.

We now apply the technique to a more practical problem.

EXAMPLE 15.3

LO2

Odometer readings and prices of used cars: Part I

XM15-03 A critical factor for used-car buyers when determining the value of a car is how far the car has been driven. There is, however, not much information about this available in the public domain. To examine this issue, a used-car dealer randomly selected 100 five-year-old Ford Lasers that had been sold at auction during the past month. Each car was in top condition and equipped with automatic transmission, GPS and air conditioning. The dealer recorded the price and the number of kilometres on the odometer. An extract of the data is listed below. The dealer wants to find the regression line.

Car	Odometer reading ('000 km)	Auction selling price (\$'000)
1	37.4	16.0
2	44.8	15.2
3	45.8	15.0
.	.	.
.	.	.
.	.	.
100	36.4	15.4

Solution**Identifying the technique**

Notice that the problem objective is to analyse the relationship between two numerical variables. Because we want to know how the odometer reading affects the selling price, we identify the former as the independent variable, which we label x , and the latter as the dependent variable, which we label y .

Calculating manually

From the given data,

$$n = 100$$

$$\sum x_i = 3601.1 \quad \sum y_i = 1623.7$$

$$\bar{x} = 36.01 \quad \bar{y} = 16.24$$

$$\sum x_i^2 = 133\,986.6 \quad \sum y_i^2 = 26\,421.9 \quad \text{and} \quad \sum x_i y_i = 58\,067.4$$

Using these summations in our shortcut formulas, we find s_{xy} and s_x^2

$$s_{xy} = \frac{1}{n-1} \left[\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n} \right] = \frac{1}{99} \left(58\,067.4 - \frac{(3601.1)(1623.7)}{100} \right) = -4.077$$

and

$$s_x^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \right] = \frac{1}{99} \left(133\,986.6 - \frac{(3601.1)^2}{100} \right) = 43.509$$

Now we find the slope coefficient,

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{-4.077}{43.509} = -0.0937$$

and determine the y -intercept as follows:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 16.24 - (-0.0937)(36.01) = 19.61$$

The sample regression line is

$$\hat{y} = 19.61 - 0.0937x$$

Interpreting the coefficients

The slope coefficient $\hat{\beta}_1$ is -0.0937 , which means that for each additional kilometre on the odometer, the price decreases by an average of $\$0.0937$ (9.37 cents). In other words, for every additional 1000 km on the odometer the selling price decreases, on average, by $\$93.70$.

The y -intercept $\hat{\beta}_0$ is 19.61. Technically, the y -intercept is the point at which the regression line and the y -axis intersect. This means that when $x = 0$ (i.e. the car was not driven at all) the selling price is $\$19.61$ ($\times 1000$) = $\$19610$. We might be tempted to interpret this number as the price of cars that have not been driven. However, in this case, the intercept is probably meaningless. Because our sample did not include any cars with zero kilometres on the odometer, we have no basis for interpreting $\hat{\beta}_0$. As a general rule, we cannot determine the value of y for a value of x that is far outside the range of the sample values of x . In this example, the smallest and largest values of x are 19.1 and 49.2 respectively. Because $x = 0$ is not in this interval, we cannot safely interpret the value of y when $x = 0$.

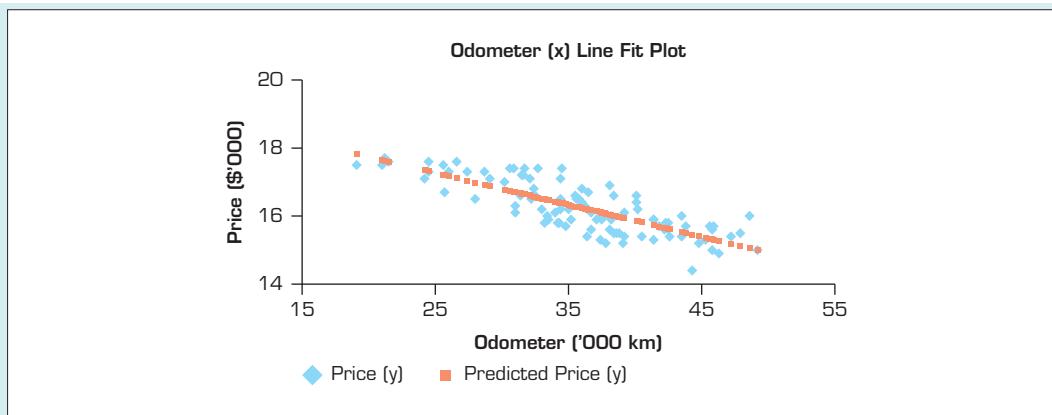
Using the computer

Using Excel Data Analysis

The complete printouts are shown below. The printouts include more statistics than we need at this point; however, we will discuss these statistics later in this chapter. We have also included the scatter diagram, which is often a first step in the regression analysis. Notice that there appears to be a straight-line relationship between the two variables.

Excel output for Example 15.3

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.8083					
5	R Square	0.6533					
6	Adjusted R Square	0.6498					
7	Standard Error	0.4526					
8	Observations	100					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	37.8211	37.8211	184.6583	0.0000	
13	Residual	98	20.0720	0.2048			
14	Total	99	57.8931				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	19.6114	0.2524	77.6965	0.0000	19.1105	20.1123
18	Odometer (x)	-0.0937	0.0069	-13.5889	0.0000	-0.1074	-0.0800



COMMANDS

- 1 Type the data into two columns, one storing the dependent variable and the other the independent variable or open the data file (**XM15-03**).
- 2 Click **DATA**, **Data Analysis**, and **Regression**. Click **OK**.
- 3 Specify the **Input Y Range** (**A1:A101**) and the **Input X Range** (**B1:B101**). Click **Labels** (if the first row contains the column headings).
- 4 Click **Output Range** and specify the output start cell reference or click **New Worksheet Ply:** and type the name of the sheet for the output **Regression**.
- 5 To draw the scatter diagram and the fitted points on the regression line, click **Line Fit Plots**, or follow the instructions provided in Chapter 4 on page 113. Click **OK**.

Using XLSTAT

	A	B	C	D	E	F	G
18	<i>Model parameters (Price (y)):</i>						
19	Source	Value	Standard error	t	Pr > t	Lower bound (95%)	Upper bound (95%)
20	Intercept	19.611	0.252	77.697	< 0.0001	19.110	20.112
21	Odometer (x)	-0.094	0.007	-13.589	< 0.0001	-0.107	-0.080

COMMANDS

- 1 Type the data or open the data file (**XM15-03**).
- 2 Click **XLSTAT**, **Modeling data**, and **Linear regression**.
- 3 In the **Quantitative** box type the input range of Y (**A1:A101**). In the **X Explanatory variables Quantitative** box type the input range of X (**B1:B101**).
- 4 Click **Outputs** and check **Analysis of variance**. Click **OK**.

In the sections that follow, we will return to this problem and the computer output to introduce other statistics associated with regression analysis.

EXERCISES

- 15.4** The term *regression* was originally used in 1885 by Sir Francis Galton in his analysis of the relationship between the heights of children and their parents. He formulated the 'law of universal regression', which specifies that 'each peculiarity in a man is shared by his kinsmen, but on average in a less degree'. (Evidently, people spoke this way in 1885.) In 1903, two statisticians, K. Pearson and A. Lee, took a random sample of 1078 father–son pairs to examine Galton's law. Their sample regression line was

$$\text{Son's height} = 33.73 + 0.516 \times \text{Father's height}$$

- a Interpret the coefficients.
- b What does the regression line tell you about the heights of sons of tall fathers?
- c What does the regression line tell you about the heights of sons of short fathers?

- 15.5 XR15-05** Consider the following data for the two variables, x and y .

x	-5	-2	0	3	4	7
y	15	9	7	6	4	1

- a Determine the least squares regression line and interpret the coefficients.
- b Calculate the sum of squared errors (SSE).

- 15.6 XR15-06** Consider the observations for the two variables, x and y :

x	1	2	3	4	5	6	7	8	9
y	5	28	17	14	27	33	39	26	30

- a Determine the least squares regression line and interpret the coefficients.
- b Calculate the sum of squares errors (SSE).

- 15.7** A set of 10 observations for the two variables x and y to be analysed by a linear regression model yields the following summations:

$$\begin{aligned}\sum x &= 31 & \sum y &= 37 & \sum xy &= 75 \\ \sum x^2 &= 103 & \sum y^2 &= 445\end{aligned}$$

Determine the least squares regression line and interpret the coefficients.

- 15.8** A set of 25 observations of two variables x and y produced the following summations:

$$\begin{aligned}\sum x &= 62.5 & \sum y &= 129.0 & \sum xy &= 141.1 \\ \sum x^2 &= 317.8 & \sum y^2 &= 936.4\end{aligned}$$

Determine the least squares regression line and interpret the coefficients.

- 15.9** In a study of the relationship between two variables x and y , the following summations were calculated:

$$\begin{array}{lll}\sum x = 105 & \sum y = 4414 & \sum xy = 37\ 525 \\ \sum x^2 = 956 & \sum y^2 = 1818\ 421 & n = 15\end{array}$$

Determine the least squares regression line and interpret the coefficients.

Applying the techniques

- 15.10 XR15-10 Self-correcting exercise.** The accompanying table exhibits the annual profit per dollar of sales y (measured in cents) for eight delicatessens, and the number of employees per store x .

x	3	6	4	5	2	5	4	1
y	22	30	20	25	18	26	22	19

- a Draw a scatter diagram.
- b Determine the least squares regression line and interpret the slope coefficient estimate.
- c Graph the regression line on the scatter diagram.
- d Does it appear that annual profit and the number of employees are linearly related?

- 15.11** A custom jobber of speciality fibreglass-bodied cars wished to estimate overhead expenses (labelled y and measured in '\$000s) as a function of the number of cars (x) produced monthly. A random sample of 12 months was recorded and the following statistics calculated:

$$\begin{array}{lll}\sum x = 157 & \sum y = 57 & \sum xy = 987 \\ \sum x^2 = 4102 & \sum y^2 = 413\end{array}$$

- a Compute the least squares regression line.
- b Interpret the slope coefficient estimate ($\hat{\beta}_1$).

- 15.12 XR15-12** Refer to Example 4.5. The real estate agent wants to analyse the relationship between the construction price of a house and the size of the house measured by the number of squares of living space in the house. He used the same data he utilised in Example 4.5, which are listed in the following table.

House size (squares) and construction price (\$'0000)

Size (squares)	Price (\$'0000)	Size (squares)	Price (\$'0000)
20	22	29	35
21	26	30	34
31	40	26	24
32	38	33	38
24	30	27	33
25	31	34	40
22	27	28	35
23	27		
$\sum x = 405$	$\sum y = 480$	$n = 15$	
$\sum x^2 = 11215$	$\sum xy = 13296$		

To show that house size influence its selling price answer the following.

- a Draw a scatter diagram.
- b Determine the least squares regression line.
- c Interpret the y -intercept and slope and comment on the relationship between the two variables.

15.13 XR15-13 Twelve secretaries at a university in Queensland were asked to take a special three-day intensive course to improve their keyboarding skills. At the beginning and again at the end of the course, they were given a particular two-page letter and asked to type it flawlessly. The data shown in the following table were recorded.

- a Compute the equation of the regression line.
- b As a check of your calculations in part (a), plot the 12 points in a scatter diagram and graph the line.
- c Does it appear that the secretaries' experience is linearly related to their improvement?

Secretary	Number of years of experience x	Improvement (words per minute) y
A	2	9
B	6	11
C	3	8
D	8	12
E	10	14
F	5	9
G	10	14
H	11	13
I	12	14
J	9	10
K	8	9
L	10	10
$\sum x = 94$	$\sum y = 133$	$n = 12$
$\sum x^2 = 848$	$\sum y^2 = 1529$	$\sum xy = 1102$

15.14 XR15-14 Advertising is often touted as the key to success. In seeking to determine just how influential advertising is, the management of a recently established retail chain collected data on sales revenue and advertising expenditures from its stores over the previous 15 weeks, with the results shown in the following table.

Advertising expenditures (\$'000) x	Sales (\$'000) y	
3.0	50	
5.0	250	
7.0	700	
6.0	450	
6.5	600	
8.0	1000	
3.5	75	
4.0	150	
4.5	200	
6.5	550	
7.0	750	
7.5	800	
7.5	900	
8.5	1100	
7.0	600	
$\sum x = 91.5$	$\sum y = 8175$	$n = 15$
$\sum x^2 = 598.75$	$\sum y^2 = 6070625$	$\sum xy = 57787.5$

- a Compute the coefficients of the regression line using the least squares method.
- b What does the value of the intercept tell you?
- c Interpret the slope coefficient estimate.
- d If the sign of the slope were negative, what would that say about the advertising?

15.15 XR15-15 Critics of television often refer to the detrimental effects that violence shown on television has on children. However, there may be another problem. It may be that watching television also reduces the amount of physical exercise, causing weight gain. A sample of 15 10-year-old children was taken. The number of kilograms by which each child was overweight was recorded (a negative number indicates the child is underweight). The number of hours of television viewing per week was also recorded. These data are listed overleaf.

Observation	1	2	3	4	5	6	7	8
Television (hours)	42	34	25	35	37	38	31	33
Overweight (kg)	18	6	0	-1	13	14	7	7

Observation	9	10	11	12	13	14	15
Television (hours)	19	29	38	28	29	36	18
Overweight (kg)	-9	8	8	5	3	14	-7

- a Draw the scatter diagram.
- b Calculate the sample regression line and describe what the coefficients tell you about the relationship between the two variables.

15.16 XR15-16 To determine how the number of housing starts is affected by mortgage rates, an economist recorded the average mortgage rate and the number of new housing starts in a suburb in Tasmania for the past 10 years. These data are listed here.

REAL-LIFE APPLICATIONS

Retaining employees

Human resource managers are responsible for a variety of tasks within organisations. Personnel or human resource managers are involved with recruiting new workers, determining which applicants are most suitable to hire, and helping with various aspects of monitoring the workforce, including absenteeism and employee turnover. For many firms, employee turnover is a costly problem. First, there is the cost of recruiting and attracting qualified workers. The firm must advertise vacant positions and make certain that applicants are judged properly. Second, the cost of training the new employees can be high, particularly in technical areas. Third, new employees are often not as productive and efficient as experienced employees. Consequently, it is in the interests of the firm to attract

Year	1	2	3	4	5
Interest rate	8.5	7.8	7.6	7.5	8.0
Housing starts	115	111	185	201	206

Year	6	7	8	9	10
Interest rate	8.4	8.8	8.9	8.5	8.0
Housing starts	167	155	117	133	150

- a Determine the regression line.
- b What do the coefficients of the regression line tell you about the relationship between mortgage rates and housing starts?

Computer/manual applications

The following exercises require the use of a computer and software. Alternatively, they may be solved manually using the sample statistics provided.

and keep the best workers. Any information that the personnel manager can obtain is likely to be useful.



Source: Shutterstock.com/ESB Professional

15.17 XR15-17 The human resource manager of a telemarketing firm is concerned about the rapid turnover of the firm's telemarketers. It appears that many telemarketers do not work very long before quitting. There may be a number of reasons, including relatively low pay, personal unsuitability for the work and the low probability of advancement. Because of the high cost of hiring and training new workers, the manager decided to examine the factors that influence workers to quit. He reviewed the work history of a random sample of workers who had quit

in the last year and recorded the number of weeks on the job before quitting and the age of each worker when originally hired.

- a Use regression analysis to describe how the work period and age are related.
- b Briefly discuss what the coefficients tell you.

Sample statistics: $n = 80$;
 $\bar{X}_{\text{Age}} = 37.28$; $\bar{Y}_{\text{Emp}} = 26.28$;
 $s_{xy} = -6.44$; $s_x^2 = 55.11$; $s_y^2 = 4.00$.

REAL-LIFE APPLICATIONS

Testing job applicants

The recruitment process at many firms involves tests to determine the suitability of candidates. There may be written tests to determine whether the applicant has sufficient knowledge in his or her area of expertise to perform well on the job. There could be oral tests to determine whether the applicant's personality matches the needs of the job. Manual or technical skills can be tested through a variety of physical tests. The test results contribute to the decision to hire. In some cases, the test result is the only criterion to hire. Consequently, it is vital to ensure that the test is a reliable predictor of job performance. If the tests are poor predictors, they should be discontinued. Statistical analyses allow personnel managers to examine the link between the test results and job performance.



Source: Shutterstock.com/Rido

15.18 XR15-18 Although a large number of tasks in the computer industry are carried out by robots, many operations require human workers. Some jobs require a great deal of dexterity to properly position components into place. A large computer maker routinely tests applicants for these jobs by giving a dexterity test that involves a number of intricate finger and hand movements. The tests are scored on a 100-point scale. Only those who have scored above 70 are hired. To determine whether the tests are valid predictors of job performance, the personnel manager drew a random sample of 45 workers who had been hired two months earlier. She recorded their test scores and the percentage of non-defective computers they had produced in the last week. Determine the regression line and interpret the coefficients.

Sample statistics: $n = 45$;
 $\bar{x}_{\text{Test}} = 79.47$; $\bar{y}_{\text{non-defective}} = 93.89$;
 $s_{xy} = 0.83$; $s_x^2 = 16.07$; $s_y^2 = 1.28$.

15.19 XR15-19 The objective of commercials is to have as many viewers as possible remember the product in a favourable way and eventually buy it. In an experiment to determine how the length of a commercial affects people's memory of it, 60 randomly selected people were asked to watch a one-hour television program. In the middle of the show, a commercial advertising a brand of toothpaste appeared. Each viewer was exposed to varying lengths of the commercial, ranging between

20 and 60 seconds, but its essential content was the same. After the television program finished, each person took a test to determine how much he or she remembered about the product. The commercial times and test scores (on a 30-point test) are recorded.

- a Draw a scatter diagram of the data to determine whether a linear model appears to be appropriate.
- b Determine the least squares line.
- c Interpret the coefficients.

Sample statistics: $n = 60$;
 $\bar{x}_{\text{Length}} = 38.0$; $\bar{y}_{\text{Test}} = 13.8$;
 $s_{xy} = 51.864$; $s_x^2 = 193.898$; $s_y^2 = 47.959$.

15.20 XR15-20 After several semesters without much success, Pat Statsdud (a student in the lower quarter of a statistics subject) decided to try to improve. Pat needed to know the secret of success for university students. After many hours of discussion with other, more successful, students, Pat postulated a rather radical theory: the longer one studied, the better one's grade. To test the theory, Pat took a random sample of 100 students in an economics subject and asked each student to report the average amount of time he or she studied economics and the final mark received. These data are stored in columns 1 (study time in hours) and 2 (final mark out of 100).

- a Determine the sample regression line.
- b Interpret the coefficients.
- c Is the sign of the slope logical? If the slope had had the opposite sign, what would that tell you?

Sample statistics: $n = 100$;
 $\bar{x}_{\text{Study time}} = 27.95$; $\bar{y}_{\text{Final mark}} = 74.06$;
 $s_{xy} = 153.950$; $s_x^2 = 82.007$; $s_y^2 = 363.939$.

- 15.21 XR15-21** Suppose that a statistics practitioner wanted to update the study described in Exercise 15.4. She collected data on 400 father–son pairs and stored the data in columns 1 (fathers' heights in centimetres) and 2 (sons' heights in centimetres).

- a Determine the sample regression line.
- b What does the value of $\hat{\beta}_0$ tell you?
- c What does the value of $\hat{\beta}_1$ tell you?

Sample statistics: $n = 400$;
 $\bar{x}_{\text{Father}} = 167.86$; $\bar{y}_{\text{Son}} = 171.74$;
 $s_{xy} = 49.188$; $s_x^2 = 102.707$; $s_y^2 = 88.383$.

- 15.22 XR15-22** A health economist is investigating whether there is a linear relationship between people's ages and their medical expenses. He gathered data concerning the age and mean daily medical expenses of a random sample of 1348 Australians during the previous 12-month period. The data are recorded (column 1 = age; column 2 = mean daily medical expense).

- a Determine the sample regression line.
- b Interpret the coefficients.

Sample statistics: $n = 1348$;
 $\bar{x}_{\text{Age}} = 56.0$; $\bar{y}_{\text{Expense}} = 6.67$;
 $s_{xy} = 51.525$; $s_x^2 = 228.256$; $s_y^2 = 179.881$.

- 15.23 XR15-23** The growing interest in and use of the internet has forced many companies into considering ways of selling their products online. Therefore, such companies are interested in determining who is using the internet. A statistics practitioner undertook a study to determine how education and internet use are connected. She took a random sample of 200 adults (20 years of age and older) and asked each to report the years of education they had completed and the number of hours of internet use in the previous week. The responses are recorded.

- a Perform a regression analysis to describe how the two variables are related.
- b Interpret the coefficients.

Sample statistics: $n = 200$;
 $\bar{x}_{\text{Education}} = 11.04$; $\bar{y}_{\text{Internet}} = 6.67$;
 $s_{xy} = 3.08$; $s_x^2 = 3.90$; $s_y^2 = 22.16$

- 15.24 XR15-24** An economist for the federal government is attempting to produce a better measure of poverty than is currently in use. To help acquire information, she

recorded the annual household income (in '\$000s) and the amount of money spent on food during one week for a random sample of households. Determine the regression line and interpret the coefficients.

Sample statistics: $n = 150$;
 $\bar{x}_{\text{Income}} = 59.42$; $\bar{y}_{\text{Food exp}} = 270.26$;
 $s_{xy} = 225.660$; $s_x^2 = 115.240$; $s_y^2 = 1797.25$.

- 15.25 XR15-25** In an attempt to determine the factors that affect the amount of energy used, 200 households were analysed. In each, the number of occupants and the amount of electricity used were measured. Determine the regression line and interpret the results.

Sample statistics: $n = 200$;
 $\bar{x}_{\text{Occupants}} = 4.75$; $\bar{y}_{\text{Electricity}} = 762.6$;
 $s_x^2 = 4.84$; $s_y^2 = 56725$; $s_{xy} = 310.0$.

- 15.26 XR15-26** Besides their known long-term effects, do cigarettes also cause short-term illnesses such as colds? To help answer this question, a sample of smokers was drawn. Each person was asked to report the average number of cigarettes smoked per day and the number of days absent from work due to colds last year.

- a Determine the regression line.
- b What do the coefficients tell you about the relationship between smoking cigarettes and sick days because of colds?

Sample statistics: $n = 231$;
 $\bar{x}_{\text{Cigarettes}} = 37.64$; $\bar{y}_{\text{Days}} = 14.43$;
 $s_x^2 = 108.3$; $s_y^2 = 19.8$; $s_{xy} = 20.55$

- 15.27 XR15-27** In an attempt to analyse the relationship between advertising and sales, the owner of a furniture store recorded the monthly advertising budget (\$ thousands) and the sales (\$ millions) for a sample of 12 months. The data are recorded and stored.

- a Draw a scatter diagram. Does it appear that advertising and sales are linearly related?
- b Calculate the least squares line and interpret the coefficients.

Sample statistics: $n = 12$;
 $\bar{x}_{\text{Advertising}} = 51.08$; $\bar{y}_{\text{Sales}} = 12.075$;
 $s_x^2 = 749.72$; $s_y^2 = 4.19$; $s_{xy} = 43.66$

15.3 Error variable: Required conditions

In the previous section, we used the least squares method to estimate the coefficients of the linear regression model. A critical part of this model is the error variable. In the next section, we will present an inferential method that determines whether there is a relationship between the dependent and independent variables. Later we will show how we use the regression equation to estimate and predict. For these methods to be valid, however, four requirements involving the probability distribution of the error variable must be satisfied.

Required conditions for the error variable

- 1 The probability distribution of ε is normal.
- 2 The mean of the distribution of ε is zero; that is, $E(\varepsilon) = 0$
- 3 The standard deviation of ε is σ_ε , which is a constant no matter what the value of x is. (Errors with this property are called homoscedastic errors.)
- 4 The errors associated with any two values of y are independent. As a result, the value of the error variable at one point does not affect the value of the error variable at another point. (Errors that do not satisfy this requirement are known as autocorrelated errors.)
- 5 The errors (ε) are independent of the independent variables.

Requirements 1, 2 and 3 can be interpreted in another way: for each value of x , y is a normally distributed random variable whose mean is

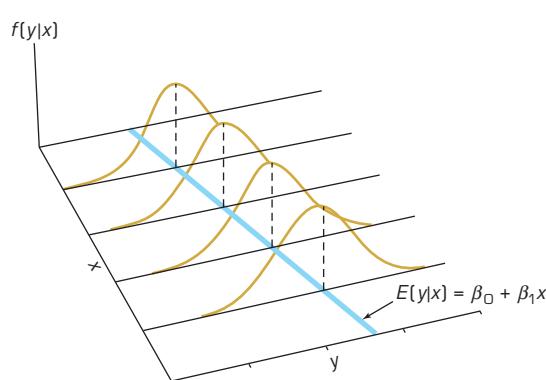
$$E(y) = \beta_0 + \beta_1 x$$

and whose standard deviation is σ_ε . Notice that the mean depends on x . To reflect this dependence, the expected value is sometimes expressed as

$$E(y|x) = \beta_0 + \beta_1 x$$

The standard deviation, however, is not influenced by x , because it is a constant over all values of x . **Figure 15.5** depicts this interpretation. In Section 15.7, we will discuss how departures from these required conditions affect the regression analysis and how they are identified.

FIGURE 15.5 Distribution of y given x



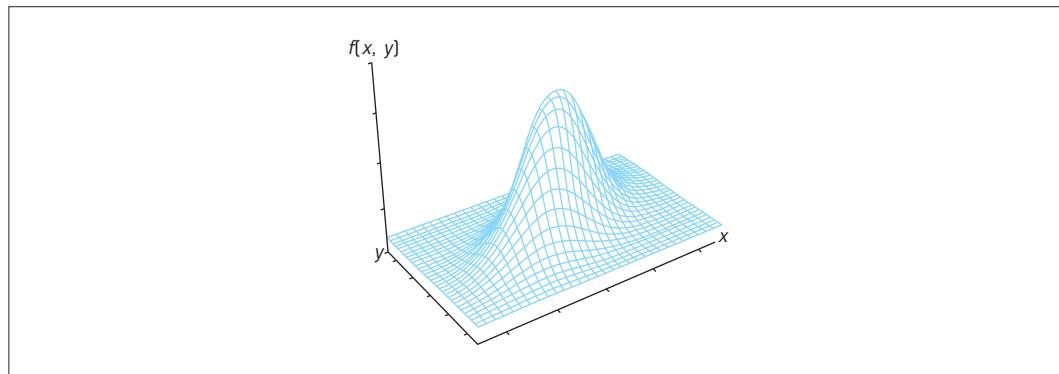
Experimental and observational data

In Chapter 2, we described the difference between observational and experimental data. We pointed out that statistics practitioners often design controlled experiments to enable them to interpret the results of their analyses more clearly than would be the case after conducting an observational study. Example 15.3 is an illustration of observational data. In that example, we merely observed the odometer reading and auction selling price of 100 randomly selected cars. If you examine Exercise 15.19, you will see experimental data gathered through a controlled experiment. To determine the effect of the length of a television commercial on its viewers' memories of the product advertised, the statistics practitioner arranged for 60 television viewers to watch a commercial of differing lengths and then tested their memories of that commercial. Each viewer was randomly assigned a commercial length. The values of x ranged from 20 to 60 and were set by the statistics practitioner as part of the experiment. For each value of x , the distribution of the memory test scores is assumed to be normally distributed with a constant variance.

We can summarise the difference between the experiment described in Example 15.3 and the one described in Exercise 15.19. In Example 15.3, both the odometer reading and the auction selling price are random variables. We hypothesise that for each possible odometer reading, there is a theoretical population of auction selling prices that are normally distributed with a mean that is a linear function of the odometer reading and a variance that is constant. In Exercise 15.19, the length of the commercial is not a random variable but a series of values selected by the statistics practitioner. For each commercial length, the memory test scores are required to be normally distributed with a constant variance.

Regression analysis can be applied to data generated from either observational or controlled experiments. In both cases, our objective is to determine how the independent variable is related to the dependent variable. However, observational data can be analysed in another way. When the data are observational, both variables are random variables. We need not specify that one variable is independent and the other is dependent. We can simply determine whether the two variables are related. The equivalent of the required conditions described in the previous box is that the two variables are distributed as a bivariate normal distribution. (Recall that in Section 7.4 we introduced the bivariate distribution, which describes the joint probability of two variables.) A bivariate normal distribution is described in **Figure 15.6**. As you can see, it is a three-dimensional bell-shaped curve. The dimensions are the variables x , y , and the joint density function $f(x, y)$.

FIGURE 15.6 Bivariate normal distribution



In Section 15.4, we will discuss the statistical technique that is used when both x and y are random variables and they are bivariate normal distributed. In Chapter 16, we will introduce a procedure applied when the normality requirement is not satisfied.

EXERCISES

- 15.28** Describe what the required conditions mean in Exercise 15.19. If the conditions are satisfied, what can you say about the distribution of memory test scores?
- 15.29** What are the required conditions for Exercise 15.23? Do these seem reasonable?
- 15.30** Assuming that the required conditions in Exercise 15.25 are satisfied, what does this tell you about the distribution of energy consumption?

15.4 Assessing the model

The least squares method produces the best straight line. However, there may in fact be no relationship or perhaps a non-linear (e.g. quadratic) relationship between the two variables. If so, the use of a linear model is pointless. Consequently, it is important for us to assess how well the linear model fits the data. If the fit is poor, we should discard the linear model and seek another one.

Several methods are used to evaluate the model. In this section, we present two statistics and one test procedure to determine whether a linear model should be employed. They are the standard error of estimate, the t -test of the slope and the coefficient of determination. All these measures are based on the sum of squares for error (SSE).

15.4a Sum of squares for error

The least squares method determines the coefficients that minimise the sum of squared deviations between the points and the line defined by the coefficients. Recall from Section 15.2 that the minimised sum of squared deviations is called the sum of squares for error, denoted SSE. In that section, we demonstrated the direct method of calculating SSE. For each value of x , we compute the value of \hat{y} . In other words, for $i = 1$ to n , we compute

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

For each point, we then compute the difference between the actual value of y and the value calculated at the line, which is the residual. We square each residual and sum the squared values. That is, for the least squares line,

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

where $e_i = y_i - \hat{y}_i$ is the residual of the i th observation.

In Example 15.2, we showed how SSE is found. Table 15.1 on page 632 shows these calculations for Example 15.1. To calculate SSE manually requires a great deal of arithmetic. We determined the value of \hat{y} for each value of x , calculated the difference between y and \hat{y} , squared the difference, and added. This procedure can be quite time consuming. Fortunately, there is a shortcut method available that uses the sample variances and the covariance.

Shortcut calculation for SSE

$$SSE = (n-1) \left[s_y^2 - \frac{s_{xy}^2}{s_x^2} \right]$$

where, as defined earlier,

$$s_y^2 = \frac{1}{(n-1)} \left[\sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i \right)^2}{n} \right]; \quad s_x^2 = \frac{1}{(n-1)} \left[\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \right]$$

and

$$s_{xy} = \frac{1}{(n-1)} \left[\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n} \right]$$

15.4b Standard error of estimate

In Section 15.3, we pointed out that the error variable ϵ is normally distributed with mean zero and standard deviation σ_ϵ . If σ_ϵ is large, some of the errors will be large, which implies that the model's fit is poor. If σ_ϵ is small, the errors tend to be close to the mean (which is zero), and, as a result, the model fits well. Hence, we could use σ_ϵ to measure the suitability of using a linear model. Unfortunately, σ_ϵ is a population parameter and, like most parameters, it is unknown. We can, however, estimate σ_ϵ from the data. The estimate is based on the sum of squares for error, SSE.

We can estimate σ_ϵ^2 by dividing SSE by the number of observations minus 2, where 2 represents the number of parameters estimated in the regression model – namely, β_0 and β_1 . That is, the sample statistic

$$s_e^2 = \frac{SSE}{n-2}$$

standard error of estimate

An estimate of the standard deviation of the error variable, which is the square root of the sum of squares error (SSE) divided by the degrees of freedom.

Standard error of estimate

$$s_e = \sqrt{\frac{SSE}{n-2}}$$

EXAMPLE 15.4

LO3

Odometer readings and prices of used cars: Part II

Find the standard error of estimate for Example 15.3.

Solution**Calculating manually**

To calculate the standard error of estimate, we need to find the sum of squares for error. This requires the calculation of s_x^2 , s_y^2 and s_{xy} . In Example 15.3, we calculated

$$s_{xy} = -4.077 \quad \text{and} \quad s_x^2 = 43.509$$

From the data, we find

$$s_y^2 = \frac{1}{(n-1)} \left[\sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i \right)^2}{n} \right] = \frac{1}{99} \left(26421.9 - \frac{(1623.7)^2}{100} \right) = 0.5848$$

We can now determine the sum of squares for error.

$$\text{SSE} = (n-1) \left[s_y^2 - \frac{s_{xy}^2}{s_x^2} \right] = (100-1) \left[0.5848 - \frac{(-4.077)^2}{43.509} \right] = 20.072$$

Thus, the standard error of estimate is

$$s_e = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{20.072}{100-2}} = 0.4526$$

Using the computer**Using Excel Data Analysis**

	A	B
5	Standard Error	0.4526

Using XLSTAT

Refer to the XLSTAT output for Example 15.3. XLSTAT reports the standard error of estimate as follows.

	A	B
5	RMSE	0.4526

Interpreting the results

The smallest value that s_e can assume is zero, which occurs when SSE = 0, that is, when all the points fall on the regression line. Thus, when s_e is small, the fit is excellent, and the linear model is likely to be an effective analytical and forecasting tool. If s_e is large, the model is a poor one, and the statistics practitioner should improve it or discard it.

We judge the value of s_e by comparing it to the values of the dependent variable y , or, more specifically, to the sample mean \bar{y} . In this example, s_e (= 0.4526) is only 2.8% relative to \bar{y} (= 16.24). Therefore, we could admit that the standard error of estimate is reasonably small. In general, the standard error of estimate cannot be used alone as an absolute measure of the model's utility.

Nonetheless, s_e is useful in comparing models. If the statistics practitioner has several models from which to choose, the one with the smallest value of s_e should generally be the one used. As you will see, s_e is also an important statistic in other procedures associated with regression analysis.

15.4c Estimating the slope and the intercept: Interval estimates

As discussed in Chapter 10, there are two types of estimators available – namely, point estimators and interval estimators – to estimate an unknown parameter. In Section 15.2 we used the least squares method to derive point estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ to estimate the intercept β_0 and the slope coefficient β_1 . Now we provide the interval estimators for β_0 and β_1 .

Confidence interval estimator of β_0 and β_1

$$\hat{\beta}_0 \pm t_{\alpha/2, n-2} s_{\hat{\beta}_0}$$

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} s_{\hat{\beta}_1}$$

where $s_{\hat{\beta}_1}$ (which is called the standard error of $\hat{\beta}_1$) is the standard deviation of $\hat{\beta}_1$ and is equal to

$$s_{\hat{\beta}_1} = \frac{s_e}{\sqrt{(n-1)s_x^2}}$$

and $s_{\hat{\beta}_0}$ is the standard error of β_0 and is equal to

$$s_{\hat{\beta}_0} = \frac{s_e \sqrt{\sum x_i^2}}{\sqrt{(n-1)s_x^2}} = s_{\hat{\beta}_1} \sqrt{\frac{\sum x_i^2}{n}}$$

EXAMPLE 15.5

LO2

Odometer readings and prices of used cars: Part III

Determine the 95% confidence interval estimate of the slope β_1 for Example 15.3.

Solution

Calculating manually

In Example 15.3, we have

$$s_{\hat{\beta}_1} = \frac{s_e}{\sqrt{(n-1)s_x^2}} = \frac{0.4526}{\sqrt{99 \times 43.509}} = 0.0069$$

Therefore, the 95% confidence interval estimate of β_1 for Example 15.3 is

$$\begin{aligned} & \hat{\beta}_1 \pm t_{0.025, 98} \times s_{\hat{\beta}_1} \\ &= -0.0937 \pm 1.984 \times 0.0069 \\ &= -0.0937 \pm 0.00137 \\ &= [-0.107, -0.080] \end{aligned}$$

Thus the 95% confidence interval estimate of the slope is the interval from -0.107 to -0.080 .

Using the computer

The output below was taken from the output for Example 15.3. Excel and XLSTAT report the confidence interval estimate in the column next to that of the p -value.



Using Excel Data Analysis

Excel output for Example 15.5

	A	B	C	D	E	F	G
	<i>Coefficients</i>		<i>Standard Error</i>	<i>t Stat</i>	<i>p-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept		19.6114	0.2524	77.6965	0.0000	19.1105	20.1123
Odometer (x)		-0.0937	0.0069	-13.5889	0.0000	-0.1074	-0.0800

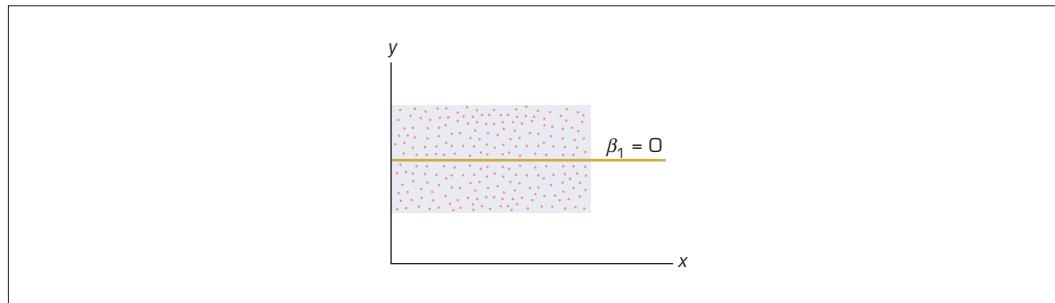
Using XLSTAT

	A	B	C	D	E	F	G
18	<i>Model parameters (Price (y)):</i>						
19	<i>Source</i>	<i>Value</i>	<i>Standard error</i>	<i>t</i>	<i>Pr > t </i>	<i>Lower bound (95%)</i>	<i>Upper bound (95%)</i>
20	Intercept	19.611	0.252	77.697	< 0.0001	19.110	20.112
21	Odometer (x)	-0.094	0.007	-13.589	< 0.0001	-0.107	-0.080

15.4d Testing the slope

To understand this method of assessing the linear model, consider the consequences of applying the regression technique to two variables that are not at all linearly related. If we could observe the entire population and draw the regression line, we would observe the scatter diagram shown in **Figure 15.7**. The line is horizontal, which means that no matter what value of x is used, we would estimate the same value for \hat{y} ; thus, y is not linearly related to x . Recall that a horizontal straight line has a slope of 0, that is, $\beta_1 = 0$.

FIGURE 15.7 Scatter diagram of entire population with $\beta_1 = 0$

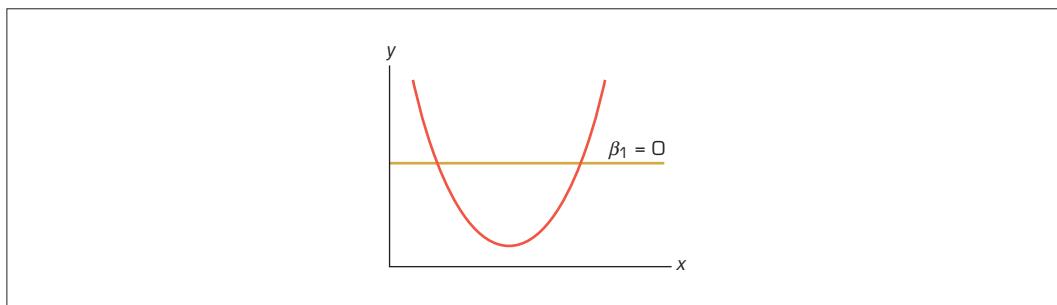


Because we rarely examine complete populations, the parameters are unknown. However, we can draw inferences about the population slope β_1 from the sample slope $\hat{\beta}_1$.

The process of testing hypotheses about β_1 is identical to the process of testing any other parameter. We begin with the hypotheses. The null hypothesis specifies that there is no linear relationship, which means that the slope is zero. Thus, we specify

$$H_0: \beta_1 = 0$$

It must be noted that if the null hypothesis is true, it does not necessarily mean that no relationship exists. It means that *no linear* relationship exists. For example, the quadratic relationship described in **Figure 15.8** may exist for which $\beta_1 = 0$.

FIGURE 15.8 Quadratic relationship

We can conduct one- or two-tail tests of β_1 . Most often, we perform a two-tail test to determine whether there is sufficient evidence to infer that a linear relationship exists. We test

$$H_A: \beta_1 \neq 0$$

If we wish to test for positive or negative linear relationships, we conduct one-tail tests. To illustrate, suppose that in Example 15.4 we wanted to know whether there is evidence of a negative linear relationship between odometer reading and auction selling price. We would specify the hypotheses as

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 < 0$$

Test statistic and sampling distribution

In Section 15.2, we pointed out that $\hat{\beta}_1$ is an unbiased estimator of β_1 ; that is,

$$E[\hat{\beta}_1] = \beta_1$$

The estimated standard error of $\hat{\beta}_1$ is

$$S_{\hat{\beta}_1} = \frac{s_e}{\sqrt{(n-1)s_x^2}}$$

where s_e is the standard error of estimate and s_x^2 is the sample variance of the independent variable. If the required conditions outlined in Section 15.3 are satisfied, the sampling distribution of the t -statistic is

$$t = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}}$$

distributed as a t distribution with degrees of freedom, $v = n - 2$. Note that 2 represents the number of parameters estimated in the regression model. Also, notice that the standard error of $\hat{\beta}_1$ decreases when the sample size increases (which makes $\hat{\beta}_1$ a consistent estimator of β_1) or the variance of the independent variable increases.

Thus, the test statistic is

$$t = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \sim t_{n-2}$$

EXAMPLE 15.6

LO4

Are odometer reading and price of used cars related?

Test to determine whether there is enough evidence in Example 15.3 to infer that a linear relationship exists between the price of a car and the odometer reading. Use a significance level of 5%.

Solution*Hypotheses:*

$$\begin{aligned} H_0: \beta_1 &= 0 \\ H_A: \beta_1 &\neq 0 \end{aligned} \quad (\text{Two-tail test})$$

Test statistic:

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \sim t_{n-2}$$

Level of significance:

$$\alpha = 0.05$$

Decision rule:

$$\text{Reject } H_0 \text{ if } |t| > t_{\alpha/2, n-2} = t_{0.025, 98} = 1.984.$$

Alternatively, reject H_0 if $p\text{-value} < \alpha = 0.05$

Calculating manually*Value of the test statistic:*

To calculate the value of the test statistic, we need $\hat{\beta}_1$ and $s_{\hat{\beta}_1}$.

In Example 15.3, we found $s_x^2 = 43.509$ and $\hat{\beta}_1 = -0.0937$.

In Example 15.4, we found $s_e = 0.4526$. Thus,

$$s_{\hat{\beta}_1} = \frac{s_e}{\sqrt{(n-1)s_x^2}} = \frac{0.4526}{\sqrt{99 \times 43.509}} = 0.0069$$

The value of the test statistic is

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} = \frac{-0.0937 - 0}{0.0069} = -13.58$$

Conclusion:

As $t = -13.58 < -1.984$, reject the null hypothesis. (Alternatively, using the computer, the $p\text{-value} = 0 < 0.05 = \alpha$, which rejects the null hypothesis.)

Interpreting the results

The value of the test statistic is $t = -13.59$, with a $p\text{-value}$ of 0. There is overwhelming evidence to infer that $\beta_1 \neq 0$ and a linear relationship exists. What this means is that the odometer reading does affect the auction selling price of the cars. As was the case when we interpreted the y -intercept, the conclusion we draw here is valid only over the range of the values of the independent variable. That is, we can infer that there is a linear relationship between odometer reading and auction price for the five-year-old Ford Laser whose odometer reading lies between 19 100 and 49 200 km (the minimum and maximum values of x ('000) in the sample). Because we have no observations outside this range, we do not know how, or even whether, the two variables are related for odometer values outside that range. This issue is particularly important to remember when we use the regression equation to estimate or forecast (see Section 15.5).

Using the computer

The output below was taken from the Excel and XLSTAT output for Example 15.3.

Using Excel Data Analysis

Excel output for Example 15.6

	A	B	C	D	E	F	G
		Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
Intercept		19.6114	0.2524	77.6965	0.0000	19.1105	20.1123
Odometer (x)		-0.0937	0.0069	-13.5889	0.0000	-0.1074	-0.0800

Using XLSTAT

	A	B	C	D	E	F	G
18	Model parameters (Price [y]):						
19	Source	Value	Standard error	t	Pr > t	Lower bound (95%)	Upper bound (95%)
20	Intercept	19.611	0.252	77.697	< 0.0001	19.110	20.112
21	Odometer (x)	-0.094	0.007	-13.589	< 0.0001	-0.107	-0.080

The output includes an estimate of β_1 , the standard error of $\hat{\beta}_1$, the t-statistic (t Stat) and the two-tail² p-value of the test (p-value). These values are -0.0937, 0.0069, -13.5889 and 0, respectively.

Suppose that in Example 15.3 we wanted to know whether there is evidence of a negative linear relationship between odometer reading and auction selling price. The value of the test statistic would be exactly the same (-13.5889). However, in this case the p-value would be the two-tail p-value divided by 2; using the p-value from the output, this would be $0.000/2 = 0.000$. As $p\text{-value} = 0.000 < 0.05 = \alpha$, we reject the null hypothesis and infer that there is a negative linear relationship between odometer reading and auction price.

Notice the output includes a test for β_0 . However, as we have pointed out before, interpreting the value of the y-intercept can lead to erroneous, if not ridiculous, conclusions. As a result, we will ignore the test of β_0 .

15.4e Coefficient of determination

coefficient of determination (R^2)

The proportion of the variation in the dependent variable that is explained by the variation in the independent variable(s).

Coefficient of determination

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}$$

With a little algebra, statisticians can show that

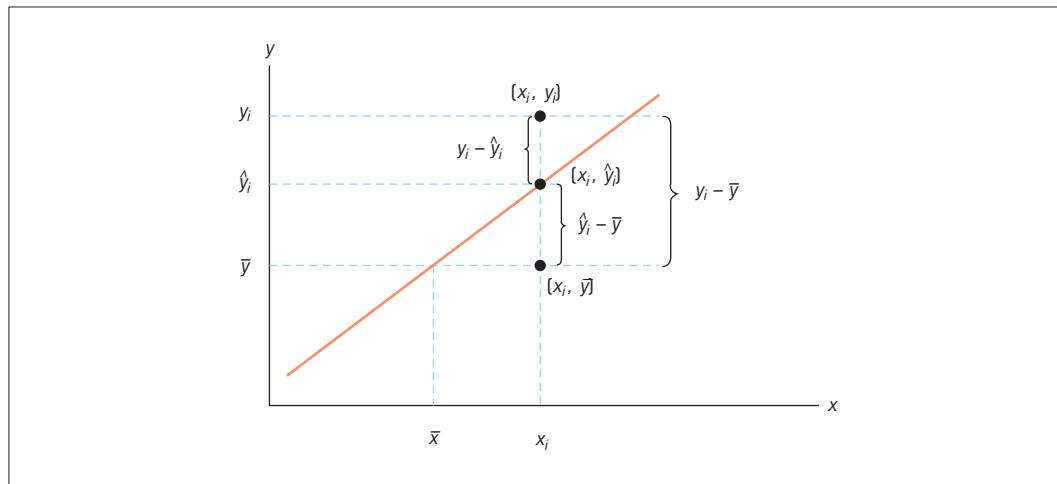
$$R^2 = 1 - \frac{\text{SSE}}{\sum (y_i - \bar{y})^2} = 1 - \frac{\text{SSE}}{(n-1)s_y^2}$$

The significance of this formula is based on the analysis of variance technique. Here, we begin the discussion by observing that the deviation between y_i and \bar{y} can be decomposed into two parts. By adding and subtracting \hat{y}_i from the deviation between y_i and \bar{y} ; that is,

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

This equation is represented graphically for the i th observation (x_i, y_i) in **Figure 15.9**.

² The p-value provided by Excel is for a two-tailed test. If using a one-tailed test, the Excel p-value should be divided by 2.

FIGURE 15.9 Analysis of the deviation

Now we ask, why are the values of y different from one another? In Example 15.3, we observe that the auction selling prices of the cars vary, and we would like to explain why. From **Figure 15.9**, we see that part of the difference between y_i and \bar{y} is the difference between \hat{y}_i and \bar{y} , which is accounted for by the difference between x_i and \bar{x} . That is, some of the price variation is *explained* by the odometer reading. The other part of the difference between y_i and \bar{y} , however, is accounted for by the difference between y_i and \hat{y}_i . This difference is the residual, which to some degree reflects variables not otherwise represented by the model. (These variables likely include the local supply and demand for this type of used car, the colour of the car and other relatively small details.) As a result, we say that this part of the difference is *unexplained* by the variation in x .

If we now square both sides of the equation, sum over all sample points and perform some algebra, we produce

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

$$\text{Total variation} = \text{Unexplained variation} + \text{Explained variation}$$

The quantity on the left side of this equation is $SST = (n-1)s_y^2$, which is the total measure of the variation in the dependent variable (selling price), also known as *sum of squares total*. The first quantity on the right side of the equation is SSE , also known as *sum of squares errors*, and the second term is SSR , also known as *sum of squares regression*. We can rewrite the equation as

$$\text{Total variation in } y = SST = SSE + SSR$$

where

$$\text{Total variation} = SST = \sum (y_i - \bar{y})^2 = (n-1)s_y^2$$

$$\text{Unexplained variation} = SSE = \sum (y_i - \hat{y}_i)^2$$

$$\text{Explained variation} = SSR = \sum (\hat{y}_i - \bar{y})^2$$

As we did in the analysis of variance, we partition the variation of y into two parts: SSE , which measures the amount of variation in y that remains *unexplained*; and SSR , which measures the amount of variation in y that is *explained* by the variation in the dependent variable (odometer reading). Therefore, from the definition of R^2 ($= SSR/SST$), it can easily be seen that R^2 measures the proportion of the variation in y that is explained by the regression model – in other words, explained by the variation in x . The higher the value of R^2 , the greater the explanatory power of the estimated regression model. We can incorporate this analysis into the definition of R^2 .

Coefficient of determination

$$R^2 = 1 - \frac{\text{SSE}}{\sum(y_i - \bar{y}_i)^2} = \frac{\sum(y_i - \bar{y}_i)^2 - \text{SSE}}{\sum(y_i - \bar{y}_i)^2} = \frac{\text{Explained variation}}{\text{Total variation in } y}$$

It follows that R^2 measures the proportion of the variation in y that can be explained by the variation in x .

We can also see that

$$R^2 = \frac{\text{Explained variation in } y}{\text{Total variation in } y} = \frac{\text{SSR}}{\text{SST}} = \frac{\text{SST} - \text{SSE}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

Because $\text{SST} \geq \text{SSR}$ and R^2 is the ratio of two sums of squares, it follows that

$$0 \leq R^2 \leq 1$$

It also follows that if SSE is small relative to SST, then R^2 will be close to 1, and the estimated regression model explains most of the variation in y . Incidentally, the notation R^2 is derived from the fact that the coefficient of determination is the coefficient of correlation squared. Recall that we introduced the sample coefficient of correlation in Chapter 5 and labelled it r . (To be consistent with the computer output, we capitalise r in the definition of the coefficient of determination.) We will discuss the coefficient of correlation in Section 15.6.

EXAMPLE 15.7

LO5

Measuring the strength of the relationship between odometer readings and prices of used cars

Find the coefficient of determination for the data in Example 15.3.

Solution

Calculating manually

The coefficient of determination is

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

where $\text{SST} = \sum(y_i - \bar{y})^2 = (n - 1)s_y^2$; $\text{SSE} = \sum(y_i - \hat{y}_i)^2$.

From Example 15.4, $\text{SST} = (n - 1)s_y^2 = 99(0.5848) = 57.89$ and $\text{SSE} = 20.072$.

Alternatively, from the computer output in Example 15.3, we have

$$\text{SST} = 57.89 \text{ and } \text{SSE} = 20.07.$$

Thus,

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{20.07}{57.89} = 1 - 0.3467 = 0.6533$$

Alternatively, we can use the alternate formula

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}$$

From Examples 15.1 and 15.4, we have

$$s_{xy} = -4.077, s_x^2 = 43.509 \text{ and } s_y^2 = 0.5848$$



Therefore,

$$R^2 = \frac{(-4.077)^2}{43.509 \times 0.5848} = 0.6533$$

Interpreting the results

We found that R^2 is equal to 65%. This statistic tells us that 65% of the variation in the auction selling prices is explained by the variation in the odometer readings. The remaining 35% is unexplained. Unlike the value of a test statistic, the coefficient of determination does not have a critical value that enables us to draw conclusions. We know that the higher the value of R^2 , the better the model fits the data. From the t -test of the significance of β_1 we know already that there is evidence of a linear relationship. The coefficient of determination merely supplies us with a measure of the strength of that relationship. As you will discover in the next chapter, when we improve the model, the value of R^2 increases.

Using the computer

Refer to the Excel output for Example 15.3. Excel reports the coefficient of determination as

	A	B
7	R Square	0.6533

Excel outputs a second R^2 statistic called the *coefficient of determination* adjusted for degrees of freedom. We will define and describe this statistic in Chapter 16.

15.4f Other parts of the computer printout

The last part of the Excel output for Example 15.3 relates to our discussion of the interpretation of the value of R^2 , where its meaning is derived from the partitioning of the variation in y . The values of SSR and SSE are shown in an analysis of variance (ANOVA) table. The general format of the table is shown below. The F -test performed in the ANOVA table will be explained in Chapter 16.

General form of the ANOVA table in the simple regression model

Source	d.f.	Sums of squares	Mean squares	F-value
Regression	1	SSR	MSR = SSR/1	$F = \text{MSR}/\text{MSE}$
Error	$n - 2$	SSE	MSE = SSE/($n - 2$)	
Total	$n - 1$	SST		

Note: Excel refers to the second source of variation 'Sum of squares errors' as 'Sum of squares residuals'.

The ANOVA table for Example 15.3 can be seen on page 634.

15.4g Developing an understanding of statistical concepts

Once again, we encounter the concept of explained variation. We first discussed the concept in Section 13.2 when we introduced the matched pairs experiment, which was designed to reduce the variation among experimental units. This concept was extended in the analysis of variance, where we partitioned the total variation into two or more sources (depending on the experimental design). And now in regression analysis, we use the concept to measure how the dependent variable is related to the independent variable. We partition the variation of the dependent variable into two sources: the variation explained by the variation in the independent variable and the unexplained variation. The greater the explained variation, the better the model is. We often refer to the coefficient of determination as a measure of the *explanatory power* of the model.

15.4h Cause-and-effect relationship

A common mistake is made by many students when they attempt to interpret the results of a regression analysis in which there is evidence of a linear relationship. They imply that changes in the independent variable *cause* changes in the dependent variable. It must be emphasised that we cannot infer a causal relationship from statistics alone. Any inference about the cause of the changes in the dependent variable must be justified by a reasonable theoretical relationship. For example, statistical tests established that the more one smoked, the greater the probability of developing lung cancer. However, this analysis did not prove that smoking causes lung cancer. It only demonstrated that smoking and lung cancer were somehow related. Only when medical investigations established the connection were scientists confidently able to declare that smoking causes lung cancer.

As another illustration, consider Example 15.3 in which we showed that the odometer reading is linearly related to the car's auction price. Although it seems reasonable to conclude that decreasing the odometer reading would cause the auction price to rise, this conclusion may not be entirely true. It is theoretically possible that the price is determined by the overall condition of the car and that the condition generally worsens when the car is driven further. Another analysis would be needed to establish the veracity of this conclusion.

Be cautious about the use of the terms 'explained variation' and 'explanatory power of the model'. Do not interpret the word 'explained' to mean 'caused'. We say that the coefficient of determination measures the amount of variation in y that is explained (not caused) by the variation in x . Thus, regression analysis can only show that a statistical relationship exists. We cannot infer that one variable causes another.

Recall that we first pointed this out in Chapter 4 (page 115) using the following sentence:

Correlation is not causation.

15.4i Applications in finance: Market model

In this section we describe one of the most important applications of simple linear regression. It is the well-known and often applied *market model*. This model assumes that the rate of return on a stock is linearly related to the rate of return on the overall market. The mathematical description of the model is

$$R = \beta_0 + \beta_1 R_m + \varepsilon$$

where R is the return on a particular share and R_m is the return on some major stock index, such as the Australian All Ordinaries Index.

The coefficient β_1 is called the share's beta coefficient, which measures how sensitive the stock's rate of return is to changes in the level of the overall market. For example, if β_1 is greater than 1, the stock's rate of return is more sensitive to changes in the level of the overall market than is the average stock. To illustrate, suppose that $\beta_1 = 2$. Then a 1% increase in the index results in an average increase of 2% in the stock's return. A 1% decrease in the index produces an average 2% decrease in the stock's return. Thus, a stock with a beta coefficient greater than 1 will tend to be more volatile than the market.

A stock's beta coefficient is determined using the statistical tools described in this chapter. The regression analysis produces $\hat{\beta}_1$, which is an estimate of a stock's beta coefficient. The coefficient of determination is also an important part of the financial-statistical analysis.

We will now present the solution for the opening example to this chapter.

SPOTLIGHT ON STATISTICS

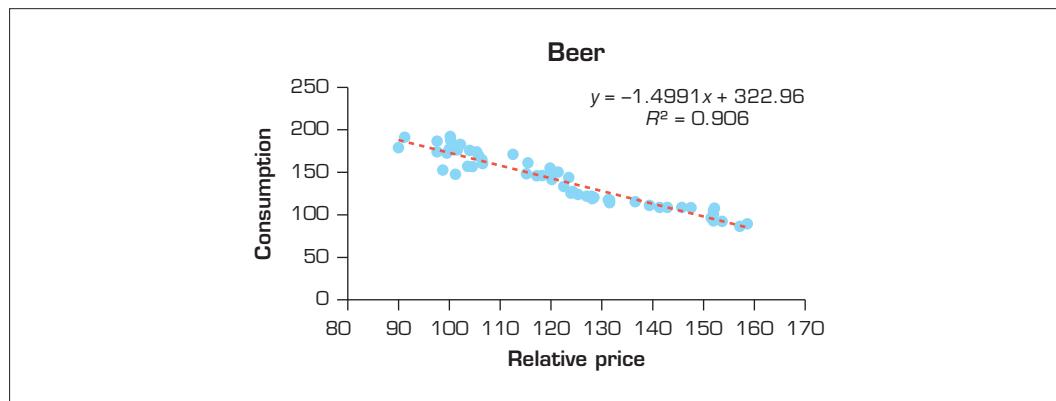
Would increasing tax on alcoholic beverages reduce consumption?: Solution

Identifying the technique

The problem objective is to analyse the relationship between two numerical (quantitative) variables. Because we want to know how beer price would affect beer consumption, the independent variable x is beer relative price (RPbeer) and the dependent variable y is beer consumption (CONSbeer). First we draw a scatter diagram to see whether a linear relationship is possible visually.



Source: iStock.com/Ilyabolotov



As can be seen, there appears to be a linear relationship between beer price and consumption. Therefore, the regression model to be estimated can be written as

$$y = \beta_0 + \beta_1 x + \varepsilon$$

We now estimate the model using the data for beer consumption and beer price. A similar analysis can be performed for wine.

Using the computer

Excel output from regression model estimation

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	<i>Regression Statistics</i>						
4	Multiple R	0.9518					
5	R Square	0.9060					
6	Adjusted R Square	0.9042					
7	Standard Error	9.5056					
8	Observations	54					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	45267.10	45267.10	500.99	0.0000	
13	Residual	52	4698.52	90.36			
14	Total	53	49965.62				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	322.9622	8.2696	39.0541	0.0000	306.3680	339.5564
18	RPBeer	-1.4991	0.0670	-22.3827	0.0000	-1.6336	-1.3647

The estimated regression equation is $\hat{y} = 322.96 - 1.499x$. The negative slope coefficient (-1.499) tells us that on average for each additional unit increase in the relative price of beer, consumption would decrease by 1.499 litres. We test to determine whether there is evidence of a negative linear relationship.

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 < 0 \quad (\text{Left one-tail test})$$

The test statistic is $t = -22.38$ and the p -value for a two-tail test is 2.33×10^{-28} which is virtually 0. The p -value for a one-tail test is $(0.00)/2 = 0.00$. Therefore, the null hypothesis should be rejected and one could conclude that there is a statistically significant negative effect on beer consumption due to increasing beer price. Hence, increasing the tax rate on beer would have the desired effect of reducing beer consumption.

The coefficient of determination is $R^2 = 0.9060$, which means that 90.6% of the variation in the consumption of beer is explained by the variation in beer price; the remaining 9.4% is not explained. This high value of R^2 close to 1 shows that the fitness of the linear model is very good.

A similar analysis can be performed for the wine data.

EXERCISES

Learning the techniques

15.31 Test each of the following hypotheses:

a $H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

$\hat{\beta}_1 = 1.87 \quad s_{\hat{\beta}_1} = 0.63 \quad n = 25 \quad \alpha = 0.05$

b $H_0: \beta_1 = 0$

$H_A: \beta_1 < 0$

$\hat{\beta}_1 = -26.32 \quad s_{\hat{\beta}_1} = 14.51 \quad n = 100 \quad \alpha = 0.01$

c $H_0: \beta_1 = 0$

$H_A: \beta_1 > 0$

$\hat{\beta}_1 = 0.056 \quad s_{\hat{\beta}_1} = 0.021 \quad n = 10 \quad \alpha = 0.05$

15.32 Estimate with 95% confidence the slope of the regression line, given the following data:

$$\sum x = 55 \quad \sum y = 325 \quad \sum xy = 929$$

$$\sum x^2 = 175 \quad \sum y^2 = 6051 \quad n = 20$$

15.33 Twelve observations of x and y produced the following summations:

$$\sum x = 65 \quad \sum y = 515 \quad \sum xy = 3085$$

$$\sum x^2 = 445 \quad \sum y^2 = 24815$$

- a Is there sufficient evidence at the 1% significance level to determine whether a positive linear relationship exists between x and y ?
- b Determine the 99% confidence interval estimate of the population slope.
- c Determine the coefficient of determination. What does this statistic tell you about the regression line?

15.34 You have been given the following data for variables x and y :

x	1	3	4	6	9	8	10
y	1	8	15	33	75	70	95

- a Draw the scatter diagram. Does it appear that x and y are related? If so, how?
- b Test to determine whether there is evidence of a linear relationship.

15.35 Suppose you have the following data for variables x and y :

x	3	5	2	6	1	4
y	25	110	9	250	3	71

- a Draw the scatter diagram. Does it appear that x and y are related? If so, how?
- b Test to determine whether there is evidence of a linear relationship.

Applying the techniques

15.36 **Self-correcting exercise.** Refer to Exercise 15.14.

- a Find the standard error of estimate. What does this value tell you about the linear regression model?
- b Determine the 95% confidence interval estimate for the slope.
- c Can we conclude at the 1% significance level that advertising is effective in increasing sales?
- d Determine the coefficient of correlation and discuss its value.

15.37 XR15-37 A new profit-sharing plan was introduced at a car-parts manufacturing plant in Geelong last year. Both management and union representatives are interested in determining how a worker's years of experience influence his or her productivity gains. After the plan had been in effect for a while, the following data were collected:

Worker	Years on assembly line (x)	Number of units manufactured daily (y)
1	15.1	110
2	7.0	105
3	18.6	115
4	23.7	127
5	11.5	98
6	16.4	103
7	6.3	87
8	15.4	108
9	19.9	112
	$\sum x = 133.9$	$\sum y = 965$
	$\sum xy = 14801.2$	$\sum x^2 = 2258.73$
		$\sum y^2 = 104469$

- a Find the least squares regression line.
- b Calculate the standard error of estimate. What does the value of s_e tell you about the relationship between x and y ?
- c Can we conclude at the 5% level of significance that a worker's years of experience influences his or her productivity gains?
- d Measure how well the linear model fits.

15.38 XR15-38 The manager of a supermarket chain performed a survey to help determine desirable locations for new stores. The manager wanted to know whether a linear relationship exists between weekly take-home pay and weekly food expenditures. A random sample of eight households produced the data shown in the following table.

Family	Weekly take-home pay (x)	Weekly food expenditure (y)
1	600	160
2	400	110
3	540	150
4	360	90
5	500	130
6	720	200
7	450	120
8	680	180

- a Find the least squares regression line.
- b Calculate the standard error of estimate.
- c Do these data provide sufficient evidence (with $\alpha = 0.01$) to allow us to conclude that a linear relationship exists between x and y ?
- d Calculate the coefficient of determination. What does the value tell you about the strength of the linear relationship?

15.39 XR15-39 In order to determine a realistic price for a new product that a company wishes to market, the company's research department selected 10 sites thought to have essentially identical sales potential and offered the product in each at a different price. The resulting sales are recorded in the following table.

Location	Price (x)	Sales (\$'000) (y)
1	15.00	15
2	15.50	14
3	16.00	16
4	16.50	9
5	17.00	12
6	17.50	10
7	18.00	8
8	18.50	9
9	19.00	6
10	19.50	5

- a Find the equation of the regression line.
- b Is there sufficient evidence at the 10% significance level to allow us to conclude that a negative linear relationship exists between price and sales?
- c Calculate the coefficients of correlation and determination.

Computer/manual applications

The following exercises require the use of a computer and software. Alternatively, they may also be completed manually using the sample statistics provided here or in the associated previous exercises.

15.40 XR15-40 Refer to Exercise 15.17. Use two statistics to measure the strength of the linear association. What do these statistics tell you?

Sample statistics: $n = 80$;
 $\bar{x}_{\text{Age}} = 37.28$; $\bar{y}_{\text{Emp}} = 26.28$;
 $s_{xy} = -6.44$; $s_x^2 = 55.11$; $s_y^2 = 4.00$.

15.41 XR15-41 Refer to Exercise 15.18, in which the personnel manager wanted to determine whether the test score is a valid predictor of job performance. Using the regression results obtained from the data, can we infer at the 5% significance level that the test score is a valid predictor? That is, can we infer that higher test scores are associated with higher percentages of non-defective units?

Sample statistics: $n = 45$;
 $\bar{x}_{\text{Test}} = 79.47$; $\bar{y}_{\text{non-defective}} = 93.89$;
 $s_{xy} = 0.83$; $s_x^2 = 16.07$; $s_y^2 = 1.28$.

15.42 XR15-42 Refer to Exercise 15.19.

- a Determine the standard error of estimate and interpret its value.
- b Can we conclude at the 5% significance level that the length of the commercial and people's memories of it are linearly related?
- c Determine the coefficient of determination. What does this statistic tell you about the regression line?

Sample statistics: $n = 60$;
 $\bar{x}_{\text{Length}} = 38.0$; $\bar{y}_{\text{Test}} = 13.8$;
 $s_{xy} = 51.864$; $s_x^2 = 193.898$; $s_y^2 = 47.959$.

15.43 XR15-43 Refer to Exercise 15.20.

- a Determine the standard error of estimate and interpret its value.
- b Test at the 10% significance level to determine whether there is evidence of a positive linear relationship between study time and the final mark.
- c Determine the coefficient of determination. What does this statistic tell you about the regression line?

Sample statistics: $n = 100$;
 $\bar{x}_{\text{Study time}} = 27.95$; $\bar{y}_{\text{Final mark}} = 74.06$;
 $s_{xy} = 153.950$; $s_x^2 = 82.007$; $s_y^2 = 363.939$.

15.44 XR15-44 Refer to Exercise 15.21.

- a Determine the standard error of estimate and describe what this statistic tells you about the regression line.
- b Can we conclude at the 1% significance level that the heights of fathers and sons are linearly related?
- c Determine the coefficient of determination and discuss what its value tells you about the two variables.

Sample statistics: $n = 400$;
 $\bar{x}_{\text{Father}} = 167.86$; $\bar{y}_{\text{Son}} = 171.74$;
 $s_{xy} = 49.188$; $s_x^2 = 102.707$; $s_y^2 = 88.383$.

15.45 XR15-45 Refer to Exercise 15.22. Use whatever statistics you think useful to describe the strength of a linear relationship between medical expenses and age. Test whether the relationship is significantly positive.

Sample statistics: $n = 1348$;
 $\bar{x}_{\text{Age}} = 56.0$; $\bar{y}_{\text{Expense}} = 6.67$;
 $s_{xy} = 51.525$; $s_x^2 = 228.256$; $s_y^2 = 179.881$.

15.46 XR15-46 Doctors often recommend exercise for their patients, particularly those who are overweight.

One benefit of regular exercise appears to be a reduction in cholesterol, a substance associated with heart disease. In order to study the relationship more carefully, a doctor took a random sample of 50 patients who did not exercise. He measured their cholesterol levels and then started them on regular exercise programs. After four months, he asked each patient how many minutes per week (on average) he or she exercised and measured their cholesterol levels. The results are recorded (column 1 = weekly exercise in minutes; column 2 = cholesterol level before exercise program; and column 3 = cholesterol level after exercise program).

- a Determine the regression line that relates exercise time with cholesterol reduction.
- b Interpret the coefficients.
- c Can we conclude at the 5% significance level that the amount of exercise is positively and linearly related to cholesterol reduction?
- d Measure how well the linear model fits.

Sample statistics: $n = 50$,
 $\bar{x}_{\text{Exercise}} = 283.14$; $\bar{y}_{\text{Reduction}} = 27.8$;
 $s_{xy} = 1240.592$; $s_x^2 = 13\,641.31$; $s_y^2 = 221.429$.

15.47 XR15-47 An economist wanted to investigate the relationship between office vacancy rate and rent.

Accordingly, he took a random sample of monthly office rents per square metre and the percentage of vacant office space in 30 different cities.

- a Determine the regression line.
- b Interpret the coefficients.
- c Can we conclude at the 5% significance level that higher vacancy rates result in lower rents?
- d Measure how well the linear model fits the data. Discuss what this measure tells you.

- 15.48 XR15-48** Refer to Exercise 15.26. Is there evidence of a linear relationship between number of cigarettes smoked and number of days absent at the 5% significance level?

Sample statistics: $n = 231$;
 $\bar{x}_{\text{Cigarettes}} = 37.64$; $\bar{y}_{\text{Days}} = 14.43$;
 $s_x^2 = 108.3$; $s_y^2 = 19.8$; $s_{xy} = 20.55$.

15.5 Using the regression equation

Using the techniques in Section 15.4, we can assess how well the linear model fits the data. If the model fits satisfactorily, we can use it to forecast and estimate values of the dependent variable. To illustrate, suppose that in Example 15.3, the used-car dealer wanted to predict the selling price of a five-year-old Ford Laser with 40 000 km on the odometer. Using the regression equation, with $x = 40$, we get

$$\hat{y} = 19.61 - 0.0937x = 19.61 - 0.0937(40) = 15.862$$

We call this value the point prediction. Thus, the dealer would predict that a car with 40 000 km on the odometer would sell for \$15 862.

By itself, however, this value does not provide any information about how closely the value will match the true selling price. To discover that information, we must use a confidence interval. In fact, we can use one of two intervals: the **prediction interval** (of a particular value of y) or the **confidence interval estimate** (of the expected value of y).

15.5a Predicting the particular value of y for a given $x = x_g$

The first interval we present is used whenever we want to predict one particular value of the dependent variable, given a specific value of the independent variable. This confidence interval, often called the *prediction interval*, is calculated in the usual way (point estimator \pm bound on the error of estimation). Here the point estimate for y is \hat{y} and the bound on the error of estimation is shown below.

prediction interval
The confidence interval for a predicted value of the dependent variable for a given value of the independent variable.

Prediction interval of a particular value of y for a given $x = x_g$

$$\hat{y} \pm t_{\alpha/2, n-2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)s_x^2}}$$

where x_g is the given value of x and $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_g$

15.5b Estimating the expected value of y for a given $x = x_g$

The conditions described in Section 15.3 imply that for a given value of x ($= x_g$), there is a population of values of y whose mean is

$$E(y) = \beta_0 + \beta_1 x_g$$

To estimate the mean of y or the long-run average value of y , we would use the following interval referred to simply as the confidence interval. Again, the point estimator is \hat{y} but the bound on the error of estimation is different from the prediction interval shown below.

Confidence interval estimator of the expected value of y for a given $x = x_g$

$$\hat{y} + t_{\alpha/2, n-2} s_e \sqrt{\frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)s_x^2}}$$

Unlike the formula for the prediction interval described above, this formula does not include the 1 under the square-root sign. As a result, the confidence interval estimate of the expected value of y will be narrower than the prediction interval for a particular value of y for the same given value of x and confidence level. This is reasonable, given that predicting a single value is more difficult than estimating the average of a population of values.

EXAMPLE 15.8

LO7

Predicting the price and estimating the mean price of used cars

Refer to Example 15.3.

A used-car dealer is about to bid on a five-year-old Ford Laser equipped with automatic transmission, air conditioner and GPS, and with 40 000 km ($x_g = 40$) on the odometer. To help him decide how much to bid, he needs to predict the selling price.

The used-car dealer alluded to above also has an opportunity to bid on a lot of cars offered by a rental company. The rental company has 250 five-year-old Ford Lasers, all equipped with automatic transmission, air conditioning and GPS. All of the cars in this lot have about 40 000 km ($x_g = 40$) on the odometer. The dealer would like an estimate of the average selling price of all the cars in the lot.

Solution

Identifying the technique

The dealer would like to predict the selling price of a *single* car. Thus, he needs to employ the prediction interval.

$$\hat{y} \pm t_{\alpha/2, n-2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)s_x^2}}$$

The dealer also wants to determine the mean price of a large lot of cars, so he needs to calculate the confidence interval estimate of the expected value.

$$\hat{y} \pm t_{\alpha/2, n-2} s_e \sqrt{\frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)s_x^2}}$$

Technically, this formula is used for infinitely large populations. However, we can interpret our problem as attempting to determine the average selling price of *all* Ford Lasers equipped as described above and with 40 000 km on the odometer. The critical factor is the need to estimate the mean price of a number of cars.

We arbitrarily select a 95% confidence level.

Calculating manually

From previous calculations, we have the following:

$$\hat{y} = 19.61 - 0.0937x_g = 19.61 - 0.0937(40) = 15.862$$

$$s_e = 0.4526, s_x^2 = 43.509 \text{ and } \bar{x} = 36.01$$

From Table 4 in Appendix B, we find

$$t_{0.025, 98} \approx 1.984$$

The 95% prediction interval for the selling price of a single car with an odometer reading of 40 000 km is

$$\begin{aligned}\hat{y} &\pm t_{\alpha/2, n-2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)s_x^2}} \\ &= 15.862 \pm 1.984 \times 0.4526 \sqrt{1 + \frac{1}{100} + \frac{(40 - 36.01)^2}{99 \times 43.509}} \\ &= 15.862 \pm 0.904 \\ &= [14.959, 16.767]\end{aligned}$$

The lower and upper limits of the prediction interval are \$14 959 and \$16 767 respectively.

The 95% confidence interval estimate of the expected selling price of all cars with an odometer reading of 40 000 km is

$$\begin{aligned}\hat{y} &\pm t_{\alpha/2, n-2} s_e \sqrt{\frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)s_x^2}} \\ &= 15.862 \pm 1.984 \times 0.4526 \sqrt{\frac{1}{100} + \frac{(40 - 36.01)^2}{99 \times 43.509}} \\ &= 15.862 \pm 0.105 \\ &= [15.758, 15.968]\end{aligned}$$

The lower and upper limits of the 95% confidence interval estimate of the expected value are \$15 758 and \$15 968 respectively.

Interpreting the results

We predict that one car will sell for between \$14 959 and \$16 768. The average selling price of the population of five-year-old Ford Lasers is estimated to lie between \$15 758 and \$15 968. Because predicting the selling price of one car is more difficult than estimating the mean selling price of all similar cars, the prediction interval is wider than the confidence interval estimate of the expected value.

Using the computer

Using Excel workbook

Excel does not calculate the prediction interval or the confidence interval estimator of the expected value automatically. However, the **Prediction** worksheet in the **Estimators workbook**³ can be used to obtain the prediction interval. The commands follow.

Excel output for Example 15.8

	A	B	C	D	E
1	Predict & Estimate of y				
2					
3	Sample mean of x	36.011	Confidence Interval Estimate of Expected value		
4	Sample variance of x	43.509	15.86	±	0.105
5	Sample size	100	Lower confidence limit		15.758
6	Regression coefficients		Upper confidence limit		15.968
7	Intercept	19.6114			
8	Slope	-0.0937	Prediction Interval		
9	SSE	20.072	15.86	±	0.904
10	Confidence level	0.95	Lower prediction limit		14.959
11	Given value of x	40	Upper prediction limit		16.768

³ This workbook is available under the Workbooks tab on the textbook companion website (accessible through <https://login.cengagebrain.com/>).



COMMANDS

- 1 Calculate the mean and variance of the independent variable x .
- 2 Conduct a regression analysis.
- 3 Open the **Estimators Workbook** and click the **Prediction** tab.
- 4 Input the sample mean and variance of X , the sample size, the regression coefficients β_0 and β_1 , SSE, the confidence level and the given value of X (40).

Using XLSTAT

103	Predictions for the new observations (Price):							
104	Observation	Odometer	Pred(Price)	Std. dev. on pred. (Mean)	Lower bound 95% (Mean)	Upper bound 95% (Mean)	Std. dev. on pred. (Observation)	Lower bound 95% (Observation)
105	PredObs1	40.000	15.863	0.053	15.758	15.968	0.456	14.959

COMMANDS

- 1 Conduct a regression analysis using the data file **XM13-03**. Type the *given value of x* in any empty cell.
- 2 Click **Options** and specify the **Confidence interval (%)** (95).
- 3 Click **Prediction**. In the **X/Explanatory variables** box, check **Quantitative** and type the cell containing the *given value of x*. Click **OK**.

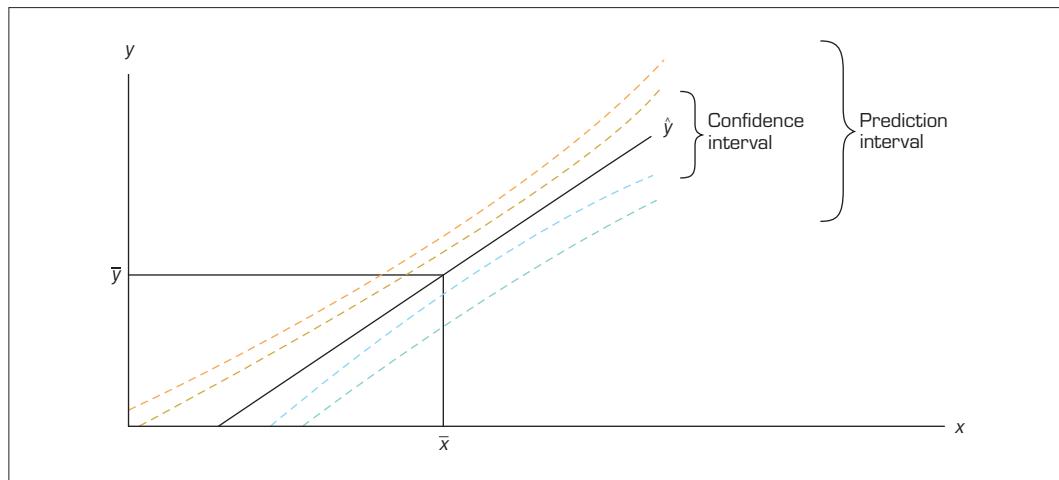
15.5c The effect of the given value of x on the confidence intervals

If the two intervals were calculated for various values of x and graphed, **Figure 15.10** would be produced. Notice that both intervals are represented by curved lines. This is due to the fact that the further the given value of $x = x_g$ is from \bar{x} , the greater the estimation error becomes. This factor is measured by

$$\frac{(x_g - \bar{x})^2}{(n-1)s_x^2}$$

which appears in both the prediction interval and the confidence interval estimate.

FIGURE 15.10 Confidence intervals and prediction intervals



EXERCISES

Learning the techniques

15.49 On the basis of eight observations of x and y , the following summations were calculated:

$$\sum x = 40, \quad \sum y = 100, \quad \sum xy = 600 \\ \sum x^2 = 250, \quad \sum y^2 = 2000$$

- a Determine the 95% confidence interval for the expected value of y when $x = 6$.
- b Determine a 95% prediction interval for the value of y when $x = 3$.

15.50 Consider the following statistics based on 25 observations:

$$\hat{\beta}_0 = 6.5, \quad \hat{\beta}_1 = -2.0, \quad s_e = 5.0, \quad s_x^2 = 4.167, \quad \bar{x} = 25.0$$

- a Find the 90% prediction interval for the value of y when $x = 22.0$.
- b Find the 99% confidence interval estimate of the expected value of y when $x = 27.0$.

15.51 What happens to the prediction interval and the confidence interval when $SSE = 0$?

Applying the techniques

15.52 Self-correcting exercise. For Exercise 15.14, predict with 90% confidence the sales when advertising expenditure equals \$5000.

15.53 XR15-53 The different interest rates charged by some financial institutions may reflect how stringent their standards are for their loan appraisals: the lower the rate, the higher the standards (and hence, the lower the default rate). The data below were collected from a sample of nine financial companies selected at random.

Interest rate (%) x	Default rate (per 1000 loans) y
7.0	38
6.6	40
6.0	35
8.5	46
8.0	48
7.5	39
6.5	36
7.0	37
8.0	44
$\sum x = 65.1$	$\sum y = 363$
$\sum xy = 2652.5$	$\sum x^2 = 476.31$
	$\sum y^2 = 14811$

- a Find the least squares regression line.
- b Do these data provide sufficient evidence to indicate that there is a positive linear relationship between the interest rate and the default rate? (Use $\alpha = 0.10$.)
- c Calculate the coefficients of correlation and determination.
- d Find a 95% prediction interval for the default rate when the interest rate is 8%.

Computer/manual applications

The following exercises require the use of a computer and software.

15.54 XR15-54 Refer to Exercise 15.19.

- a Predict with 95% confidence the memory test score of a viewer who watched a 36-second commercial.
- b Estimate with 95% confidence the mean memory test score of people who watch 36-second commercials.

15.55 XR15-55 Refer to Exercise 15.20.

- a Predict with 90% confidence the final mark of a student who studied for 25 hours.
- b Estimate with 90% confidence the average mark of all students who study for 25 hours.

15.56 XR15-56 Refer to Exercise 15.21. A statistics practitioner wants to produce an interval estimate of the height of a man whose father is 180 cm tall. What formula should be used? Produce such an interval using a confidence level of 99%.

15.57 XR15-57 Refer to Exercise 15.22.

- a Predict with 95% confidence the daily medical expense of an average 65-year-old Australian.
- b Estimate with 95% confidence the mean daily medical expense of all 65-year-old Australians.

15.58 XR15-58 Refer to Exercise 15.46.

- a Predict with 95% confidence the reduction in cholesterol level of an individual who plans to exercise for 100 minutes per week for a total of four months.
- b An individual whose cholesterol level is 250 is planning to exercise for 250 minutes per week. Predict with 95% confidence his cholesterol level after four months.

15.59 XR15-59 Refer to Exercise 15.46. Predict with 95% confidence the reduction in the cholesterol level of an individual who plans to exercise for 300 minutes per week for a total of four months.

15.60 XR15-60 Refer to Exercise 15.47. Predict with 95% confidence the monthly office rent in a city when the vacancy rate is 10%.

coefficient of correlation (Pearson)

A measurement of the strength and direction of a linear relationship between two numerical variables.

15.6 Testing the coefficient of correlation

In Section 15.4, we noted that the coefficient of determination is the coefficient of correlation squared. When we introduced the coefficient of correlation (also called the **coefficient of correlation (Pearson)**) in Chapter 5, we pointed out that it is used to measure the strength of association between two variables. Why then do we use the coefficient of determination as our measure of the regression model's fit? The answer is that the coefficient of determination is a better measure than the coefficient of correlation because the values of R^2 can be interpreted precisely. That is, R^2 is defined as the proportion of the variation in y that is explained by the variation in x . Except for $r = -1, 0$ and 1 , the coefficient of correlation cannot be interpreted. (When $r = -1$ or 1 , every point falls on the regression line, and when $r = 0$, there is no linear pattern.) However, the coefficient of correlation can be useful in another way. We can use it to test for a linear relationship between two variables.

When we are interested in determining *how* the independent variable affects the dependent variable, we estimate and test the linear regression model. The t -test of the slope β_1 allows us to determine whether a linear relationship actually exists. The test requires that for each value of x, y there is a normally distributed random variable with mean $E(y|x) = \beta_0 + \beta_1x$ and constant standard deviation s_e (see **Figure 15.5**). This condition is required whether the data are experimental or observational.

There are many circumstances in which we are interested only in determining *whether* a linear relationship exists and not the form of the relationship. In some cases, we cannot even identify which variable is the dependent variable and which is the independent variable. In such applications, we can calculate the coefficient of correlation and use it to test for linear association.

As we pointed out in Chapter 5, the population coefficient of correlation is denoted ρ (the Greek letter *rho*). Because ρ is a population parameter, we must estimate its value from the data. The sample coefficient of correlation is denoted r and is defined as follows.

Sample coefficient of correlation

$$r = \frac{s_{xy}}{s_x s_y}$$

15.6a Testing the coefficient of correlation

When there is no linear relationship between two variables, $\rho = 0$. To determine whether we can infer that ρ is zero, we test the following hypotheses:

$$\begin{aligned} H_0: \rho &= 0 \\ H_A: \rho &\neq 0 \end{aligned}$$

The test statistic is defined as

$$t = \frac{r - \rho}{s_T}$$

where

$$s_T = \sqrt{\frac{1-r^2}{n-2}}$$

Test statistic for testing ρ

$$t = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}} \sim t_{n-2}$$

which is Student-*t* distributed, with $v = n - 2$ degrees of freedom, provided that the variables are bivariate normally distributed.

The *t*-test of ρ and the *t*-test of β_1 (in which under the null hypothesis both parameters are set equal to 0) produce identical results. If we applied the *t*-test of ρ to Example 15.3, we would produce the same *t*-statistic, *p*-value and conclusion. Hence, practically speaking, it doesn't matter which test we employ.

EXAMPLE 15.9

LO8

Are odometer reading and price of used cars linearly related? Testing the significance of the coefficient of correlation

Using the data in Example 15.3, test to determine whether we can infer that a linear relationship exists between selling price and odometer reading. Use a significance level of 5%.

Solution

The problem objective is to analyse the relationship between two numerical (quantitative) variables. Because we are not interested in the form of the linear relationship but only in whether a linear relationship exists between the two variables and the data are observational, the parameter of interest is the coefficient of correlation.

We test the following hypotheses:

Hypotheses: $H_0: \rho = 0$
 $H_A: \rho \neq 0$ (Two-tail test)

Test statistic: $t = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}} \sim t_{n-2}$

Level of significance: $\alpha = 0.05$

Decision rule: Reject H_0 if $|t| > t_{\alpha/2, n-2} = t_{0.025, 98} \approx 1.984$

In other words, $t < -1.984$ or $t > 1.984$

Alternatively, reject H_0 if *p*-value $< \alpha = 0.05$

Value of the test statistic:

Calculating manually

The value of r is calculated from the covariance and two standard deviations calculated in Examples 15.3 and 15.4:

$$s_{xy} = -4.077, \quad s_x = 6.596, \quad s_y = 0.765$$

Therefore,

$$r = \frac{s_{xy}}{s_x s_y} = \frac{-4.077}{6.596 \times 0.765} = -0.8083$$

The value of the test statistic is

$$t = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{(-0.8083 - 0)}{\sqrt{\frac{1-(-0.8083)^2}{100-2}}} = -13.59$$

Notice that this is the same value we produced in the *t*-test of the slope in Example 15.6. Because both sampling distributions are Student *t* with 98 degrees of freedom, the *p*-value and conclusion are also identical.

Conclusion: As $t = -13.59 < -1.984$, we reject the null hypothesis. Alternatively, from the Excel output, *p*-value = 0, which is less than $\alpha = 0.05$. Hence we reject H_0 .

Interpreting the results

There is overwhelming evidence to infer that the odometer reading and the price of a used Ford Laser are linearly related.

Using the computer

Using Excel workbook

Excel does not perform this test automatically. However, the **t-Test_Correlation** worksheet in the **Test Statistics workbook** can be used to test the significance of the coefficient of correlation ρ . The output is given below and the Excel commands follow.

Excel output for Example 15.9

	A	B	C	D
1	t-Test of Correlation Coefficient			
2				
3	Sample correlation	0.8083	t Stat	
4	Sample size	100	P(T<=t) one-tail	13.59
5	Alpha	0.05	t Critical one-tail	1.41E-24
6			P(T<=t) two-tail	1.6604
7			t Critical two-tail	2.82E-24
8				1.9842

COMMANDS

- 1 Calculate the coefficient of correlation. (See page 181 for instructions.)
- 2 Open the **Test Statistics Workbook** and click the **t-Test_Correlation** tab.
- 3 Input the coefficient of correlation, the sample size and the value of α .

Using XLSTAT

	A	B	C
8	Correlation matrix (Pearson):		
9			
10	Variables	Price (y)	Odometer (x)
11	Price (y)	1	-0.808
12	Odometer (x)	-0.808	1
13			
14			
15	p-values (Pearson):		
16			
17	Variables	Price (y)	Odometer (x)
18	Price (y)	0	< 0.0001
19	Odometer (x)	< 0.0001	0

COMMANDS

- 1 Type the data or open the data file (**XM15-03**).
- 2 Click **XLSTAT**, **Correlation/Association test**, and **Correlation tests**.
- 3 In the **Observations/variables table** box type the input range (**A1:B101**). Specify **Type of correlation: Pearson**.
- 4 Click **Outputs** and check **Correlations** and **p-values**. Click **OK**.

If you review Example 15.6 in which we tested the slope coefficient β_1 , you will find the same value of the test statistic, the same rejection region and, of course, the same conclusion as we produced above. This is not a coincidence; the two tests are identical. This should be no surprise, since the data and the objective are the same: to determine whether two variables are linearly related. Hence, it is necessary to perform only one test, either the t -test of β_1 or the t -test of ρ . (We performed both tests to show you that they are identical.)

15.6b Violation of the required condition

When the normality requirement is unsatisfied, we can use a nonparametric technique – the Spearman rank correlation coefficient to replace the t -test of ρ .

EXERCISES

Learning the techniques

- 15.61 XR15-61** Given the following observations, determine the Pearson correlation coefficient.

x	0	5	-1	0	3	2	4
y	3	10	2	4	6	5	7

- 15.62** Test each of the following sets of hypotheses:

- a $H_0: \rho = 0$
 $H_A: \rho > 0$
 $r = 0.30 \quad n = 20 \quad \alpha = 0.05$
- b $H_0: \rho = 0$
 $H_A: \rho < 0$
 $r = -0.28 \quad n = 10 \quad \alpha = 0.01$
- c $H_0: \rho = 0$
 $H_A: \rho \neq 0$
 $r = 0.48 \quad n = 18 \quad \alpha = 0.05$

- 15.63 XR15-63** You are given the following data:

x	115	220	86	99	50	110
y	1.0	1.3	0.6	0.8	0.5	0.7

- a Calculate the Pearson correlation coefficient.
- b Test to determine whether we can infer that a linear relationship exists between the two variables. (Use $\alpha = 0.05$.)

Applying the techniques

- 15.64 Self-correcting exercise.** Refer to Exercise 15.19.

- a Determine the coefficient of correlation.
- b Conduct a test at the 5% significance level to determine whether a linear relationship exists between length of commercial and memory test score.

- 15.65 XR15-65** The weekly returns of two shares are recorded for a 13-week period.

- a Determine the coefficient of correlation.
- b Assuming that the returns are normally distributed, can we infer at the 5% significance level that the shares are correlated?

15.7 Regression diagnostics – I

In Section 15.3, we described the conditions required for the validity of regression analysis. Simply put, the error variable must be normally distributed with a constant variance, and the errors must be independent of each other. In this section, we show how to diagnose violations. Additionally, we discuss how to deal with observations that are unusually large or small. Such observations must be investigated to determine if an error was made in recording them.

15.7a Residual analysis

Most departures from required conditions can be diagnosed by examining the residuals, which we discussed in Section 15.2. Most computer packages allow you to output the values of the residuals and apply various graphical and statistical techniques to this variable. If using manual calculations, the residuals for a simple regression model can be calculated as $e_i = y_i - \hat{y}_i$, where y_i is the i th observed value and $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the i th fitted value.

We can also calculate the standardised residuals. We standardise residuals in the same way we standardise all variables, by subtracting the mean and dividing by the standard deviation. The mean of the residuals is zero, and because the standard deviation σ_e is unknown, we must estimate its value. The simplest estimate is the standard error of estimate s_e . Thus:

$$\text{Standardised residual} = \frac{e_i}{s_e}$$

A partial list of this standardised residual calculated in Excel for Example 15.3 is shown below.

Using the computer

Excel calculates the standardised residuals by dividing the residuals by the standard deviation of the residuals. (The difference between the standard error of estimate and the standard deviation of the residuals is that in the formula of the former, the denominator is $n - 2$, whereas in the formula for the latter, the denominator is $n - 1$.)

Part of the Excel output (we show only the first five and last five values) for Example 15.3 is shown.

Using Excel Data Analysis

Excel output of the residuals: Example 15.3

	A	B	C	D
1	Observation	Predicted Price (y)	Residuals	Standard Residuals
2	1	16.1068	-0.1068	-0.2373
3	2	15.4134	-0.2134	-0.4740
4	3	15.3197	-0.3197	-0.7101
5	4	16.7159	0.6841	1.5192
6	5	16.6410	0.7590	1.6857
7
8	96	16.2193	0.0807	0.1792
9	97	16.4067	-0.6067	-1.3474
10	98	16.5004	-0.7004	-1.5555
11	99	15.9382	0.1618	0.3594
12	100	16.2005	-0.8005	-1.7779

COMMANDS

Proceed with the first five steps of regression analysis described on page 635. Then select **Residuals** and **Standardised Residuals** and click **OK**. The predicted values, residuals and standardised residuals will be printed.

Using XLSTAT

	A	B	C	D
1	Observation	Predicted Price (y)	Residuals	Standard Residuals
2	Obs1	16.107	-0.107	-0.236
3	Obs2	15.413	-0.213	-0.472
4	Obs3	15.320	-0.320	-0.706
5	Obs4	16.716	0.684	1.512
6	Obs5	16.641	0.759	1.67
7
8	Obs96	16.219	0.081	0.178
9	Obs97	16.407	-0.607	-1.341
10	Obs98	16.500	-0.700	-1.548
11	Obs99	15.938	0.162	0.358
12	Obs100	16.201	-0.801	-1.769

COMMANDS

- 1 Open the data file (**XM15-03**). Conduct a regression analysis.
- 2 Click **Outputs** and check **Predictions** and **residuals**. Click **OK**.

We can also standardise by calculating the standard deviation of each residual. Statisticians have determined that the standard deviation of the residual at point i is defined as follows.

Standard deviation of the i th residual

$$s_{e_i} = s_e \sqrt{1 - h_i}$$

where

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}$$

The quantity h_i should look familiar; it was used in the formula for the prediction interval and the confidence interval estimate of the expected value of y in Section 15.5. Excel calculates this version of the standardised residuals. This version is more generally accepted. Below we list some of the residuals and standardised residuals for Example 15.3.

Partial list of residuals and standardised residuals from Excel for Example 15.3

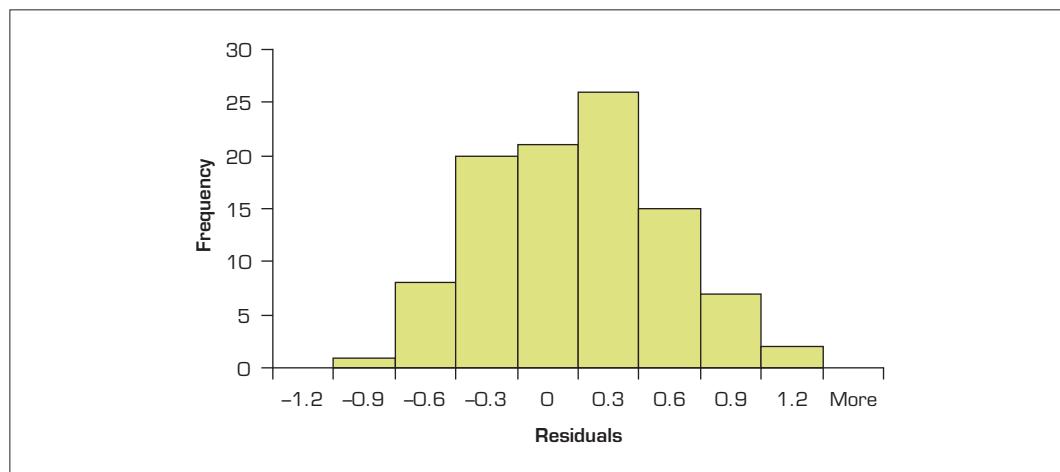
Observation	Residuals	Standardised residuals
1	-0.1068	-0.2373
2	-0.2134	-0.4740
3	-0.3197	-0.7101
.	.	.
99	0.1618	0.3594
100	-0.8006	-1.7779

An analysis of the residuals will allow us to determine if the error variable is non-normal, whether the error variance is constant, and whether the errors are independent. We begin with non-normality.

15.7b Non-normality

In Chapter 8, we introduced the normal probability distribution with the comment that it is the most important distribution in statistics. By now you must agree, if for no other reason than the number of techniques that require normality. As we have done throughout this book, we can check for normality by producing a histogram of the residuals to see if it appears that the error variable is normally distributed. If the histogram appears to at least resemble a bell shape, it is probably safe to assume the normality requirement has been met.

Excel histogram of residuals: Example 15.3



Excel's histogram of the residuals for Example 15.3 is shown above. The histogram suggests that the error variable is approximately normally distributed. It should be noted that the tests applied in regression analysis are robust, which means that only when the error variable is quite non-normal are the test results called into question.

15.7c Heteroscedasticity

heteroscedasticity

The condition under which the variance of the error variables is not constant.

homoscedasticity

The condition under which the variance of the error variables is constant.

The variance of the error variable σ_e^2 is required to be constant. When this requirement is violated, the condition is called **heteroscedasticity**. (You can impress friends and relatives by using this term. If you can't pronounce it, try **homoscedasticity**, which refers to the condition in which the requirement is satisfied.) One method of diagnosing heteroscedasticity is to plot the residuals against the predicted values of y , \hat{y} . We then look for a change in the spread of variation of the plotted points. **Figure 15.11** describes such a situation. Notice that, in this illustration, σ_e^2 appears to be small when \hat{y} is small, and large when \hat{y} is large. Of course, many other patterns could be used to depict this problem.

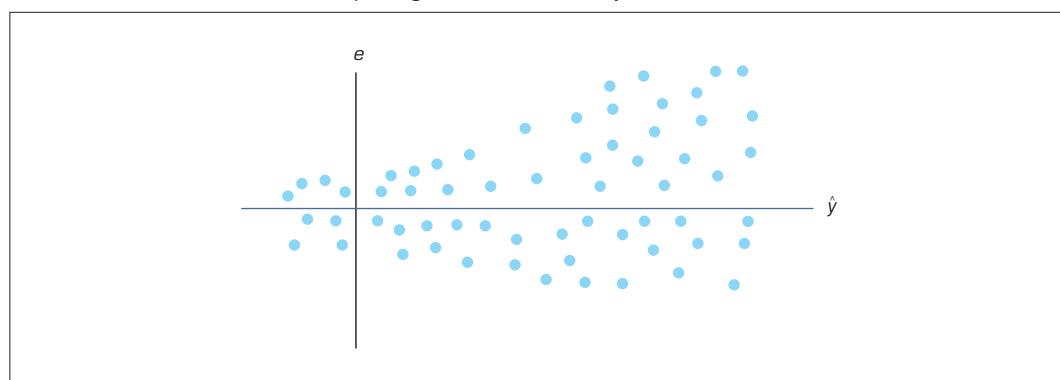
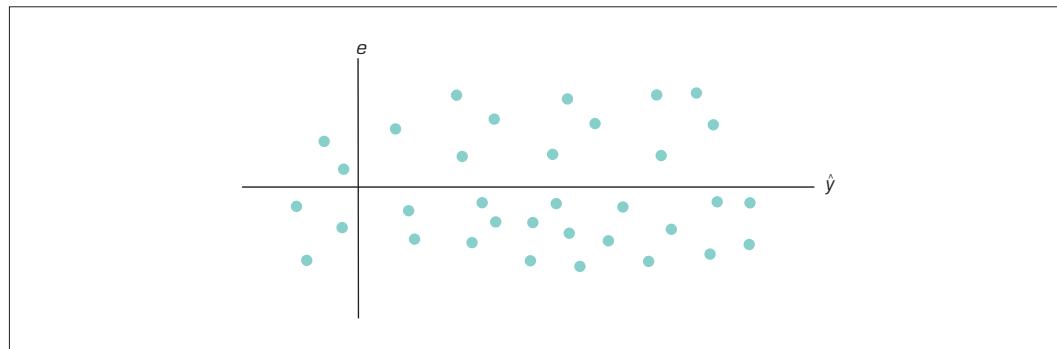
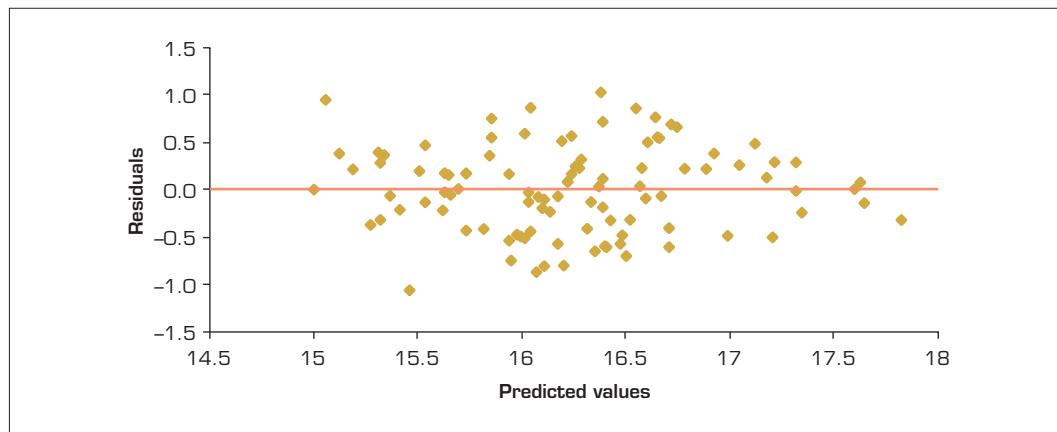
FIGURE 15.11 Plot of residuals depicting heteroscedasticity

Figure 15.12 illustrates a case in which σ_e^2 is constant. As a result, there is no apparent change in the variation of the residuals. Excel's plot of the residuals e_i versus the predicted values of y , \hat{y} , for Example 15.3 is shown (**Figure 15.13**). There does appear to be a decrease in the variance for larger values of \hat{y} . However, it is far from clear that there is a problem here.

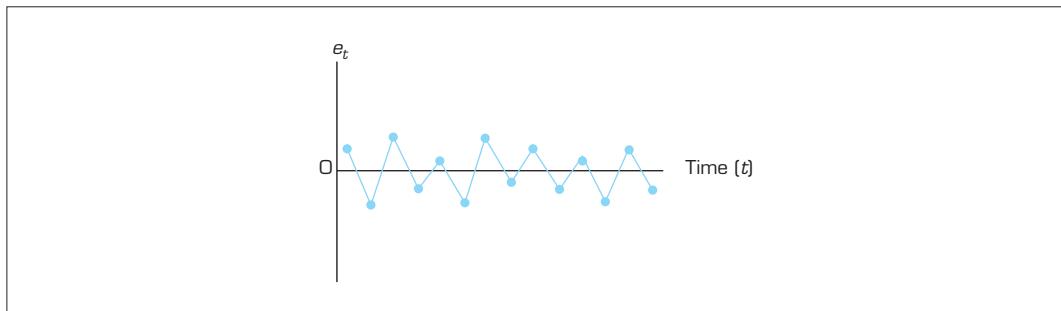
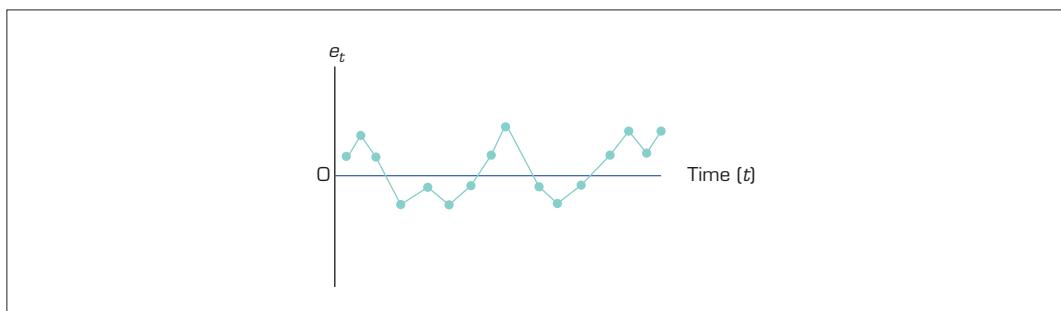
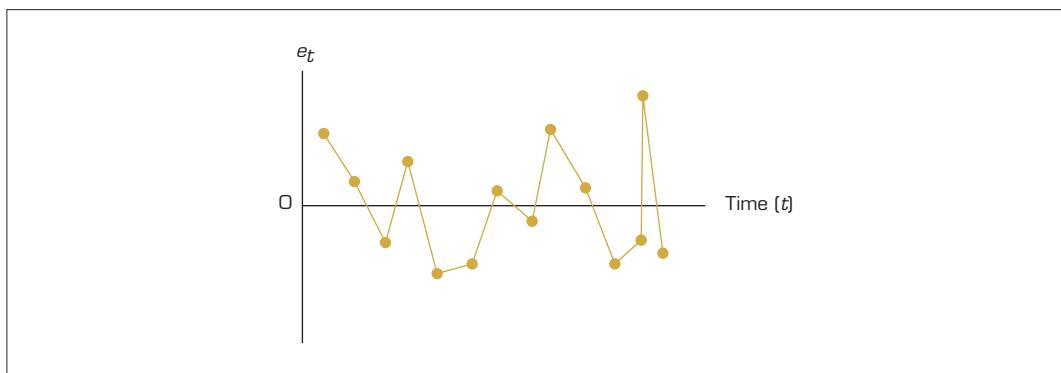
FIGURE 15.12 Plot of residuals depicting homoscedasticity**FIGURE 15.13** Excel plot of residuals versus predicted values: Example 15.3

15.7d Non-independence of the error variable

In Chapter 2, we briefly described the difference between cross-sectional and time-series data. Cross-sectional data are observations made at approximately the same time, whereas a time series is a set of observations taken at successive points of time. The data in Example 15.3 are cross-sectional because all of the prices and odometer readings were taken at about the same time. If we were to observe the auction price of cars every week for (say) a year that would constitute a time series.

Condition 4 states that the values of the error variable are independent. When the data are time series, the errors often are correlated. Error terms that are correlated over time are said to be *autocorrelated* or serially correlated. For example, suppose that, in an analysis of the relationship between annual gross profits and some independent variable, we observe the gross profits for the years 1997 to 2016. The observed values of y are denoted y_1, y_2, \dots, y_{20} , where y_1 is the gross profit for 1997, y_2 is the gross profit for 1998 and so on. If we label the residuals e_1, e_2, \dots, e_{20} , then – if the independence requirement is satisfied – there should be no relationship among the residuals. However, if the residuals are related, it is likely that autocorrelation exists.

We can often detect autocorrelation by graphing the residuals against the time periods. If a pattern emerges, it is likely that the independence requirement is violated. **Figures 15.14** and **15.15** exhibit patterns indicating negative (the sign of consecutive residuals changes very frequently) and positive autocorrelation (the sign of consecutive residuals remains the same mostly for some time before the sign changes), respectively, while **Figure 15.16** exhibits no pattern and, thus, likely represents independent errors.

FIGURE 15.14 Plot of residuals versus time indicating negative autocorrelation (alternating)**FIGURE 15.15** Plot of residuals versus time indicating positive autocorrelation (increasing)**FIGURE 15.16** Plot of residuals versus time indicating independence (or no autocorrelation)

In Chapter 16, we introduce the Durbin–Watson test, which is another statistical test to determine if one form of this problem is present.

We will describe a number of remedies to violations of the required conditions in Chapter 16.

15.7e Outliers

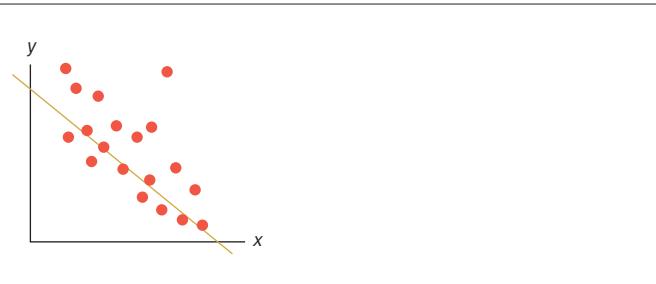
An *outlier* is an observation that is unusually small or unusually large. To illustrate, consider Example 15.3, where the range of odometer readings was 19 100 to 49 200 km. If we had observed a value of 5000 km, we would identify that point as an outlier. There are several possibilities that we need to investigate.

- 1 There was an error in recording the value. To detect an error we would check the point or points in question. In Example 15.3, we could check the car's odometer to determine if a mistake was made. If so, we would correct it before proceeding with the regression analysis.

- 2 The point should not have been included in the sample. Occasionally, measurements are taken from experimental units that do not belong with the sample. We can check to ensure that the car with the 5000km odometer reading was actually five years old. We should also investigate the possibility that the odometer had been rolled back. In either case, the outlier should be discarded.
- 3 The observation was simply an unusually large or small value that belongs to the sample and that was recorded properly. In this case we would do nothing to the outlier. It would be judged to be valid.

Outliers can be identified from the scatter diagram. **Figure 15.17** depicts a scatter diagram with one outlier. The statistics practitioner should check to determine if the measurement was recorded accurately and whether the experimental unit should be included in the sample.

FIGURE 15.17 Scatter diagram with one outlier



The standardised residuals also can be helpful in identifying outliers. Large absolute values of the standardised residuals should be investigated thoroughly.

15.7f Influential observations

Occasionally, in a regression analysis, one or more observations have a large influence on the statistics. **Figure 15.18** describes such an observation and the resulting least squares line. If the point had not been included, the least squares line in **Figure 15.19** would have been produced. Obviously, one point has had an enormous influence on the results. Influential points can be identified by the scatter diagram. The point may be an outlier, and as such must be investigated thoroughly. Observations with large standardised residuals and observations whose values of x give them a large influence are considered as unusual observations in the sample. These are points to be checked for accuracy and to ensure that the observations belong to the sample.

FIGURE 15.18 Scatter diagram with one influential observation

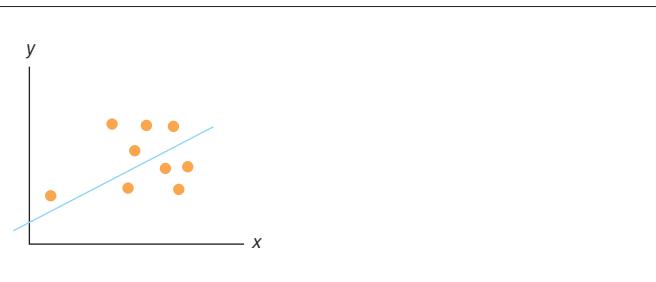
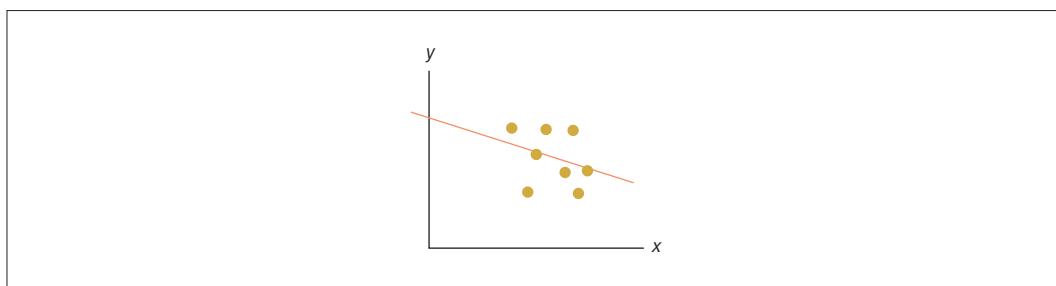


FIGURE 15.19 Scatter diagram without the influential observation

List of unusual observations

	A	B	C	D	E
1	<i>Observation</i>	<i>Price</i>	<i>Odometer</i>	<i>Residual</i>	<i>Standardised residual</i>
2		14	17.4	32.7	1.0214
3		19	16.5	35.8	0.9426
4		78	16.5	34.4	-1.0603

Points 14, 19 and 78 have standardised residuals that are greater than $|2|$, and so are judged to be unusual.

	A	B	C	D	E
1	<i>Observation</i>	<i>Price</i>	<i>Odometer</i>	<i>Residual</i>	<i>Standardised residual</i>
2		8	17.5	19.1	-0.3216
3		63	17.7	21.2	0.0751
4		74	17.5	21.0	-0.1436
5		86	17.6	21.5	0.0033

Notice that points 8, 63, 74 and 86 are identified as points that have a large influence. These points are the four smallest odometer readings. Their removal would change the regression equation in a substantive way.

15.7g Procedure for regression diagnostics

The order of the material presented in this chapter is dictated by pedagogical requirements. Consequently, we presented the least squares method, methods of assessing the model's fit, predicting and estimating using the regression equation, coefficients of correlation and, finally, the regression diagnostics. In a practical application, the regression diagnostics would be conducted earlier in the process.

It is appropriate to investigate violations of the required conditions when the model is assessed and before using the regression equation to predict and estimate. The following steps describe the entire process.

- 1 Develop a model that has a theoretical basis. That is, for the dependent variable in question find an independent variable to which you believe it is linearly related.
- 2 Gather data for the two variables. Ideally, conduct a controlled experiment. If that is not possible, collect observational data.
- 3 Draw the scatter diagram to determine whether a linear model appears to be appropriate. Identify possible outliers.
- 4 Determine the regression equation.
- 5 Calculate the residuals and check the required conditions.

Is the error variable non-normal?

Is the error variance constant?

Are the errors independent?

Check the outliers and influential observations.

6 Assess the model's fit.

Calculate the standard error of estimate.

Test to determine whether there is a linear relationship. (Test β_1 or ρ .)

Calculate the coefficient of determination.

7 If the model fits the data, use the regression equation to predict a particular value of the dependent variable and/or estimate its mean.

EXERCISES

Learning the techniques

- 15.66 XR15-66** Observations of two variables were recorded as shown below.

x	-5	-2	0	3	4	7
y	15	9	7	6	4	1

- a Determine the regression equation.
- b Use the regression equation to determine the predicted values of y .
- c Use the predicted and actual values of y to calculate the residuals.
- d Calculate the standardised residuals.
- e Identify possible outliers.

- 15.67 XR15-67** Observations of two variables were recorded as shown below.

x	1	2	3	4	5	6	7	8	9
y	5	28	17	14	27	33	39	26	30

- a Calculate the regression equation.
- b Use the regression equation to determine the predicted values of y .
- c Use the predicted and actual values of y to calculate the residuals.
- d Calculate the standardised residuals.
- e Identify possible outliers.
- f Plot the residuals against the predicted values of y . Does the variance appear to be constant? Explain.

- 15.68 XR15-68** Each of the following pairs of values represents an actual value of y and a predicted value of y (based on a simple regression model). Graph the predicted values of y (on the horizontal axis) versus the residuals (on the vertical axis). In each case, determine from the graph whether the requirement that the variance of the error variable be constant is satisfied.

a	y	\hat{y}	b	y	\hat{y}	c	y	\hat{y}
	155	143		10	7		46	48
	112	108		22	21		40	43
	163	180		29	29		53	54
	130	133		15	13		60	63
	143	146		24	25		56	54
	182	193		13	16		62	65
	160	140		17	19		44	46
	104	101		23	22		49	47
	125	126		11	14		52	49
	161	176		27	27		59	56
	189	200		19	17		45	41
	102	97		26	27		55	53
	142	145		20	22		47	44
	149	151		14	11		61	57
	180	158					42	45
							57	62
							50	51

Applying the techniques

- 15.69 XR15-69 Self-correcting exercise.** Refer to Exercise 15.10.

- a Use the regression equation you produced to determine the predicted values of y .
- b Use the predicted and actual values of y to calculate the residuals.
- c Calculate the standardised residuals.
- d Identify possible outliers.
- e Plot the residuals against the predicted values of y . Does the variance appear to be constant? Explain.

- 15.70 XR15-70** Refer to Exercise 15.15.

- a Use the regression equation you produced to determine the predicted values of y .
- b Use the predicted and actual values of y to calculate the residuals.
- c Calculate the standardised residuals.

- d** Identify possible outliers.
- e** Plot the residuals against the predicted values of y . Does the variance appear to be constant? Explain.

Computer applications

The following exercises require the use of a computer and software.

15.71 XR15-71 Refer to Exercise 15.19.

- a** Determine the residuals and the standardised residuals.
- b** Draw the histogram of the residuals. Does it appear that the errors are normally distributed? Explain.
- c** Identify possible outliers.
- d** Plot the residuals versus the predicted values of y . Does it appear that heteroscedasticity is a problem? Explain.

15.72 XR15-72 Refer to Exercise 15.20.

- a** Determine the residuals and the standardised residuals.
- b** Draw the histogram of the residuals. Does it appear that the errors are normally distributed? Explain.

- c** Identify possible outliers.
- d** Plot the residuals versus the predicted values of y . Does it appear that heteroscedasticity is a problem? Explain.

15.73 XR15-73 Refer to Exercise 15.21.

- a** Determine the residuals and the standardised residuals.
- b** Draw the histogram of the residuals. Does it appear that the errors are normally distributed? Explain.
- c** Identify possible outliers.
- d** Plot the residuals versus the predicted values of y . Does it appear that heteroscedasticity is a problem? Explain.

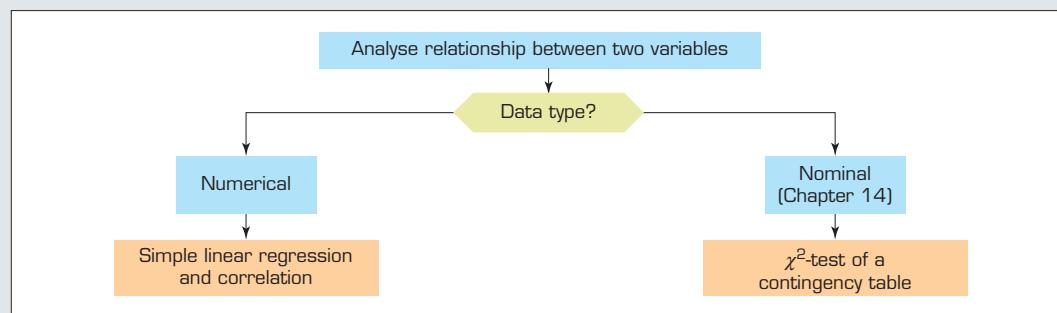
15.74 XR15-74 Refer to Exercise 15.22.

- a** Determine the residuals and the standardised residuals.
- b** Draw the histogram of the residuals. Does it appear that the errors are normally distributed? Explain.
- c** Identify possible outliers.
- d** Plot the residuals versus the predicted values of y . Does it appear that heteroscedasticity is a problem? Explain.

Study Tools

CHAPTER SUMMARY

Simple linear regression and correlation are techniques for analysing the relationship between two numerical variables. Regression analysis assumes that the two variables are linearly related. The *least squares method* produces estimates of the *y*-intercept and the *slope* of the regression line. Considerable effort is expended in assessing how well the linear model fits the data. We calculate the *standard error of estimate*, which is an estimate of the standard deviation of the error variable. We test the slope to determine whether there is sufficient evidence of a linear relationship. The strength of the linear association is measured by the *coefficient of correlation* and the *coefficient of determination*. When the model provides a good fit, we can use it to predict the particular value and to estimate the expected value of the dependent variable. We can also use the *Pearson correlation coefficient* to measure and test the linear relationship between two bivariate normally distributed variables. We completed this chapter with a discussion of how to diagnose violations of the required conditions.



The data files for Examples, Exercises and Cases are provided on the companion website (accessible through <https://login.cengagebrain.com/>).

SYMBOLS

Symbol	Pronounced	Represents
β_0	Beta-sub-zero or beta-zero	<i>y</i> -intercept coefficient
β_1	Beta-sub-one or beta-one	Slope coefficient
ε	Epsilon	Error variable
\hat{y}	<i>y</i> -hat	Fitted or calculated value of <i>y</i>
$\hat{\beta}_0$	Beta-hat-sub-zero or beta-zero-hat	Sample <i>y</i> -intercept coefficient
$\hat{\beta}_1$	Beta-hat-sub-one or beta-one-hat	Sample slope coefficient
σ_ε	Sigma-sub-epsilon or sigma-epsilon	Standard deviation of error variable
s_e	<i>s</i> -sub- <i>e</i> or <i>s</i> - <i>e</i>	Standard error of estimate
$s_{\hat{\beta}_0}$	<i>s</i> -sub-beta-hat-sub-zero or <i>s</i> -beta-zero-hat	Standard error of $\hat{\beta}_0$
$s_{\hat{\beta}_1}$	<i>s</i> -sub-beta-hat-sub-one or <i>s</i> -beta-one-hat	Standard error of $\hat{\beta}_1$
R^2	<i>R</i> -squared	Coefficient of determination
x_g	<i>x</i> -sub- <i>g</i> or <i>x</i> - <i>g</i>	Given value of <i>x</i>
ρ	<i>rho</i>	Pearson coefficient of correlation
r	<i>r</i>	Sample coefficient of correlation
e_i	<i>e</i> -sub- <i>i</i> or <i>e</i> - <i>i</i>	Residual of <i>i</i> th point

SUMMARY OF FORMULAS

Variance of x	$s_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = \frac{1}{n-1} \left(\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right)$
Variance of y	$s_y^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2 = \frac{1}{n-1} \left(\sum y_i^2 - \frac{(\sum y_i)^2}{n} \right)$
Covariance between x and y	$s_{xy} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} \left(\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} \right)$
Slope coefficient estimate	$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$
y -intercept estimate	$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
Residual	$e_i = y_i - \hat{y}_i$
Sum of squares for error	$SSE = \sum e_i^2 = (n-1) \left[s_y^2 - \frac{s_{xy}^2}{s_x^2} \right]$
Standard error of estimate	$s_e = \sqrt{\frac{SSE}{n-2}}$
Standard error of $\hat{\beta}_1$ and $\hat{\beta}_0$	$s_{\hat{\beta}_1} = \frac{s_e}{\sqrt{(n-1)s_x^2}}, s_{\hat{\beta}_0} = \frac{s_e \sqrt{\sum x_i^2}}{\sqrt{(n-1)s_x^2}} = s_{\hat{\beta}_1} \sqrt{\frac{\sum x_i^2}{n}}$
Test statistic for the slope coefficient β_1	$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$
Sample coefficient of correlation	$r = \frac{s_{xy}}{s_x s_y}$
Test statistic for testing $\rho = 0$	$t = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}}$
Coefficient of determination	$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{SSE}{(n-1)s_y^2} = \frac{s_{xy}^2}{s_x^2 s_y^2}$
Predicted value of y	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
Prediction interval of a particular value of y	$\hat{y} \pm t_{\alpha/2, n-2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)s_x^2}}$
Confidence interval estimator of the expected value of y	$\hat{y} \pm t_{\alpha/2, n-2} s_e \sqrt{\frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)s_x^2}}$

SUPPLEMENTARY EXERCISES

15.75 XR15-75 A computer dating service typically asks for various pieces of information such as height, weight, income, and so on. One such service requested the length of index fingers. The only plausible reason for this request is to act as a proxy on height. Women have often complained that men lie about their height. If there is a strong relationship between height and length of index finger, the information can be used to 'correct' false claims about height. To test the relationship between the two variables, researchers gathered the height and length of index finger (in centimetres) of 121 students.

- a Graph the relationship between the two variables.
- b Estimate a linear relationship between height and length of index finger.
- c Is there sufficient evidence to infer that height and length of index finger are linearly related?
- d Predict with 95% confidence the height of someone whose index finger is 6.5 cm long. Is this prediction likely to be useful? Explain.

(The authors would like to thank Howard Waner for supplying the problem and data.)

15.76 XR15-76 The store manager of Colonial Furniture Ltd has been reviewing quarterly advertising expenditures. Television ads in particular caught her eye, because they were the main expenditure item. In order to maximise cost effectiveness, she would like to get a better idea of the relationship between the television advertisements the company sponsors and the number of people who visit the store because of them. To this end, she has compiled and recorded the data of which some are presented in the following table:

Quarterly number of television ads x	Quarterly number of people y	
7	42	
5	32	
.	.	
.	.	
6	37	
5	33	
$\Sigma x = 96$	$\Sigma y = 559$	$n = 16$
$\Sigma x^2 = 676$	$\Sigma y^2 = 23037$	$\Sigma xy = 3930$

- a Find the sample regression line that expresses the number of people coming to the store as a function of the number of television ads run.
- b Is there enough evidence to allow the manager to conclude that a linear relationship exists between the two variables? (Use $\alpha = 0.10$.)
- c What proportion of the variability in the number of people coming into the store is explained by the variability in the number of television ads?
- d Find a 99% prediction interval for the number of people entering the store if the store manager intends to sponsor five television ads this quarter.

15.77 XR15-77 In recent years, fishermen have suffered financial hardship because of shortened fishing seasons, reduced catches and lower market prices. Moreover, fishermen have complained about price fluctuations and have called for a system of minimum prices. One suggestion made was that the size of the catch had an immediate impact on prices, and that this relationship should be clarified before potential solutions were discussed. To investigate this issue, a random 12-week period was selected to study the price of fish versus the average daily catch. The data were collected for analysis and recorded. Part of the data is presented below.

Average daily catch x ('00 kg)	Price per kg (\$) y	
357	1.95	
618	1.05	
.	.	
.	.	
695	1.10	
710	1.05	
$\Sigma x = 6970$	$\Sigma y = 16.45$	$n = 12$
$\Sigma x^2 = 4315894$	$\Sigma y^2 = 24.94$	$\Sigma xy = 8972.8$

- a Determine the sample regression line that shows the price per kilogram as a function of average daily catch.
- b Calculate the standard error of estimate. What does this value tell you about the relationship between the two variables?
- c Do these data provide sufficient evidence at the 5% significance level to allow you to conclude

- that large catches result in lower prices?
- d Calculate the coefficient of determination. What does this value tell you about the relationship between the two variables?
- e Find a 90% confidence interval estimate for the expected value of the price per kilogram if the daily catch is 75000 kg.

15.78 XR15-78 The head office of a life insurance company believed that regional managers should have weekly meetings with their salespeople, not only to keep them abreast of current market trends but also to provide them with important facts and figures that would help them with their sales. Furthermore, the company felt that these meetings should be used for pep talks. One of the points management felt strongly about was the high value of new contact initiation and follow-up phone calls. To dramatise the importance of phone calls on prospective clients and (ultimately) on sales, the company undertook the following small study. Twenty randomly selected life insurance salespeople were surveyed to determine the number of weekly sales calls they made and the number of policy sales they concluded. The data partly shown in the following table were collected and stored.

Weekly calls x	Weekly sales y	
66	20	
43	15	
.	.	
.	.	
51	17	
44	14	
$\Sigma x = 902$	$\Sigma y = 270$	$n = 20$
$\Sigma x^2 = 44318$	$\Sigma y^2 = 4120$	$\Sigma xy = 13432$

- a Find the least squares regression line that expresses the number of sales as a function of the number of calls.
- b What do the coefficients tell you?
- c Is there enough evidence (with $\alpha = 0.05$) to indicate that the larger the number of calls, the larger the number of sales?
- d What proportion of the variability in the number of sales can be attributed to the variability in the number of calls?
- e Find a 90% confidence interval estimate of the mean number of sales made by all the salespeople who each make 50 calls.

- f Predict with 99% confidence the number of sales concluded by a salesperson who makes 30 calls.

15.79 XR15-79 An agriculture student at the University of New England pulled from his father's farm records some data relating crop yield to the amount of fertiliser used, the mean seasonal rainfall, the mean number of hours of sunshine and the mean daily temperature. As a first approximation, he wishes to regress crop yield on the amount of fertiliser used. Some of the data are provided in the following table.

Fertiliser (kg/hectare) x	Crop yield ('000kg/hectare) y	
220	36	
450	72	
.	.	
.	.	
410	79	
450	75	
$\Sigma x = 4900$	$\Sigma y = 825$	$n = 14$
$\Sigma x^2 = 1825600$	$\Sigma y^2 = 51891$	$\Sigma xy = 307190$

- a Graph the data and comment on the suitability of using a linear regression model.
- b Find the least squares regression line for these data.
- c Test to determine whether a linear relationship exists between the two variables. (Use $\alpha = 0.05$.)
- d Calculate the coefficient of determination and interpret its value.
- e Forecast the crop yield, with 99% confidence, based on using 500kg of fertiliser. How does this compare to the actual yield when 500kg of fertiliser was used?

15.80 XR15-80 In response to the complaints of both students and parents about the high cost of school materials for Year 9–12 students, the school council attempted to keep track of the cost. It selected three students at random from Year 9 and tracked the cost of their school materials over the four years (Years 9–12), with the results shown in the following table.

Year	Student		
	1	2	3
9	415	400	410
10	410	405	420
11	410	420	435
12	425	425	415

- a Find the equation of the regression line.
- b Do these data provide enough evidence to indicate that students in higher years incur higher costs? (Use $\alpha = 0.10$.)
- c Predict with 95% confidence the annual cost of putting a child through Year 11.
- d Estimate with 98% confidence the mean annual cost of putting children through Year 9.

Computer/manual applications

The following exercises require the use of a computer and software. Alternatively, they may be solved using the sample statistics provided.

- 15.81 XR15-81** In an attempt to further investigate their advertising, the store manager of Colonial Furniture Ltd has been reviewing weekly expenditures. During the past six months all advertisements for the store have appeared in the local newspaper. The number of ads per week has varied from one to seven. The store's sales staff have been tracking the number of customers who enter the store each week. The number of ads and the number of customers per week for the past 26 weeks have been recorded.
- a Determine the sample regression line.
 - b Interpret the coefficients.
 - c Can the manager infer at the 5% significance level that the larger the number of ads, the larger the number of customers?
 - d Find and interpret the coefficient of determination.
 - e In your opinion, is it a worthwhile exercise to use the regression equation to predict the number of customers who will enter the store, given that Colonial Furniture intends to advertise in the newspaper five times per week? If so, find a 95% prediction interval. If not, explain why not.

Sample statistics: $n = 26$,
 $\bar{x}_{\text{Ads}} = 4.115$; $\bar{y}_{\text{Customer}} = 384.808$;
 $s_{xy} = 74.024$; $s_x^2 = 3.466$; $s_y^2 = 18552.08$.

- 15.82 XR15-82** The production manager of a company that manufactures car seats has been concerned about the number and cost of machine breakdowns. The problem is that the machines are old and are becoming quite unreliable. However, the cost of replacing them is quite high, and the manager is not certain that the cost can be recouped given the slow economy. To help make a decision about replacement, he gathered data about

last month's costs for repairs and the ages (in months) of the plant's 20 welding machines.

- a Find the sample regression line.
- b Interpret the coefficients.
- c Determine the standard error of estimate and discuss what this statistic tells you.
- d Conduct a test at whatever significance level you deem suitable to determine whether the age of a machine and its monthly cost of repair are linearly related.
- e Find and interpret the coefficient of determination.
- f Is the fit of the simple linear model good enough to allow the production manager to predict the monthly repair cost of a welding machine that is 120 months old? If so, find a 95% prediction interval. If not, explain why not.

Sample statistics: $n = 20$,
 $\bar{x}_{\text{Age}} = 113.35$; $\bar{y}_{\text{Repairs}} = 395.21$;
 $s_{xy} = 936.842$; $s_x^2 = 378.768$; $s_y^2 = 4094.793$.

- 15.83 XR15-83** Several years ago, Coca-Cola attempted to change its 100-year-old recipe. One reason why the company's management thought this was necessary was competition from Pepsi Cola. Surveys of Pepsi drinkers indicated that they preferred Pepsi because it was sweeter than Coke. As part of the analysis that led to Coke's ill-fated move, the management of Coca-Cola performed extensive surveys in which consumers tasted various versions of the new Coke. Suppose that a random sample of 200 cola drinkers were given versions of Coke with different amounts of sugar. After tasting the product, each drinker was asked to rate the taste quality. The possible responses were as follows:

1 = poor; 2 = fair; 3 = average; 4 = good;
 5 = excellent

The responses and the sugar content (percentage by volume) of the version tasted were recorded in columns 1 and 2 respectively. Can management infer at the 5% significance level that sugar content affects drinkers' ratings of the cola?

Sample statistics: $n = 200$,
 $\bar{x}_{\text{Sugar}} = 12.305$; $\bar{y}_{\text{Rating}} = 2.91$;
 $s_{xy} = 6.565$; $s_x^2 = 37.007$; $s_y^2 = 1.861$.

- 15.84 XR15-84** One general belief held by observers of the business world is that taller men earn more money than shorter men. In a study reported in the *Wall Street*

Journal, 30 MBA graduates, all about 30 years old, were surveyed and asked to report their annual incomes and their heights. These responses are recorded.

- a Determine the sample regression line and interpret the coefficients.
- b Find the standard error of estimate and interpret its value.
- c Do these data provide sufficient statistical evidence to infer at the 5% significance level that taller MBAs earn more money than shorter ones?
- d Provide a measure of the strength of the linear relationship between income and height.
- e Do you think that this model is good enough to be used to estimate and predict income on the basis of height? If not, explain why not. If so,
 - i estimate with 95% confidence the mean income of all 183 cm men with MBAs
 - ii predict with 95% confidence the income of a man 175 cm tall with an MBA.

Sample statistics: $n = 30$,

$$\bar{x}_{\text{Height}} = 174.0; \bar{y}_{\text{Income}} = 72\ 639.33;$$

$$s_{xy} = 14325.172; s_x^2 = 55.431; s_y^2 = 32\ 560\ 206.4$$

The following exercises require the use of a computer and software.

15.85 XR15-85 An agronomist wanted to investigate the factors that determine crop yield. Accordingly, she undertook an experiment in which a farm was divided into 30 half-hectare plots. The amount of fertiliser applied to each plot was varied. Wheat was then planted, and the amount harvested at the end of the season was recorded.

- a Find the sample regression line and interpret the coefficients.
- b Can the agronomist conclude that there is a linear relationship between the amount of fertiliser and the crop yield?
- c Find the coefficient of determination and interpret its value.
- d Does the simple linear model appear to be a useful tool in predicting crop yield from the amount of fertiliser applied? If so, produce a 95% prediction

interval of the crop yield when 300 kg of fertiliser is applied. If not, explain why not.

15.86 XR15-86 Car manufacturers are required to test the exhaust gases of their vehicles for a variety of pollutants. The amount of pollutant varies even among identical vehicles, so several vehicles must be tested. The engineer in charge of testing has collected data (in grams per kilometre driven) on the amounts of two pollutants, carbon monoxide and nitrous oxide, for 50 identical vehicles. The engineer believes the company can save money by testing for only one of the pollutants because the two pollutants are closely linked. That is, if a car is emitting a large amount of carbon monoxide, it will also emit a large amount of nitrous oxide. Do the data support the engineer's belief?

15.87 XR15-87 Society in general and the judicial system in particular have altered their opinions on the seriousness of drink driving. In most jurisdictions, driving an automobile with a blood alcohol level in excess of 0.05 is an offence. Because of a number of factors, it is difficult to provide guidelines on when it is safe for someone who has consumed alcohol to drive a car. In an experiment to examine the relationship between blood alcohol level and the weight of a drinker, 50 men of varying weights were each given three beers to drink, and one hour later their blood alcohol levels were measured. These data were recorded and stored. If we assume that the two variables are normally distributed, can we conclude that blood alcohol level and weight are related?

15.88 XR15-88 In Exercise 15.15, we considered the relationship between the amount of time spent watching television and the high level of obesity among children. Now, we may have to add financial problems to the list. A sociologist theorised that people who watch television frequently are exposed to many commercials, which in turn lead them to buy, resulting in increasing debt. To test this belief, a sample of 430 families was drawn. For each, the total debt and the number of hours the television is turned on per week were recorded. Perform a statistical procedure to help test this theory.

Case Studies

CASE 15.1 Does unemployment rate affect weekly earnings in New Zealand?

C15-01 Wages in the labour market are very much influenced by the demand and supply of labour. In any profession or industry, when there is an oversupply of labour, the workers will be at a disadvantage and will not be able to demand high wages and vice versa. In New Zealand, in the last few decades the average weekly earnings have fluctuated, depending on the state of the economy, especially on the level of labour supply. When the number of unemployed persons increases, it is expected that the average hourly earnings would fall. Quarterly data for the average weekly earnings and the rate of unemployment in New Zealand during the period March 2012 to September 2019 were recorded. Is there any evidence to support the proposition that the higher (lower) the rate of unemployment the lower (higher) the average weekly earnings in New Zealand?

CASE 15.2 Tourism vs tax revenue

C15-02 For many local governments in New Zealand, tourism is one source of raising their revenue to maintain a good livelihood for their taxpayers. One way of raising tourism revenue is via some form of tax from the hotels where the tourists stay. A tourism hotel operator is interested in investigating the relationship between guest nights and local government tax revenue. The data for both variables for the period 2009–18 are recorded. Analyse the relationship between the two variables.

CASE 15.3 Does unemployment affect inflation in New Zealand?

C15-03 A social science research student is interested in investigating the relationship between unemployment and inflation in New Zealand. The student found annual data on the rate of unemployment and the rate of inflation for the period 1990–2017 on the government website, *Statistics New Zealand*. Can you help the student to establish the relationship between the two variables using the data collected?

CASE 15.4 Does domestic market capital influence stock prices?

C15-04 A business manager would like to study the behaviour of Australian stock prices. On the basis of previous knowledge, he believes that the value of the domestic market capital would influence stock prices. In order to check his belief he has collected and recorded monthly data on the two variables. Analyse the data and the relationship between the two variables.

CASE 15.5 Book sales vs free examination copies

C15-05 The academic book business is different from most other businesses because of the way in which purchasing decisions are made. The customer, who is usually a student studying a university or TAFE subject, buys a specific book because the lecturer of the subject adopts

(chooses to use) that book. Sales representatives of publishers sell their products by persuading lecturers to adopt their books. Unfortunately, judging the quality of textbooks is not easy. To help with the decision process, sales representatives give free examination copies to lecturers so that they can review the book and decide whether or not to adopt it. In many universities, textbook review committees meet to make the adoption decision.

The senior publishing editor, Mr Geoff Howard, at Cengage Learning Australia was examining the latest data on the sales of recently published textbooks. He noted that the number of examination copies was quite large, which can be a serious problem given the high cost of producing books. The editor wonders whether his sales representatives were giving away too many free books, or perhaps not enough. The data contain the gross revenues from the sales of the books, and the number of free copies for a sample of 30 books. The publishing editor would like to know if there is a direct link between the number of free copies of a book and the gross revenue generated by that book.

Perform an analysis to provide the publishing editor with the required information.

CASE 15.6 Does increasing per capita income lead to increase in energy consumption?

C15-06 When people's incomes increase, their purchasing power increases, and they will be in a position to purchase more electrical goods, which leads to an increase in energy consumption. At a student energy forum, one science student argued that although the claim that increasing household income increases household energy consumption is true, it is also true that, in recent times, due to innovations in technology, electrical goods have become more energy efficient and this would have led to a reduction in energy consumption. You would like to investigate the validity of these arguments using income and energy consumption data for Australia. Data on energy use (kg of oil equivalent per capita) and GDP per capita (current \$AU) for the years 1970 to 2015 are available in the World Development Indicators (WDI) database. Investigate the validity of these arguments considering the relationship between the two variables.

Source: World Bank. CC BY 4.0 <https://creativecommons.org/licenses/by/4.0/>

CASE 15.7 Market model of share returns

C15-07 A well-known model in finance, called the *market model*, assumes that the rate of return on a share (R) is linearly related to the monthly rate of return on the overall market R_m . The mathematical description of the model is

$$R = \beta_0 + \beta_1 R_m + \varepsilon$$

where the error term ε is assumed to satisfy the requirements of the linear regression model. For practical purposes, R_m is taken to be the annual rate of return on some major stock market index, such as the Australian All Ordinaries Index.

The coefficient β_1 , called the share's beta coefficient, measures how sensitive the share's rate of return is to changes in the level of the overall market. For example, if $\beta_1 > 1$ (or $\beta_1 < 1$), the share's rate of return is more (or less) sensitive to changes in the level of the overall market than is the average share.

The annual rates of return for five shares (gold, energy, retail, banks and property trusts) and for the overall market over a 17-year period are recorded. (Columns 2–6 store the annual

percentage return for the five shares; column 7 stores the All Ordinaries Index.) For each share, determine the following:

- a What is the sample regression line?
- b Is there sufficient evidence to infer at the 5% significance level that there is a linear relationship between the return on the individual share and the return on the total market?
- c Is there sufficient evidence to infer at the 5% significance level that an individual share is less sensitive than the average share?
- d Discuss the significance of the findings.

CASE 15.8 Life insurance policies

C15-08 Most Australians purchase life insurance to help their families in case of unexpected events, such as accidents, that could cost them their lives. From the insurance companies' point of view, they would like to see their customers live longer so that the chances of their making a pay claim will be smaller. In light of this, life insurance companies are keenly interested in predicting how long their customers will live, because their premiums and profitability will depend on such information. It is a common belief among actuarial researchers that a person's lifetime is very much related to the lifetime of his or her mother, father and grandparents.

An actuary for an insurance company, however, doesn't believe that these variables are significantly correlated. He gathered data from 100 recently deceased male customers in order to investigate it further. He recorded the age at death of these male customers and the age at death of the customers' fathers. Can he conclude that a significant relationship exists between male customers' age at death and their fathers' age at death?

CASE 15.9 Education and income: How are they related?

C15-09 If you're taking this course, you're probably a student in an undergraduate or graduate business or economics program. Your plan is to graduate, get a good job and draw a high salary. You have probably assumed that more education equals better job equals higher income. Is this true? A social survey recorded the data for two variables – annual income and years of education – to help determine whether education and income are related and, if so, what the value of an additional year of education might be. Perform the analysis.

CASE 15.10 Male and female unemployment rates in New Zealand – Are they related?

C15-10 A social scientist is interested in investigating the relationship between the rate of unemployment of males and females and intends to predict the female rate of unemployment based on male rate of unemployment. The researcher uses quarterly data recorded from 2007–19 to establish the relationship between male and female rates of unemployment in New Zealand. Test whether male rate of unemployment affects the female rate of unemployment.

Multiple regression

Learning objectives

This chapter extends the concepts and techniques used for a simple regression model in Chapter 15 to regression models with two or more independent variables called the multiple regression models.

At the completion of this chapter, you should be able to:

- L01** develop a multiple regression model, use a computer and program to estimate the model and interpret the estimated coefficients
- L02** understand the adjusted coefficient of determination and assess the fitness of the model
- L03** test the significance of the individual coefficients and the overall utility of the model
- L04** use the estimated regression model to make predictions and perform diagnostic checks for the regression model assumptions.

CHAPTER OUTLINE

Introduction

16.1 Model and required conditions

16.2 Estimating the coefficients and assessing the model

16.3 Regression diagnostics – II

16.4 Regression diagnostics – III (time series)

SPOTLIGHT ON STATISTICS

Determinants of income I

If you're taking this course, you're probably a student in an undergraduate or graduate business or economics program. Your plan is to graduate, get a good job and draw a high salary. You have probably assumed that more education equals better job equals higher income. Is this true? Using a simple linear regression model, we investigated this issue using data from a social survey and found that income and education are linearly related. However, the fit of the model measured by the coefficient of determination R^2 was only 0.14. This raises the question of whether the model can be improved by incorporating other variables into the model. What other variables affect one's income?

To answer this question, a survey – in addition to the respondent's income and years of education – gathered information on the following list of numerical variables: age, hours of work per week of respondent and of his/her spouse, occupation prestige score of respondent, number of children, and years with current employer.

These data are stored in file **CH16:XM16-00**. The goal is to utilise a multiple regression model that includes all the variables that you believe affect a person's income. Estimate such a relationship to determine the factors that influence one's income.

See pages 704–07 for a solution.



Source: iStock.com/sorbetto

Introduction

In the previous chapter, we employed the simple linear regression model to analyse how one numerical variable (the dependent variable y) is affected by another numerical variable (the independent variable x). The restriction of using only one independent variable was motivated by the need to simplify the introduction to regression analysis. Although there are a number of applications where we purposely develop a model with only one independent variable (see Case 15.9, for example), in general we prefer to include as many independent variables as can be shown to significantly affect the dependent variable. Arbitrarily limiting the number of independent variables also limits the usefulness of the model.

In this chapter, we allow for any number of independent variables. In so doing, we expect to develop models that fit the data better than would a simple linear regression model. We will proceed in a manner similar to that in Chapter 15. We begin by describing the multiple regression model and listing the required conditions. We let the computer produce the required statistics and use them to assess the fitness of the model and diagnose violations of the required conditions. We will employ the model by interpreting the coefficients, predicting the particular value of the dependent variable and estimating its expected value.

16.1 Model and required conditions

We now assume that k independent variables are potentially related to the dependent variable. Thus, the model is represented by the following equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$$

where y is the dependent variable, x_1, x_2, \dots, x_k are the independent variables, $\beta_0, \beta_1, \dots, \beta_k$ are the coefficients and ε is the error variable. The independent variables may actually be functions of other variables. For example, we might define some of the independent variables as follows:

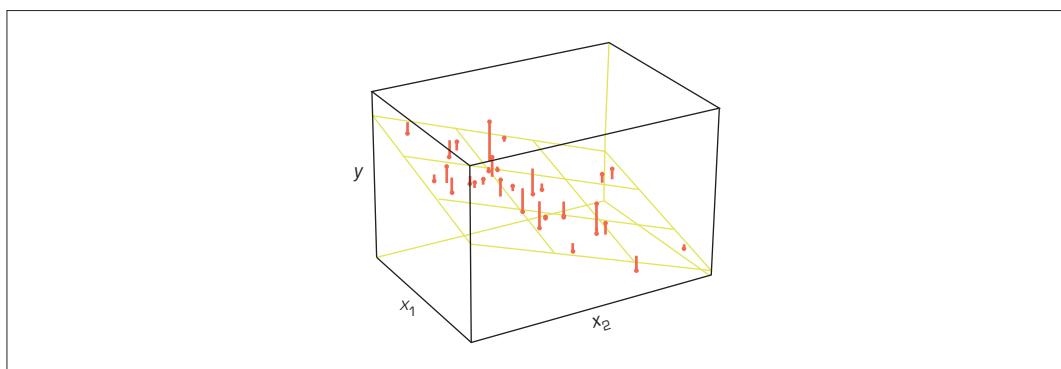
$$\begin{aligned}x_2 &= x_1^2 \\x_5 &= x_3 \cdot x_4 \\x_7 &= \log(x_6)\end{aligned}$$

The error variable is retained because, even though we have included additional independent variables, deviations between values in the model and the actual values of y will still occur. Incidentally, when there is more than one independent variable in the regression analysis, we refer to the graphical depiction of the equation as a **response surface** rather than as a straight line. **Figure 16.1** depicts a scatter diagram of a response surface with $k = 2$. (When $k = 2$, the regression equation creates a plane.) Of course, whenever k is greater than 2, we can only imagine the response surface – we cannot draw it.

An important part of the regression analysis comprises several statistical techniques that evaluate how well the model fits the data. These techniques require the following conditions, which we introduced in the previous chapter.

response surface

The graphical depiction of the regression equation when there is more than one independent variable; when there are two independent variables, the response surface is a plane.

FIGURE 16.1 Scatter diagram and response surface with $k = 2$ 

IN SUMMARY

Required conditions for the error variable

- 1 The probability distribution of the error variable ε is normal.
- 2 The mean of the error variable $E(\varepsilon)$ is zero.
- 3 The standard deviation of ε is σ_ε , which is a constant.
- 4 The errors are independent.
- 5 The errors are independent of the independent variables.

In Section 15.7, we discussed how to recognise when the requirements are not satisfied. Those same procedures can be used to detect violations of required conditions in the multiple regression model.

We now proceed as we did in Chapter 15: we discuss how the model's coefficients are estimated and how we assess the fitness of the model. However, there is one major difference between Chapters 15 and 16. In Chapter 15, we allowed for the possibility that some students will perform the calculations manually. The multiple regression model involves so many complicated calculations that it is almost impossible to conduct the analysis without a computer. All analyses in this chapter will be performed by Excel. Your job will be to interpret the output.

16.2 Estimating the coefficients and assessing the model

The sample multiple regression model is expressed similarly to the sample simple regression model. The general form of a sample multiple regression model is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k$$

The procedures introduced in Chapter 15 are extended to the multiple regression model. However, in Chapter 15, we discussed how to interpret the coefficients first, followed by a discussion of how to assess the fitness of the model. In practice, we reverse the process. That is, the first step is to determine how well the model fits. If the fitness of the model is poor, there is no point in a further analysis of the coefficients of that model. A much higher priority is assigned to the task of improving the model.

16.2a Six steps of regression analysis

In this section, we will show how a regression analysis is performed. The steps we will use are as follows:

- Step 1** Select variables that you believe are linearly related to the dependent variable.
- Step 2** Use a computer and software to generate the coefficients and the statistics used to assess the model.
- Step 3** Assess the fitness of the model. Three statistics that perform this function are the standard error of estimate, the coefficient of determination, and the *F*-test to assess the validity of the model. The first two were introduced in Chapter 15; the third will be introduced here.
- Step 4** Diagnose violations of the required conditions. If there are problems, attempt to remedy them.
- Step 5** If we are satisfied with the fitness of the model and that the required conditions are met, we can test for the significance of the individual coefficients as we did in Chapter 15.
- Step 6** Interpret the coefficients and use the model to predict a value of the dependent variable or estimate the expected value of the dependent variable.

When deciding on the independent variables to include in a multiple regression model, you may be wondering why we don't simply include all the variables that are available to us. There are three reasons for this. First, the objective is to determine whether our hypothesised model is valid and whether the independent variables in the model are linearly related to the dependent variable. That is, we should screen the independent variables and include only those that in theory affect the dependent variable. Second, by including large numbers of independent variables we increase the probability of Type I errors. For example, if we include 100 independent variables, none of which are related to the dependent variable, we're likely to conclude that five of them are linearly related to the dependent variable. Third, because of a problem called multicollinearity (described in Section 16.3), we may conclude that none of the independent variables are linearly related to the dependent variable when in fact one or more are.

We now illustrate the above six steps with the following examples.

EXAMPLE 16.1

L01 L02

Selecting sites for a motel chain I

XM16-01 A moderately priced chain of motels is located in major cities around the world. Its market is the frequent business traveller. The chain recently launched a campaign to increase market share by building new motels. The management of the chain is aware of the difficulty in choosing locations for new motels. Moreover, making decisions without adequate information often results in poor decisions. Consequently, the chain's management acquired data on 100 randomly selected motels belonging to the chain. The objective was to predict which sites are likely to be profitable. To measure profitability, the management used the operating margin, which is the ratio of the sum of profit, depreciation and interest expenses divided by total revenue. (Although occupancy is often used as a measure of a motel's success, the company statistician concluded that occupancy was too unstable, especially during economic turbulence.) The higher the operating margin, the greater the success of the motel. The chain defines profitable motels as those with an operating margin in excess of 50% and unprofitable motels as those with margins of less than 30%.

Step 1: Select the independent variables that you believe may be related to the dependent variable

In the above example, suggest suitable independent variables that would affect the operating margin.

Solution

The following variables can be considered more relevant in explaining the variation in the operating margin.

(a) Degree of competition

The degree of competition can be measured using the total number of motel and hotel rooms within 5 kilometres of the motel.

x_1 = Total number of motel and hotel rooms within 5 km of the motel (Number)

(b) Market awareness

Market awareness can be measured by the number of kilometres to the closest competing motel.

x_2 = Number of kilometres to closest competitor (Nearest)

(c) Demand generators

Two variables that represent sources of customers are chosen. The amount of office space and university enrolment in the surrounding community are demand generators. Both of these are measures of economic activity.

x_3 = Office space in thousands of square metres in surrounding community (OfficeSpace)

x_4 = University enrolment (in thousands) in nearby university (Enrolment)

(d) Demographics

A demographic variable that describes the community can be the median household income.

x_5 = Median household income (in \$'000) in surrounding community (Income)

(e) Physical quality

Finally, as a measure of the physical qualities of the location, we chose the distance to the Central Business District (CBD).

x_6 = Distance (in km) to the CBD (Distance)

These data are stored using the following format:

- Column 1: y (Margin)
- Column 2: x_1 (Number)
- Column 3: x_2 (Nearest)
- Column 4: x_3 (OfficeSpace)
- Column 5: x_4 (Enrolment)
- Column 6: x_5 (Income)
- Column 7: x_6 (Distance)

Some of these data are shown here.

Margin (y) (%)	Number (x_1)	Nearest (x_2) (km)	Office space (x_3) ('000 sq m)	Enrolment (x_4) ('000)	Income (x_5) (\$'000)	Distance (x_6) (km)
55.5	3203	4.2	54.9	8.0	40	4.3
33.8	2810	2.8	49.6	17.5	38	23.0
49.0	2890	2.4	25.4	20.0	38	4.2
...
40.0	3397	1.6	85.5	19.5	35	5.0
39.8	3823	3.6	20.2	17.0	41	7.7
35.2	3251	1.7	27.5	13.0	38	6.9

Step 2: Use computer software to compute all the coefficient estimates and other statistics

EXAMPLE 16.2

Selecting sites for a motel chain II

Use a computer software package to estimate the relationship between the operating margin and the selected independent variables.

Solution

Excel is used to produce the outputs below. From this printout we can see that the sample regression equation is

$$\hat{y} = 38.66 - 0.0076x_1 + 1.656x_2 + 0.198x_3 + 0.213x_4 + 0.366x_5 - 0.142x_6$$

or

$$\text{Margin} = 38.66 - 0.0076(\text{Number}) + 1.656(\text{Nearest}) + 0.198(\text{OfficeSpace}) \\ + 0.213(\text{Enrolment}) + 0.366(\text{Income}) - 0.142(\text{Distance})$$

Using the computer

Using Excel Data Analysis

Excel output for Example 16.1

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.7231					
5	R Square	0.5229					
6	Adjusted R Square	0.4921					
7	Standard Error	5.5248					
8	Observations	100					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	6	3110.80	518.47	16.99	0.0000	
13	Residual	93	2838.66	30.52			
14	Total	99	5949.46				
15							
16		Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
17	Intercept	38.6643	6.9690	5.5480	0.0000	24.825	52.5034
18	Number (x1)	-0.076	0.0013	-6.0586	0.000	-0.0101	-0.0051
19	Nearest (x2)	1.6564	0.6345	2.6051	0.0105	0.3964	2.9164
20	Office Space (x3)	0.1980	0.0342	5.7933	0.0000	0.1302	0.2659
21	Enrolment (x4)	0.2131	0.1338	1.5921	0.1148	-0.0527	0.4788
22	Income (x5)	0.3660	0.1271	2.8803	0.0049	0.1137	0.6184
23	Distance (x6)	-0.1424	0.1119	-1.2725	0.2064	-0.3647	0.0798

COMMANDS

- Type or import the data (**XM16-01**). Arrange the columns so that the first column is the dependent variable and the independent variables are in adjacent columns. Delete rows that have blanks in any of the columns.
- Click **DATA, Data Analysis, and Regression**.
- Specify the input Y Range (**A1:A101**), the input X Range (**B1:G101**), and a value for α (**0.05**). Click **Labels** (if necessary).
- Click **Output Range** and specify the output start cell reference (**J1**), or click **New Worksheet Ply**: and type the name of the sheet for the output (**Regression output**). Click **OK**



Using XLSTAT

	A	B	C	D	E	F	G
1	Regression of variable Margin (y):						
2	Goodness of fit statistics (Margin (y)):						
3	Observations	100					
4	Sum of weights	100					
5	DF	93					
6	R ²	0.523					
7	Adjusted R ²	0.492					
8	MSE	30.523					
9	RMSE	5.525					
10	DW	2.117					
11							
12	Analysis of variance (Margin (y)):						
13	Source	df	SS	MS	F	Pr > F	
14	Model	6	3110.80	518.47	16.99	< 0.0001	
15	Error	93	2838.66	30.52			
16	Corrected Total	99	5949.46				
17							
18	Model parameters (Margin (y)):						
19	Source	Value	Standard Error	t	Pr > t	Lower bound (95%)	Upper bound (95%)
20	Intercept	38.664	6.969	5.548	< 0.0001	24.825	52.503
21	Number (x1)	-0.008	0.001	-6.059	< 0.0001	-0.010	-0.005
22	Nearest (x2)	1.656	0.635	2.611	0.011	0.396	2.916
23	Office Space (x3)	0.198	0.034	5.793	< 0.0001	0.130	0.266
24	Enrolment (x4)	0.213	0.134	1.592	0.115	-0.053	0.479
25	Income (x5)	0.366	0.127	2.880	0.005	0.114	0.618
26	Distance (x6)	-0.142	0.112	-1.273	0.206	-0.365	0.080

COMMANDS

- Type the data or open the data file (**XM16-01**). Copy the relevant variables into a new spreadsheet. Arrange the columns so that the dependent variable is in the first column and independent variables are in adjacent columns.
- Click **XLSTAT**, **Modeling data**, and **Linear regression**. In the **Quantitative** box type the input range of Y (**A1:A101**). In the **X Explanatory variables** and **Quantitative** box type the input range of X (**B1:G101**).
- Click **Outputs** and check **Analysis of variance**.
- Click Missing data and select Remove the observations and **Across all Ys**. Click **OK**.

Step 3: Assess the fitness of the model

We assess the fitness of the model using the following three statistics:

- a standard error of estimate
- b adjusted coefficient of determination, and
- c F-test to assess the validity of the model.

Standard error of estimate

Recall that σ_ϵ is the standard deviation of the error variable ϵ and that, because σ_ϵ is a population parameter, it is necessary to estimate its value by using s_ϵ . In multiple regression, the standard error of estimate is defined as follows:

Standard error of estimate

$$s_e = \sqrt{\frac{SSE}{n - k - 1}}$$

where n is the sample size and k is the number of independent variables in the model.

Recall that we judge the magnitude of the standard error of estimate relative to the values of the dependent variable, and particularly to the mean of y , \bar{y} .

Coefficient of determination

Recall from Chapter 15 that the *coefficient of determination* can be calculated as

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{SSE}{(n-1)s_y^2}$$

where

$$SST = (n-1)s_y^2 = \sum y_i^2 - n\bar{y}^2$$

The Excel output for regression provides the value of the standard error of estimate and coefficient of determination automatically.

Notice that Excel outputs a second R^2 statistic, called the **coefficient of determination adjusted for degrees of freedom (adjusted R^2)**, which has been adjusted to take into account the sample size and the number of independent variables. The rationale for this statistic is that, if the number of independent variables k is large relative to the sample size n , the unadjusted R^2 value may be unrealistically high. To understand this point, consider what would happen if the sample size is 2 in a simple linear regression model. The line will fit the data perfectly, resulting in $R^2 = 1$ when, in fact, there may be no linear relationship. To avoid creating a false impression, the adjusted R^2 is often calculated. Its formula follows.

coefficient of determination adjusted for degrees of freedom (adjusted R^2)

A goodness-of-fit measure of the relationship between the dependent and independent variables, adjusted to take into account the number of independent variables in a multiple regression model.

Coefficient of determination adjusted for degrees of freedom

$$\text{Adjusted } R^2 = 1 - \frac{SSE / (n - k - 1)}{SST / (n - 1)} = 1 - \frac{SSE / (n - k - 1)}{s_y^2}$$

If n is considerably larger than k , the actual and adjusted R^2 values will be similar. But if SSE is quite different from zero and k is large compared to n , the actual and adjusted values of R^2 will differ substantially. If such differences exist, the analyst should be alerted to a potential problem in interpreting the coefficient of determination.

Testing the validity of the model

In the simple linear regression model, we tested the slope coefficient to determine whether sufficient evidence existed to allow us to conclude that there was a linear relationship between the independent variable and the dependent variable. However, because there is only one independent variable in that model, the t -test also tested to determine whether that model is useful. When there is more than one independent variable, we need another method to test the overall utility of the model. The technique is a version of the analysis of variance, which we introduced in Chapter 15.

To test the utility of the regression model, we specify the following hypotheses:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0.$$

$$H_A: At least one \beta_i is not equal to zero$$

If the null hypothesis is true, none of the independent variables x_1, x_2, \dots, x_k is linearly related to y , and therefore the model is useless. If at least one β_i is not equal to zero, the model does have some utility.

When we introduced the coefficient of determination in Chapter 15, we noted that the total variation (SST) in the dependent variable (measured by $(n-1)s_y^2$) can be decomposed into two parts: the explained variation (measured by SSR) and the unexplained variation (measured by SSE). That is,

$$SST = SSR + SSE$$

Furthermore, we established that, if SSR is large relative to SSE, the coefficient of determination will be high – signifying a good model. On the other hand, if SSE is large, most of the variation will be unexplained, which indicates that the model provides a poor fit and consequently has little utility.

The test statistic is the same one we encountered in Section 14.1, in which we tested for the equivalence of k population means. In order to judge whether SSR is large enough relative to SSE to allow us to infer that at least one coefficient is not equal to zero, we calculate the ratio of the two mean squares. (Recall that the *mean square* is the sum of squares divided by its degrees of freedom; recall, too, that the ratio of two mean squares is F -distributed, as long as the underlying population is normal – a required condition for this application.) The calculation of the test statistic is summarised in an analysis of variance (ANOVA) table, which, in general, appears as follows:

Analysis of variance table for regression analysis

Source of variation	Degrees of freedom	Sums of squares	Mean squares	F -statistic
Regression	k	SSR	$MSR = SSR/k$	$F = MSR/MSE$
Residual	$n - k - 1$	SSE	$MSE = SSE/(n - k - 1)$	
Total	$n - 1$	SST		

A large value of F indicates that most of the variation in y is explained by the regression equation and that the model is useful. A small value of F indicates that most of the variation in y is unexplained. The rejection region allows us to determine whether F is large enough to justify rejecting the null hypothesis. For this test, the rejection region is

$$F > F_{\alpha, k, n-k-1}$$

s_e , adjusted- R^2 and F -test

Although each assessment measurement offers a different perspective, all agree in their assessment of how well the model fits the data, because all are based on the sum of squares for error, SSE. The standard error of estimate is

$$s_e = \sqrt{\frac{SSE}{n - k - 1}}$$

and the adjusted coefficient of determination is

$$\text{Adjusted } R^2 = 1 - \frac{SSE / (n - k - 1)}{SST / (n - 1)}$$

where $SST = (n-1)s_y^2$. When the response surface hits every single point, SSE = 0. Hence, $s_e = 0$ and adjusted $R^2 = 1$.

If the model provides a poor fit, we know that SSE will be large (its maximum value is SST), s_e will be large and (since SSE is close to SST) the adjusted R^2 will be close to zero.

The F -statistic also depends on SSE. Specifically,

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR} / k}{\text{SSE} / (n - k - 1)} = \frac{(\text{SST} - \text{SSE}) / k}{\text{SSE} / (n - k - 1)}$$

When SSE = 0,

$$F = \frac{\text{SST} / k}{0 / (n - k - 1)}$$

which is infinitely large. When SSE is large, SSE is close to SST and F is quite small.

The relationship among s_e , adjusted R^2 and F is summarised in **Table 16.1**.

TABLE 16.1 Relationship among s_e , adjusted R^2 and F

SSE	s_e	Adjusted R^2	F	Assessment of model
0	0	1	∞	Perfect
Small	Small	Close to 1	Large	Good
Large	Large	Close to 0	Small	Poor
SST	$\sqrt{\frac{\text{SSE}}{n - k - 1}}$ *	0	0	Not useful

*When n is large and k is small, this quantity is approximately equal to the standard deviation of y .

EXAMPLE 16.3

Selecting sites for a motel chain III

Assess the fit of the model in Example 16.1, using the standard error of estimate, the adjusted coefficient of determination and the F -test of analysis of variance.

Solution

Partial Excel Data Analysis regression output

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.723					
5	R Square	0.523					
6	Adjusted R Square	0.492					
7	Standard Error	5.525					
8	Observations	100					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	6	3110.80	518.47	16.99	0.0000	
13	Residual	93	2838.66	30.52			
14	Total	99	5949.46				

► Partial XLSTAT regression output

	A	B	C	D	E	F	G
10	Regression of variable Margin [y]:						
11	Goodness of fit statistics [Margin [y]]:						
12	Observations	100					
13	Sum of weights	100					
14	DF	93					
15	R ²	0.523					
16	Adjusted R ²	0.492					
17	RMSE	30.523					
18		5.525					
19	Analysis of variance [Margin [y]]:						
20	Source	df	SS	MS	F	Pr > F	
21	Model	6	3110.80	518.47	16.99	< 0.0001	
22	Error	93	2838.66	30.52			
23	Corrected Total	99	5949.46				

From the Excel output in Example 16.1, the standard error of estimate s_e is 5.525. Recall that we judge the magnitude of the standard error of estimate relative to the values of the dependent variable, and particularly to the mean of y . We have $\bar{y} = 45.7$ (not shown in the printout); therefore

$$\frac{s_e}{\bar{y}} = \frac{5.525}{45.7} \times 100 = 12.08\%$$

It appears that the standard error of estimate is not particularly small.

From the output, the coefficient of determination R^2 is 0.523. This means that 52.3% of the variation in operating margins is explained by the six independent variables, while 47.7% remains unexplained.

Also from the output, the adjusted coefficient of determination is 0.492. That is, when adjusted for degrees of freedom, 49.2% of the variation in operating margins is explained by the model, while 50.8% remains unexplained. This indicates that, no matter how we measure the coefficient of determination, the model's fit is moderately good.

As you can see from the output above, the value of the test statistic for testing the overall utility of the model, $F = 16.99$. The output also includes the p -value of the test, which is 0.000.

The rejection region (assuming $\alpha = 0.05$) is

$$F > F_{\alpha, k, n-k-1} = F_{0.05, 6, 93} \approx F_{0.05, 6, 90} = 2.20 \quad (\text{using Table 6(a) Appendix B})$$

As $F = 16.99 > 2.20$ or as $p\text{-value} = 0.00 < 0.05 = \alpha$, we reject the null hypothesis. Therefore, there is a great deal of evidence to infer that the model is useful at the 5% level of significance.

Step 4: Diagnose violations of required conditions

EXAMPLE 16.4

LO3

Selecting sites for a motel chain IV

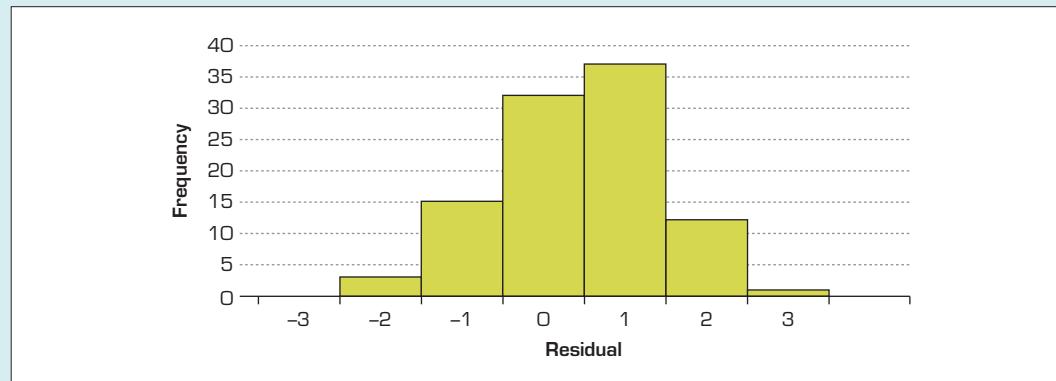
Now we determine whether the error terms satisfy the required assumptions of the regression model. Consider the estimated multiple regression model for the operating margin of the motel chain in Example 16.1. Perform the regression diagnostics of the estimated model.

Solution

As in Section 15.7, we perform regression diagnostics to see that the error term satisfies the following three assumptions.

- 1 The probability distribution of the error variable ε is normal.

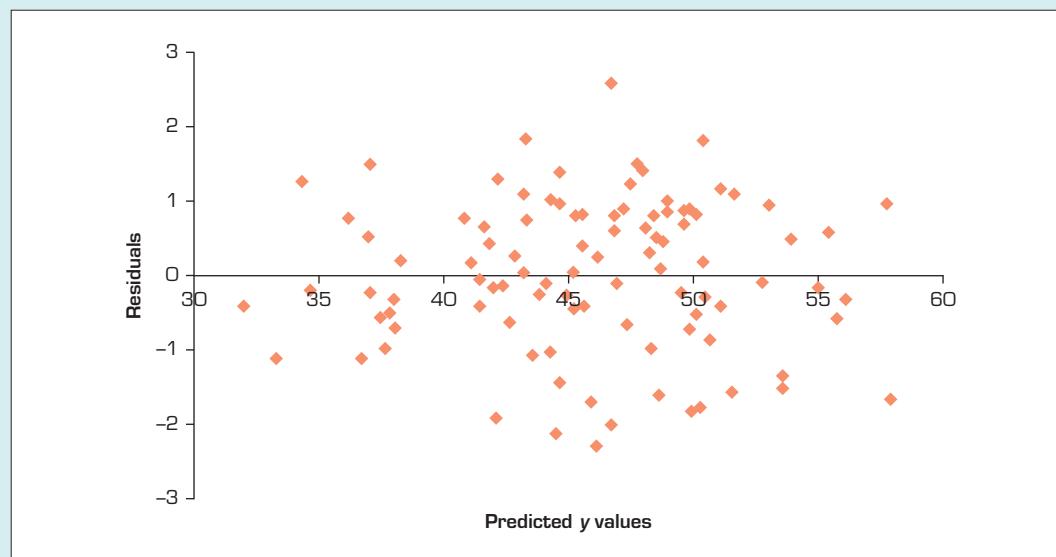
FIGURE 16.2 Histogram of the residuals



As can be seen from the histogram, the distribution of the errors is reasonably close to a normal distribution.

- 2 The errors are homoscedastic (standard deviation of ε , σ_ε , is a constant).

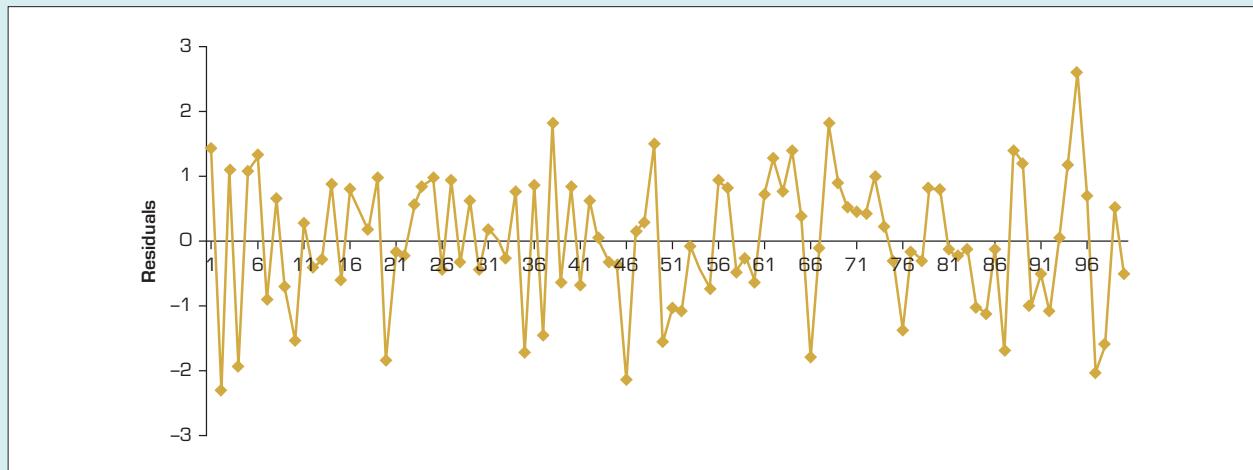
FIGURE 16.3 Plot of the residuals against the predicted y values



As can be seen from **Figure 16.3**, there is no evidence of heteroscedasticity.

- 3 The errors are independent.

FIGURE 16.4 Plot of the residuals against time



As can be seen from **Figure 16.4**, there is no evidence of autocorrelation.

Step 5: Testing the significance of the individual coefficient estimates

If we are satisfied with the fitness of the model and that the required conditions are met, we can test for the significance of the individual coefficients and interpret the coefficients.

In Section 15.4, we described how to test to determine whether there is sufficient evidence to infer that in the simple linear regression model x and y are linearly related. The null and alternative hypotheses were

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 \neq 0$$

The test statistic was

$$t = \frac{\hat{\beta}_i - \beta_i}{s_{\hat{\beta}_i}}$$

which is Student- t distributed with $n - 2$ degrees of freedom.

In the multiple regression model, we have more than one independent variable; for each such variable, we can test to determine if there is enough evidence of a linear relationship between it and the dependent variable for the entire population when the other independent variables are included in the model.

Testing the coefficients

$$H_0: \beta_i = 0$$

$$H_A: \beta_i \neq 0 \text{ (for } i = 1, 2, \dots, k\text{)}$$

The test statistic is

$$t = \frac{\hat{\beta}_i - \beta_i}{s_{\hat{\beta}_i}}$$

which is Student- t distributed with $d.f. = n - (k + 1)$. Note that here $(k + 1)$ represents the number of coefficients (including the intercept term) or k represents the number of independent variables in the model.

EXAMPLE 16.5

LO3

Selecting sites for a motel chain V

Consider the estimated multiple regression model for the operating margin of the motel chain in Example 16.1. Perform the test of significance for each of the coefficients in the model.

Solution

The tests that follow are performed just as all other tests in this book have been performed. We set up the null and alternative hypotheses, identify the test statistic, and use the computer to calculate the value of the test statistic and its p -value. For each independent variable, we test ($i = 1, 2, 3, 4, 5, 6$).

$$\begin{aligned} H_0: \beta_i &= 0 \\ H_A: \beta_i &\neq 0 \end{aligned} \quad (\text{Two-tail test})$$

Refer to pages 689–690 and examine the computer output for Example 16.1. The output includes the t -tests of β_i . The results of these tests pertain to the entire population of the motel chain considered in this study. It is also important to add that these test results were determined when the other independent variables were included in the model. We add this statement because a simple linear regression will very likely result in different values of the test statistics and possibly the conclusion.

16		Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
17	Intercept	38.664	6.969	5.548	0.000	24.825	52.503
18	Number (x1)	-0.008	0.001	-6.059	0.000	-0.010	-0.005
19	Nearest (x2)	1.656	0.635	2.611	0.011	0.396	2.916
20	Office Space (x3)	0.198	0.034	5.793	0.000	0.130	0.266
21	Enrolment (x4)	0.213	0.134	1.592	0.115	-0.053	0.479
22	Income (x5)	0.366	0.127	2.880	0.005	0.114	0.618
23	Distance (x6)	-0.142	0.112	-1.273	0.206	-0.365	0.080

Test of $\hat{\beta}_1$ (Coefficient of the number of motel and hotel rooms)

Value of the test statistic: $t = -6.059$; p -value = 0.000.

As the p -value = 0 < 0.05 = α ; reject $H_0: \beta_1 = 0$.

There is overwhelming evidence to infer that the number of motel and hotel rooms within 5 kilometres of the motel chain and the operating margin are linearly related.

Test of $\hat{\beta}_2$ (Coefficient of the distance to nearest competitor)

Value of the test statistic: $t = 2.611$; p -value = 0.011.

As the p -value = 0.011 < 0.05 = α ; reject $H_0: \beta_2 = 0$.

There is evidence to conclude that the distance to the nearest motel and the operating margin of the motel chain are linearly related.

Test of $\hat{\beta}_3$ (Coefficient of office space)

Value of the test statistic: $t = 5.793$; p -value = 0.000.

As the p -value = 0 < 0.05 = α ; reject $H_0: \beta_3 = 0$.

This test allows us to infer that there is a linear relationship between the operating margin and the amount of office space around the motel.

Test of $\hat{\beta}_4$ (Coefficient of university enrolment)

Value of the test statistic: $t = 1.592$; p -value = 0.115.

As the p -value = 0.115 > 0.05 = α , do not reject $H_0: \beta_4 = 0$.

There is not enough evidence to conclude that university enrolments is linearly related to the operating margin.





Test of $\hat{\beta}_5$ (Coefficient of median household income)

Value of the test statistic: $t = 2.88$; p -value = 0.005.

As the p -value = 0.0049 < 0.05 = α ; reject $H_0: \beta_5 = 0$.

There is sufficient statistical evidence to indicate that the operating margin and the median household income are linearly related.

Test of $\hat{\beta}_6$ (Coefficient of distance to CBD)

Value of the test statistic: $t = -1.273$; p -value = 0.206.

As the p -value = 0.206 > 0.05 = α ; do not reject $H_0: \beta_6 = 0$.

There is not enough evidence to infer the existence of a linear relationship between the distance to the CBD and the operating margin of the motel chain in the presence of the other independent variables.

Overall, there is sufficient evidence at the 5% significance level to infer that each of the following variables is linearly related to operating margin:

x_1 = Total number of motel and hotel rooms within 5 kilometres of the motel chain (Number)

x_2 = Number of kilometres to closest competitor (Nearest)

x_3 = Office space in thousands of square metres in surrounding community (OfficeSpace)

x_5 = Median household income in surrounding community (Income).

In this model, there is not enough evidence to conclude that any of the following variables are linearly related to income:

x_4 = University enrolment in nearby university (Enrolment)

x_6 = Distance (in km) to the CBD (Distance).

This may mean that there is no evidence of a linear relationship between operating margin and university enrolment and distance to the CBD. However, it may also mean that there is a linear relationship between profit margin and one or both of these variables, but because of a condition called *multicollinearity*, the t -test revealed no linear relationship. We will discuss multicollinearity in Section 16.3.

Step 6: Interpreting the coefficient estimates and using the model for prediction

Based on the above discussion of the test results, we found that the model is useful. We can now interpret the estimated coefficients. Then we use the model for prediction.

EXAMPLE 16.6

LO3

Selecting sites for a motel chain VI

Consider the estimated multiple regression model for the operating margin of a motel chain in Example 16.1.

Interpret the estimated coefficients in the model.

Interpreting the coefficients

Intercept

The intercept is $\hat{\beta}_0 = 38.66$. This is the average operating margin when all the independent variables are zero. As we observed in Chapter 15, it is often misleading to try to interpret this value, particularly if 0 is outside the range of the values of the independent variables (as is the case here).

Number of motel and hotel rooms (x_1)

The relationship between operating margin and the number of motel and hotel rooms within 5 kilometres is described by $\hat{\beta}_1 = -0.008$. From this number, we learn that, in this model, for each additional room within 5 kilometres of the motel, the operating margin decreases on average by 0.008%, assuming that the other





independent variables in this model are held constant. Changing the units we can interpret to say that for each additional 1000 rooms, the margin decreases by 8%.

Distance to nearest competitor (x_2)

The coefficient $\hat{\beta}_2 = 1.656$ specifies that for each additional kilometre of the nearest competitor from the chain motel, the average operating margin increases by 1.66%, assuming the constancy of the other independent variables.

The nature of the relationship between operating margin and the number of motel and hotel rooms and between operating margin and the distance to the nearest competitor was expected. The closer the competition, the lower the operating margin becomes.

Office space (x_3)

The relationship between office space and operating margin is expressed by $\hat{\beta}_3 = 0.198$. Because office space is measured in thousands of square metres, we interpret this number as the average increase in operating margin for each additional 1000 square metres of office space, keeping the other independent variables fixed. So, for every extra 1000 square metres of office space, the operating margin increases on average by 0.198%.

University enrolment (x_4)

The relationship between operating margin and college and university enrolment is described by $\hat{\beta}_4 = 0.213$, which we interpret to mean that for each additional 1000 students the average operating margin increases by 0.21% when the other variables are constant.

Both office space and enrolment produced positive coefficients, indicating that these measures of economic activity are positively related to the operating margin.

Median household income (x_5)

The relationship between operating margin and median household income is described by $\hat{\beta}_5 = 0.366$. For each additional thousand dollar increase in median household income, the average operating margin increases by 0.37%, holding all other variables constant. This statistic suggests that motels in more affluent communities have higher operating margins.

Distance to CBD (x_6)

The last variable in the model is distance to the CBD. Its relationship with operating margin is described by $\hat{\beta}_6 = -0.142$. This tells us that for each additional kilometre from the CBD, the operating margin decreases on average by 0.14%, keeping the other independent variables constant. Given that the chain's market is the 'frequent business traveller', it may be that frequent business travellers prefer to stay at motels that are closer to town centres.

Using the regression equation

As was the case with simple linear regression, we can use the multiple regression equation in two ways: we can produce the prediction interval for a particular value of y , and we can produce the confidence interval estimate of the expected value of y . Like the other calculations associated with multiple regression, we use the computer to do the work.

EXAMPLE 16.7

LO4

Selecting sites for a motel chain VII

Suppose that in Example 16.1 a manager investigated a potential site for the motel chain and found the following characteristics. There are 3815 rooms within 5 kilometres of the site ($x_1 = 3815$) and the closest other hotel or motel is 0.9 kilometres away ($x_2 = 0.9$). The amount of office space is 47 600 square metres ($x_3 = 47.6$). There are two universities nearby with a total enrolment of 24 500 students ($x_4 = 24.5$). From the census, the manager learns that the median household income in the area (rounded to the nearest thousand) is \$38 000 ($x_5 = 38$).



Finally, the distance to the CBD has been measured at 17.9 kilometres ($x_6 = 17.9$). The manager wants to predict the operating margin if and when the motel is built.

Solution

Calculating manually

The predicted annual gross revenue (in thousands) is

$$\begin{aligned}\hat{y} &= 38.66 - 0.0076(3815) + 1.656(0.9) + 0.198(47.6) + 0.213(24.5) + 0.366(38) - 0.142(17.9) \\ &= 37.076\end{aligned}$$

Therefore, the predicted operating margin from the proposed motel is 37.076%.

Using the computer

Excel Data Analysis does not perform this part of the analysis automatically. However, using the estimated regression coefficients from the Excel output and the appropriate formula, we can obtain the above predicted value. The output below shows how XLSTAT¹ can be used to do these calculations. XLSTAT outputs the (95%) prediction interval for one motel and the interval estimate of the expected (average) operating margin for all sites with the given variables.

XLSTAT output for Example 16.7

	B	C	D	E	F	G	H	I
103	Predictions for the new observations (Price):							
104	Observation	Pred (Margin [y])	Std. dev. on pred. (Mean)	Lower bound 95% (Mean)	Upper bound 95% (Mean)	Std. dev. on pred. (Observation)	Lower bound 95% (Observation)	Upper bound 95% (Observation)
105	PredObs1	37.076	2.080	32.945	41.207	5.903	25.353	48.799

COMMANDS

- Type or import the data (**XM16-01**) and conduct a regression analysis (use the same commands as in Example 16.2, page 691). Type the given values of the independent variables into the next available row in the columns containing the independent variables. (**3815, 0.9, 47.6, 24.5, 38.0, 17.9** in **B102:G102**).
- Click **Predictions**. In the **X/Explanatory variables** box check **Quantitative** and type the range of the cells containing the given values of the independent variables. (**B102:G102**)
- Click **OK**.

Interpreting the results

As you can see, the 95% prediction interval for the operating margin falls between 25.4 and 48.8. This interval is quite wide, confirming the need to have extremely well-fitting models to make accurate predictions. However, management defines a profitable motel as one with an operating margin greater than 50% and an unprofitable motel as one with an operating margin below 30%. Because the entire prediction interval is below 50 and part of it is below 30, the management of the motel chain will pass on this site.

The expected operating margin of all sites that fit this category is estimated to be between 32.9 and 41.2. We interpret this to mean that if we built motels on an infinite number of sites that fit the category described above, the mean operating margin would fall between 32.9 and 41.2. In other words, the average motel would not be profitable.

16.2b A cautionary note about interpreting the results

Care should be taken when interpreting the results of this and other regression analyses. We might find that in one model there is enough evidence to conclude that a particular independent variable is linearly related to the dependent variable, but that in another model no such evidence exists. Consequently, whenever a particular *t*-test is not significant, we

¹ The Prediction interval option in the Excel add-in Data Analysis Plus can be used for this purpose. Instructions on how to download the *Data Analysis Plus* Add-in is given in an Appendix to Chapter 1.

state that there is not enough evidence to infer that the independent and dependent variable are linearly related in this model. The implication is that another model may yield different conclusions.

Furthermore, if one or more of the required conditions are violated, the results may be invalid. In Section 15.7 we introduced the procedures that allow the statistics practitioner to examine the model's requirements. We will add to this discussion in Section 16.3. We also remind you that it is dangerous to extrapolate far outside the range of the observed values of the independent variables.

16.2c *t*-tests and the analysis of variance

The *t*-tests of the individual coefficients allow us to determine whether $\beta_i \neq 0$ (for $i = 1, 2, \dots, k$), which tells us whether a linear relationship exists between x_i and y . There is a *t*-test for each independent variable. Consequently, the computer automatically performs k *t*-tests. (It actually conducts $k + 1$ *t*-tests, including the one for β_0 , which we usually ignore.) The *F*-test in the analysis of variance combines these k *t*-tests into a single test. That is, we test whether $\beta_1 = \beta_2 = \dots = \beta_k = 0$ at one time to determine if at least one of them is not equal to zero. The question naturally arises, why do we need the *F*-test if it is nothing more than the combination of the previously performed *t*-tests? Using a series of *t*-tests, we increase the probability of making a Type I error. That means that even when there is no linear relationship between each of the independent variables and the dependent variable, multiple *t*-tests will likely show that some are significant. As a result, you will conclude erroneously that, since at least one β_i is not equal to zero, the model has some utility. The *F*-test, on the other hand, is performed only once. Because the probability that a Type I error will occur in a single test is equal to α , the chance of erroneously concluding that the model is useful is substantially less with the *F*-test than with multiple *t*-tests.

There is another reason why the *F*-test is superior to multiple *t*-tests. Because of a commonly occurring problem called *multicollinearity*, the *t*-tests may indicate that some independent variables are not linearly related to the dependent variable, when in fact they are. The problem of multicollinearity does not affect the *F*-test, nor does it inhibit us from developing a model that fits the data well. As mentioned above, multicollinearity will be discussed in Section 16.3.

16.2d The *F*-test and the *t*-test in the simple linear regression model

It is useful for you to know that we can use the *F*-test to test the utility of the simple linear regression model. However, this test is identical to the *t*-test of β_1 . The *t*-test of β_1 in the simple linear regression model tells us whether that independent variable is linearly related to the dependent variable. However, because there is only one independent variable, the *t*-test of β_1 also tells us whether the model is useful, which is the purpose of the *F*-test.

The relationship between the *t*-test of β_1 and the *F*-test can be explained mathematically. Statistics practitioners can show that if we square a *t*-statistic with v degrees of freedom, we produce an *F*-statistic with 1 and v degrees of freedom. (We briefly discussed this relationship in Chapter 15.) To illustrate, consider Example 15.3 in Chapter 15. We found the *t*-statistic for β_1 to be -13.59 (see Excel output for Example 15.3, page 634), with degrees of freedom equal to 98 ($d.f. = n - 2 = 100 - 2 = 98$). The *p*-value was 0. The output included the analysis of variance table in which $F = 184.66$ and *p*-value = 0. The *t*-statistic squared is $t^2 = (-13.59)^2 = 184.66$. Notice that the degrees of freedom of the *F*-statistic are 1 and 98. Thus, we can use either test to test the utility of the simple linear regression model.

Now let us return to the opening example in this chapter.

SPOTLIGHT ON STATISTICS

Determinants of income I: Solution

Step 1: Select the independent variables

Here are the variables we believe may be linearly related to one's income.

- Age (AGE): For most people income increases with age as they become more experienced.
- Years of education (EDUC): Obviously, the higher the level of education, the higher the likelihood of earning more income.
- Hours of work per week (HRS): Obviously, more hours of work should equal more income.
- Spouse's hours of work (SPHRS): It is possible that if one's spouse works more and earns more, the other spouse may choose to work less and thus earn less.
- Occupation prestige score (PRESTG): Occupations with higher prestige scores tend to pay more.
- Number of children (CHILDS): Children are expensive. Additional childcare responsibilities may reduce the hours worked, hence reducing one's income.
- Years with current employer (CUREMPYR): This variable could be positively related to income as loyalty and/or more years of experience would increase the likelihood of earning more.



Source: iStock.com/sorbetto

Step 2: Use a computer to compute all coefficients and other statistics

Excel Data Analysis output

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2		Regression Statistics					
3	Multiple R	0.584					
4	R Square	0.341					
5	Adjusted R Square	0.323					
6	Standard Error	32726.6					
7	Observations	272					
8							
9	ANOVA						
10		df	SS	MS	F	Significance F	
11	Regression	7	1.46148E+11	20878350989	19.49	0.000	
12	Residual	264	2.82752E+11	1071031858			
13	Total	271	4.28901E+11				
14							
15		Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
16	Intercept	-56018.72	19516.06	-2.87	0.004	-94445.67	-17591.78
17	AGE	504.97	236.61	2.13	0.034	39.08	970.86
18	EDUC	4003.11	877.88	4.56	0.000	2274.58	5731.64
19	HRS	752.77	175.69	4.28	0.000	406.85	1098.70
20	SPHRS	-839.94	183.26	-4.58	0.000	-1200.77	-479.11
21	PRESTG	605.99	176.66	3.43	0.001	258.15	953.82
22	CHILDS	244.10	1514.57	0.16	0.872	-2738.08	3226.28
23	CUREMPYR	323.03	237.42	1.36	0.175	-144.45	790.50

XLSTAT output

	B	C	D	E	F	G	H
12	Regression of variable INCOME:						
13	<i>Goodness of fit statistics (INCOME):</i>						
14	Observations	272					
15	Sum of weights	272					
16	DF	264					
17	R ²	0.341					
18	Adjusted R ²	0.323					
19	MSE	1071031857.9					
20	RMSE	32726.6					
21	MAPE	49.314					
22	DW	2.046					
23	Cp	8.000					
24	AIC	5663.3					
25	SBC	5692.1					
26	PC	0.699					
27							
28	<i>Analysis of variance (Margin (y)):</i>						
29	Source	df	SS	MS	F	Pr > F	
30	Model	7	146148456926.4	20878350989.5	19.49	<0.0001	
31	Error	264	282752410490.9	1071031857.9			
32	Corrected Total	271	428900867417.3				
33							
34	<i>Model parameters (INCOME):</i>						
35	Source	Value	Standard error	t	Pr > t	Lower bound (95%)	Upper bound (95%)
36	Intercept	-56018.72	19516.06	-2.87	0.004	-94445.67	-17591.78
37	AGE	504.97	236.61	2.13	0.034	39.08	970.86
38	EDUC	4003.11	877.88	4.56	<0.0001	2274.58	5731.64
39	HRS	752.77	175.69	4.28	<0.0001	406.85	1098.70
40	SPHRS	-839.94	183.26	-4.58	<0.0001	-1200.77	-479.11
41	PRESTG	605.99	176.66	3.43	0.001	258.15	953.82
42	CHILDS	244.10	1514.57	0.16	0.872	-2738.08	3226.28
43	CUREMPYR	323.03	237.42	1.36	0.175	-144.45	790.50

The regression model is estimated by

$$\hat{Y}_{\text{INCOME}} = -56019 + 505(\text{AGE}) + 4003(\text{EDUC}) + 753(\text{HRS}) - 840(\text{SPHRS}) + 606(\text{PRESTG}) \\ + 244(\text{CHILDS}) + 323(\text{CUREMPYR})$$

Step 3: Assess the fitness of the model

We assess the model in three ways: the standard error of estimate, the coefficient of determination, and the F-test of the analysis of variance (presented subsequently).

Standard error of estimate

Recall that we judge the magnitude of the standard error of estimate relative to the values of the dependent variable, and particularly to the mean of y . In this example, $\bar{y} = 56301$ (not shown in printouts). It appears that the standard error of estimate (32 726.6), from the above output, is quite large.

Adjusted coefficient of determination

The adjusted coefficient of determination is 32.3% (from the above output). There is an improvement in the adjusted R^2 value from 0.14 (simple linear regression model) to 0.32. This means that, after adjusting for degrees of freedom, 32.3% of the total variation in income is explained by the variation in the seven independent variables, whereas 67.7% remains unexplained. Therefore, the model's fit is still not very good and needs to be improved.

Testing the validity of the model

The value of the F-test statistic and the p-value for the test are presented in the analysis of variance (ANOVA) section of the above output. For this example, the rejection region (assuming $\alpha = 0.05$) is

$$F > F_{\alpha, k, n-k-1} = F_{0.05, 7, 264} = 2.01$$

As you can see from the above output, $F = 19.49$, which is greater than the critical value of 2.01. The output also includes the p-value of the test, which is 0.000. Obviously, there is a great deal of evidence to infer that the model is valid.

Step 4: Checking the model assumptions

The model assumptions were investigated (as in Example 16.4) and found that there is no evidence of violation of the assumptions.

Step 5: Testing for the significance of the individual coefficients

Testing the coefficients

To test to determine whether there is sufficient evidence to infer that the independent variables in the regression model and y are linearly related, the null and alternative hypotheses are

$$H_0: \beta_i = 0$$

$$H_A: \beta_i \neq 0, \quad i = 1, \dots, 7$$

Refer to page 704 and examine the output. The output includes the t -test results of β_i ($i = 1, \dots, 7$):

Test of β_1 (Coefficient of age)

Value of the test statistic: $t = 2.13$; p -value = 0.034

Test of β_2 (Coefficient of education)

Value of the test statistic: $t = 4.56$; p -value = 0

Test of β_3 (Coefficient of number of hours of work per week)

Value of the test statistic: $t = 4.28$; p -value = 0

Test of β_4 (Coefficient of spouse's number of hours of work per week)

Value of the test statistic: $t = -4.58$; p -value = 0

Test of β_5 (Coefficient of occupation prestige score)

Value of the test statistic: $t = 3.43$; p -value = 0.0007

Test of β_6 (Coefficient of number of children)

Value of the test statistic: $t = 0.161$; p -value = 0.872

Test of β_7 (Coefficient of years with current employer)

Value of the test statistic: $t = 1.36$; p -value = 0.175

Based on the p -values, there is sufficient evidence at the 5% significance level to infer that each of the following variables is linearly related to income:

- Age (+ve)
- Education (+ve)
- Number of hours of work per week (+ve)
- Spouse's number of hours of work per week (-ve)
- Occupation prestige score (+ve)

There is not enough evidence to conclude that each of the following variables is linearly related to income:

- Number of children
- Number of years with current employer

Note that this may mean that there is no evidence of a linear relationship between income and these two independent variables. However, it may also mean that there is a linear relationship between income and one or both of these variables, but because of multicollinearity the t -tests revealed no linear relationship.

Step 6: Interpretation and using the regression model for prediction

Interpreting the coefficient estimates

The coefficients $\hat{\beta}_i$, $i = 0, 1, 2, \dots, 7$, describe the relationship between each of the independent variables and the dependent variable in the sample. We need to use inferential methods to draw conclusions about the population. In our example, the sample consists of the 272 observations.

Intercept: The intercept is $\hat{\beta}_0 = -56018.7$. This is the average income when all the independent variables are zero.

As we observed in Chapter 15, it is often misleading to try to interpret this value, particularly if 0 is outside the range of the values of the independent variables (as is the case here).

Age: The relationship between income and age is described by $\hat{\beta}_1 = 505$. From this number, we learn that for each additional year of age in this model, income increases on average by \$505, assuming that the other independent variables in this model are held constant.

Education: The relationship between income and education is measured by the coefficient $\hat{\beta}_2 = 4003$, which specifies that for each additional year of education the income increases on average by \$4003, assuming the constancy of the other independent variables.

Hours of work: The relationship between hours of work per week is expressed by $\hat{\beta}_3 = 753$. We interpret this number as the average increase in annual income for each additional hour of work per week, keeping the other independent variables fixed in this sample.

Spouse's hours of work: The relationship between annual income and a spouse's hours of work per week is described in this sample by $\hat{\beta}_4 = -840$, which we interpret to mean that for each additional hour a spouse works per week, the respondent's income decreases on average by \$840 when the other variables are constant.

Occupation prestige score: In this sample, the relationship between annual income and occupation prestige score is described by $\hat{\beta}_5 = 606$. For each additional unit increase in the occupation prestige score, annual income increases on average by \$606, holding all other variables constant.

Number of children: The relationship between annual income and number of children is expressed by $\hat{\beta}_6 = 244$, which tells us that for each additional child, annual income increases on average by \$244 in this sample. However, this effect is not significant.

Number of years with current job: The coefficient of the last independent variable in this model is $\hat{\beta}_7 = 323$. This number means that for each additional year of job tenure with the current company, annual income increases on average by \$323, keeping the other independent variables constant in this sample. However, this effect is also not significant.

Using the regression equation

We can now produce the prediction interval for a particular value of y , and we can also produce the confidence interval estimate of the expected value of y . Like the other calculations associated with multiple regression, we call on the computer to do the work.

To illustrate, we'll predict the income of a 50-year-old, with 12 years of education, who works 40 hours per week, has a spouse who also works 40 hours per week, has an occupation prestige score of 50, has two children and has worked for the same company for five years.²

$$\begin{aligned}\hat{y}_{\text{INCOME}} &= -56018.7 + 505(50) + 4003.1(12) + 752.8(40) - 839.9(40) + 606(50) + 244.1(2) + 323(5) \\ &= \$46183.15\end{aligned}$$

Using the **XLSTAT** option **Predictions**, we present the Excel output for the point prediction, 95% prediction interval and 95% interval estimate of the expected value of incomes for all people with the given variables in the output below.

	B	C	D	E	F	G	H	I
98	Predictions for the new observations (INCOME):							
99	Observation	Pred (INCOME)	Std. dev. on pred. (Mean)	Lower bound 95% (Mean)	Upper bound 95% (Mean)	Std. dev. on pred. (Observation)	Lower bound 95% (Observation)	Upper bound 95% (Observation)
100	PredObs1	46183.4	3765.8	38768.6	53598.1	32942.6	-18680.2	111047.0

The predicted annual income is \$46 183 for a 50-year-old, with 12 years of education, who works 40 hours per week, has a spouse who works 40 hours per week, has an occupation prestige score of 50, has two children and has worked for the same company for five years.

The prediction interval is [-18680, 111047]. It is so wide as to be completely useless. To be useful in predicting values, the model must be considerably better. The confidence interval estimate of the expected income of a population is [\$38 769, \$53 598].

² There may be slight difference due to rounding off of the estimated coefficients when calculating by hand.

EXERCISES

Learning the techniques

- 16.1** Do the results in the following table allow us to conclude that linear relationships exist between each of x_1 , x_2 and x_3 and y ? (Use $\alpha = 0.05$ and $n = 100$.)

	Coefficient	Standard error
Constant	-30.0	10.0
x_1	-7.5	4.2
x_2	18.0	10.1
x_3	-0.5	0.2

- 16.2** In estimating the regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

with 50 observations, a statistics practitioner obtained the computer output shown in the following table.

The estimated regression equation is

$$\hat{y} = 110.5 + 32.8x_1 - 56.3x_2 + 85.0x_3 - 27.6x_4$$

Variable	Coefficient	Standard error	t-ratio
Constant	110.5	52.1	2.12
x_1	32.8	12.6	2.60
x_2	-56.3	48.5	-1.16
x_3	85.0	69.1	1.23
x_4	-27.6	5.6	-4.93

- a** Do these data allow us to conclude at the 5% significance level that a linear relationship exists between x_1 and y ?
- b** Can we conclude (with $\alpha = 0.01$) that there is a negative linear relationship between x_2 and y ?
- c** Can we conclude at the 10% significance level that there is a positive linear relationship between x_3 and y ?

- 16.3** Suppose that, in an attempt to estimate the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

a researcher obtained 28 observations, which produced the following analysis of variance table:

Source	d.f.	SS	MS
Regression	4	126.30	31.58
Residual	23	269.10	11.70
Total	27	395.40	

Test (with $\alpha = 0.01$) the following hypotheses to confirm the utility of the model:

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$H_A: \text{At least one } \beta_i \neq 0.$$

- 16.4** In calculations undertaken to estimate the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

50 observations of the variables y , x_1 and x_2 produced the following statistics:

$$\text{SST} = 321.2 \quad \text{SSE} = 259.0$$

- a** Find SSR.
- b** Calculate the standard error of estimate.
- c** Calculate the adjusted coefficient of determination.
- d** Test the overall utility of the model, with $\alpha = 0.01$.

- 16.5** A random sample of 100 observations was taken to estimate the regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

A software package was employed to produce the output shown in the accompanying table. Because of a printer malfunction, however, some values from the analysis of variance table are missing; at present, these are replaced by the letters (a) to (f).

Source	d.f.	SS	MS
Regression	(a)	573.6	(e)
Residual	(b)	(d)	(f)
Total	(c)	925.9	

- a** Fill in the missing values.
- b** Test to determine whether the model is useful with $\alpha = 0.05$.

Applying the techniques

- 16.6 Self-correcting exercise.** The owner of a plaster-manufacturing plant wants to predict the monthly demand for plaster (in hundreds of 4 m × 8 m sheets) as a function of the number of building permits issued, the five-year term mortgage rates, and overall economic activity as measured by per capita GDP (in \$'000s). Taking monthly data over the last three years, he produced the results shown below for the regression model

$$y = \beta_0 + \beta_1(\text{PERMITS}) + \beta_2(\text{RATES}) + \beta_3(\text{GDP}) + \varepsilon$$

Variable	Coefficient	Standard error	t-ratio
Constant	5.127	1.325	3.869
Permits	0.062	0.030	2.067
Rates	-1.212	0.659	-1.839
GDP	0.022	0.005	4.400

- a Interpret the coefficients $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_3$.
- b Test to determine whether a linear relationship exists between each of the independent variables and the dependent variable (with $\alpha = 0.05$).

REAL-LIFE APPLICATIONS

Operations management: Location analysis

Location analysis is one function of operations management. Deciding where to locate a plant, warehouse or retail outlet is a crucial decision for any organisation. A large number of variables must be considered in this decision problem. For example, a production facility must be located close to suppliers

of raw resources and supplies, skilled labour and transportation to customers. Retail outlets must consider the type and number of potential customers. In the next example, we describe an application of regression analysis to find profitable locations for some holiday cottages.

16.7 XR16-07 A developer who specialises in holiday cottage properties is considering purchasing a large tract of land adjoining a lake. The current owner of the tract has already subdivided the land into separate building lots and has prepared the lots by removing some of the trees. The developer wants to forecast the value of each lot. From previous experience, she knows that the most important factors affecting the price of the lots are size, number of mature trees and distance to the lake. From a nearby area, she gathers the relevant data for 60 recently sold lots. These data are recorded (Column 1 = price in thousands of dollars, column 2 = lot size in hundreds of square metres, column 3 = number of mature trees, and column 4 = distance to the lake in metres). A multiple

regression analysis was performed and the Excel results are shown below.

- a What is the standard error of estimate? Interpret its value.
- b What is the coefficient of determination? What does this statistic tell you?
- c What is the coefficient of determination, adjusted for degrees of freedom? Why does this value differ from the coefficient of determination? What does this tell you about the model?
- d Test the overall utility of the model.
- e Interpret each of the coefficients.
- f Test to determine whether each of the independent variables is linearly related to the price of the lot.

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.4924					
5	R Square	0.2425					
6	Adjusted R Square	0.2019					
7	Standard Error	40.2435					
8	Observations	60					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	3	29029.7	9676.57	5.9749	0.0013	
13	Residual	56	90694.3	1619.54			
14	Total	59	119724.0				
15							
16		Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
17	Intercept	51.3912	23.5165	2.1853	0.0331	4.2820	98.5004
18	Lot-size	0.6999	0.5589	1.2524	0.2156	-0.4196	1.8194
19	Trees	0.6788	0.2293	2.9603	0.0045	0.2195	1.1382
20	Distance	-0.3784	0.1952	-1.9380	0.0577	-0.7695	0.0127

- 16.8 XR16-08** After analysing the results of Exercise 15.20, Pat decided that a certain amount of studying could actually improve final grades. However, too much studying would not be warranted, since Pat's ambition (if that's what one could call it) was to ultimately graduate with the absolute minimum level of work. Pat was registered in a statistics subject that had only three weeks to go before the final exam, and for which the final grade was determined in the following way:

Assignment	20%
Mid-semester test	30%
Final exam	50%

In order to determine how much work to do for the remaining three weeks, Pat needed to be able to predict the final exam mark on the basis of the assignment mark and the mid-semestern mark. Pat's marks on these were 12/20 and 14/30 respectively. Accordingly, Pat undertook the following analysis. The final exam mark, assignment mark and mid-semestern test mark for 30 students who took the statistics subject last year were collected. These data are recorded in columns 1 to 3, respectively. A multiple regression analysis was performed using Excel with the following results.

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.8734					
5	R Square	0.7629					
6	Adjusted R Square	0.7453					
7	Standard Error	3.7524					
8	Observations	30					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	2	1223.1837	611.5919	43.4343	0.0000	
13	Residual	27	380.1830	14.0809			
14	Total	29	1603.3667				
15							
16		Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
17	Intercept	13.0091	3.5278	3.6876	0.0010	5.7707	20.2475
18	Assignment	0.1940	0.2004	0.9678	0.3417	-0.2173	0.6053
19	Mid-SEM	1.1121	0.1219	9.1196	0.0000	0.8619	1.3623

- a What is the standard error of estimate? Briefly describe how you interpret this statistic.
- b What is the coefficient of determination? What does this statistic tell you?
- c What is the coefficient of determination, adjusted for degrees of freedom? What do this statistic and the one alluded to in part (b) tell you about the model?
- d Test the overall utility of the model.
- e Interpret each of the coefficients.
- f Can Pat infer from these results that the assignment mark is linearly related to the final grade?
- g Can Pat infer from these results that the mid-semestern mark is linearly related to the final grade?
- h Which variable appears to be a better predictor of the final exam mark? Explain. Suggest several reasons for your answer.

- 16.9 XR16-09** The manager of a company that manufactures plasterboard wants to analyse the factors that affect demand for his product. Plasterboard is used to construct walls in houses and offices. Consequently, the manager decides to develop a regression model in which the dependent variable is monthly sales of plasterboard (in hundreds of 4m × 8m sheets) and the independent variables are:

- number of building permits used in the state
- five-year mortgage rates (in percentage points)
- vacancy rate in apartments (in percentage points)
- vacancy rate in office buildings (in percentage points).

To estimate a multiple regression model, the manager took the monthly observations from the past two years. The data are recorded in columns 1 to 5 respectively. A computer was used to produce the output below:

	A	B	C	D	E	F	G
1	<i>Regression Statistics</i>						
2	Multiple R	0.9453					
3	R Square	0.8935					
4	Adjusted R Square	0.8711					
5	Standard Error	40.1324					
6	Observations	24					
7							
8	ANOVA						
9		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
10	Regression	4	256793.4	64198.4	39.86	0.0000	
11	Residual	19	30601.6	1610.6			
12	Total	23	287395.0				
13							
14		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
15	Intercept	-111.83	134.34	-1.832	0.416	-393.01	169.35
16	Permits	4.763	0.395	12.057	0.000	3.936	5.590
17	Mortgage	16.989	15.159	1.121	0.276	-14.739	48.716
18	A_Vacancy	-10.528	6.394	-1.646	0.116	-23.911	2.856
19	O_Vacancy	1.308	2.791	0.469	0.645	-4.533	7.149

- a What is the standard error of estimate? Can you use this statistic to assess the fitness of the model? If so, how?
- b What is the coefficient of determination, and what does it tell you about the regression model?
- c What is the coefficient of determination, adjusted for degrees of freedom? What do this statistic and the statistic referred to in part (b) tell you about how well this model fits the data?
- d Test the overall utility of the model.
- e Interpret each of the coefficients.
- f Test to determine whether each of the independent variables is linearly related to plasterboard demand.
- g Which independent variable appears to be the best predictor of plasterboard demand? Explain.

16.10 XR16-10 Suppose that the statistics practitioner who did the analysis described in Exercise 15.21 wanted to investigate other factors that determine height. As part of the same study, she recorded the heights

of the mothers in column 3. (Columns 1 and 2 contain the data from Exercise 15.21.) The multiple regression outputs are shown in the table below.

- a What is the standard error of estimate? What does this statistic tell you?
- b What is the coefficient of determination? What does this statistic tell you?
- c What is the coefficient of determination, adjusted for degrees of freedom? What do this statistic and the one referred to in part (b) tell you about how well the model fits the data?
- d Test the overall utility of the model. What does the test result tell you?
- e Interpret each of the coefficients.
- f Do these data allow the statistics practitioner to infer that the heights of sons and their fathers are linearly related?
- g Do these data allow the statistics practitioner to infer that the heights of sons and their mothers are linearly related?

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	<i>Regression Statistics</i>						
4	Multiple R	0.5209					
5	R Square	0.2713					
6	Adjusted R Square	0.2676					
7	Standard Error	8.0453					
8	Observations	400					
9							
10	ANOVA						
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	2	9568.11	4784.06	73.91	0.0000	
13	Residual	397	25696.87	64.73			
14	Total	399	35264.98				
15							
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>p-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	100.3469	8.7024	11.5309	0.0000	83.2383	117.4555
18	Father	0.5068	0.0433	11.6955	0.0000	0.4216	0.5920
19	Mother	-0.0879	0.0545	-1.6135	0.1074	-0.1951	0.0192

Computer applications

The following exercises require the use of a computer and software.

16.11 XR16-11 Deciding where to locate a new retail store is one of the most important decisions that a manager can make. The manager of a chain of pizza shops plans to use a regression model to help select a location for a new shop. He decides to use annual gross revenue as a measure of success, which is the dependent variable. The manager believes that determinants of success include the following variables:

- Number of people living within 1 km of the store (PEOPLE)
- Mean income of households within 1 km of the store (INCOME)
- Number of competitors within 1 km of the store (COMPTORS)
- Average price of a large pizza (PRICE)

The manager randomly selects 50 pizza shops and records the values of each of the variables listed above plus the annual gross revenue (REVENUE). He proposes the following multiple regression model:

$$\text{REVENUE} = \beta_0 + \beta_1(\text{PEOPLE}) + \beta_2(\text{INCOME}) \\ + \beta_3(\text{COMPTORS}) + \beta_4(\text{PRICE}) + \varepsilon$$

- a Estimate the model and find the estimated regression equation.
- b What are the coefficient of determination (R^2) and the coefficient of determination adjusted for degrees of freedom? What do these statistics tell you about the regression equation?
- c What does the standard error of estimate tell you about the regression model?
- d Test the overall utility of the model. Discuss your results.
- e Which independent variables are linearly related to the revenue? Explain.

16.12 XR16-12 The marketing manager for a chain of hardware stores needed more information about the effectiveness of the three types of advertising that the chain used. These are localised direct mailing (in which flyers describing sales and featured products are distributed to homes in the area surrounding a store), newspaper advertising and local television advertisements. To determine which type is most effective, the manager collected one week's data from 25 randomly selected stores. For each store, the following variables were recorded:

- Weekly gross sales
- Weekly expenditures on direct mailing

- Weekly expenditures on newspaper advertising
 - Weekly expenditures on television commercials
- All variables were recorded in thousands of dollars.

- a Find the regression equation.
- b What are the coefficient of determination (R^2) and the coefficient of determination adjusted for degrees of freedom (\bar{R}^2)? What do these statistics tell you about the regression equation? Using the ANOVA table, calculate R^2 , (\bar{R}^2) and s_e and confirm that these values are the same as those shown in the computer output.
- c What does the standard error of estimate tell you about the regression model?
- d Test the overall utility of the model. Discuss your results.
- e Which independent variables are linearly related to weekly gross sales? Explain.
- f Predict next week's gross sales if a local store spent \$800 on direct mailing, \$1200 on newspaper advertisements and \$2000 on television commercials.
- g Estimate the mean weekly gross sales for all stores that spend \$800 on direct mailing, \$1200 on newspaper advertising and \$2000 on television commercials.

16.13 XR16-13 In an effort to explain to customers why their electricity bills have been so high lately, and how, specifically, they could save money by reducing the thermostat settings on both space heaters and water heaters, an electric utility company has collected total kilowatt consumption figures for last year's winter months, as well as thermostat settings on space and water heaters, for 100 homes. The data are recorded in columns 1 (consumption), 2 (space heater thermostat setting) and 3 (water heater thermostat setting).

- a Determine the regression equation.
- b Determine the standard error of estimate, and comment about what it tells you.
- c Determine the coefficient of determination, and comment about what it tells you.
- d Test whether the model is useful.
- e Predict the electricity consumption of a house whose space heater thermostat is set at 21 and whose water heater thermostat is set at 54.
- f Estimate the average electricity consumption for houses whose space heater thermostat is set at 21 and whose water heater thermostat is set at 54.

REAL-LIFE APPLICATIONS

Human resources management: Severance pay

In most firms, the entire issue of compensation falls into the domain of the human resources manager. The manager must ensure that the method used to determine compensation contributes to the firm's objectives. Moreover, the firm needs to ensure that discrimination or bias of any kind is not a factor. Another function of the personnel manager is to

develop severance packages for employees whose services are no longer needed because of downsizing or mergers. The size and nature of the package is rarely part of any working agreement and must be determined by a variety of factors. Regression analysis is often useful in this area.

- 16.14 XR16-14** When one company buys another company, it is not unusual for some workers to be made redundant. The severance benefits offered to the laid-off workers are often the subject of dispute. During the Ansett Airlines crisis of the late 1990s, Airline A, a leading airline in Australia (which cannot be named for legal reasons), bought one of Ansett's regional airlines, Ansett R, and subsequently terminated the contracts of 20 of Ansett R's employees. As part of the buyout agreement, it was promised that the severance packages offered to the former Ansett R employees would be equivalent to those offered to Airline A employees whose contract had been terminated in the past year. Thirty-six-year-old Bill Smith, an Ansett R employee for the past 10 years and earning \$32 000 per year, was one of those let go. His severance package included an offer of five weeks' severance pay. Bill complained that this offer was less than that offered to Airline A's employees when they were laid off, in contravention of the buyout agreement. A statistics practitioner was called in to settle the dispute. The statistics practitioner was told that severance packages should be determined by three factors: age, length of service with the company, and salary. To determine how generous the severance package had been, a random sample of 50 Airline A ex-employees was taken. For each, the following variables were recorded:
- Number of weeks of severance pay
 - Age of employee
 - Number of years with the company
 - Annual salary (in thousands of dollars)
- a** Determine the regression equation. Interpret the coefficients.
- b** Comment on how well the model fits the data.
- c** Do all the independent variables belong in the equation? Explain.

- d** Perform an analysis to determine whether Bill is correct in his assessment of the severance package.

- 16.15 XR16-15** The admissions officer of a private university is trying to develop a formal system of deciding which students to admit to the university. She believes that determinants of success include the standard variables – higher school certificate (HSC) grades and tertiary admission exam (TAE) score. However, she also believes that students who have participated in extracurricular activities are more likely to succeed than those who have not. To investigate the issue, she randomly sampled 100 fourth-year university students and recorded the following variables:

- Grade point average (GPA) for first three years at university
 - HSC grade from high school
 - TAE score
 - Number of hours per week on average spent in organised extracurricular activities in the last year of high school
- a** Develop a model that helps the admissions officer decide which students to admit, and use the computer to generate the usual statistics.
- b** What is the standard error of estimate? What does this statistic tell you?
- c** What is the coefficient of determination? Interpret its value.
- d** What is the coefficient of determination, adjusted for degrees of freedom? Interpret its value.
- e** Test whether the model is useful.
- f** Interpret each of the coefficients.
- g** Test to determine whether each of the independent variables is linearly related to the dependent variable.

- h** Predict with 95% confidence the GPA for the first three years of university for a student whose HSC grade is 10, whose TAE score is 600, and who worked an average of two hours per week on organised extracurricular activities in the last year of high school.
- i** Estimate with 90% confidence the mean GPA for the first three years of university for all students who achieved an HSC grade of 8, a TAE score of 550, and who worked an average of 10 hours per week on organised extracurricular activities in the last year of high school.

16.3 Regression diagnostics – II

In Section 15.7 we discussed how to determine whether the required conditions are satisfied. The same procedures can be used to diagnose problems in the multiple regression model. Here is a brief summary of the diagnostic procedure we described in Chapter 15.

16.3a Informal diagnostic procedures

Calculate the residuals and check the following:

- 1** *Is the error variable non-normal?* Draw the histogram of the residuals.
- 2** *Is the error variance constant?* Plot the residuals versus the predicted values of $y(\hat{y})$.
- 3** *Are the errors independent (time-series data)?* Plot the residuals versus the time periods. Also, check for outliers:
- 4** *Are there observations that are inaccurate or do not belong to the target population?* Double-check the accuracy of outliers and influential observations using a scatter diagram or other descriptive measures.

If the error is non-normal and/or the variance is not a constant, there are several remedies that can be attempted. These are described at the end of this section.

Outliers and influential observations are checked by examining the data in question to ensure accuracy.

Non-independence of a time series can sometimes be detected by graphing the residuals and the time periods and looking for evidence of autocorrelation. In Section 16.4, we introduce a formal test, the **Durbin–Watson test**, which tests for one form of autocorrelation. We will offer a corrective measure for non-independence.

There is another problem that is applicable to multiple regression models only. Multicollinearity is a condition in which the independent variables are highly correlated. Multicollinearity distorts the t -tests of the coefficients, making it difficult to determine whether any of the independent variables are linearly related to the dependent variable. It also makes interpreting the coefficients problematic. We will discuss this condition and its remedy next.

16.3b Multicollinearity

Durbin–Watson test

A test for autocorrelation.

multicollinearity

Condition in which the independent variables in a regression model are correlated.

Multicollinearity (also called *collinearity* and *intercorrelation*) is a condition that exists when the independent variables are correlated with one another. The adverse effect of multicollinearity is that the estimated regression coefficients $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ tend to have large sampling variability. That is, the standard errors are unusually large. Consequently, when the coefficients are tested, the t -statistics will be small, which implies that there is no linear relationship between the affected independent variables and the dependent variable. In some cases, this inference will be wrong. Fortunately, multicollinearity does not affect the F -test of the analysis of variance. We will illustrate the effects and remedy with the following example.

EXAMPLE 16.8

Modelling the sale price of a house

XM16-08 A real estate agent wanted to develop a model to predict the selling price of a house. The agent believed that the most important variables in determining the price of a house are its size, number of bedrooms and lot size. Accordingly, he took a random sample of 100 homes that had been recently sold and recorded the selling price, the number of bedrooms, the size (in squares) and the lot size (in square metres). Analyse the relationship among the four variables.

Solution

The proposed multiple regression model is

$$\text{PRICE} = \beta_0 + \beta_1(\text{BEDROOMS}) + \beta_2(\text{H-SIZE}) + \beta_3(\text{LOT-SIZE}) + \varepsilon$$

Using the commands given in Example 16.1, we estimate the multiple regression model using Excel Data Analysis or XLSTAT. The regression output (shown below) reveals that none of the independent variables is significantly related to the selling price (as p -value $> \alpha = 0.05$). However, the F -test ($F = 40.73$ and p -value = 0) indicates that the complete model fits quite well.

Excel output for Example 16.8³

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.7483					
5	R Square	0.5600					
6	Adjusted R Square	0.5462					
7	Standard Error	50045.4					
8	Observations	100					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	3	3.06007E+11	1.02002E+11	40.73	4.57E-17	
13	Residual	96	2.40436E+11	2504543586			
14	Total	99	5.46443E+11				
15							
16		Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
17	Intercept	175435.2	28353.5	6.187	0.000	119154.0	231716.4
18	Bedrooms	4612.2	13988.4	0.330	0.742	-23154.6	32378.9
19	H-Size	1485.9	1059.6	1.402	0.164	-617.3	3589.2
20	Lot-Size	-87.3	340.5	-0.256	0.798	-763.1	588.6

If we run three simple regression models for which the independent variable is (1) the number of bedrooms, (2) the house size and (3) the lot size, the output below is produced. This result tells us that each of the independent variables is strongly related to selling price (as the p -values of the slope coefficients for all three regressions are almost 0, rejecting the null hypothesis of no linear relationship).



³ Using the XLSTAT commands provided in Example 16.2, we can obtain the regression output. As the XLSTAT regression output is similar to the one from EXCEL Data Analysis, it is not provided here.

Excel output for Example 16.8 (simple linear regressions)

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.6454					
5	R Square	0.4166					
6	Adjusted R Square	0.4106					
7	Standard Error	57037.3					
8	Observations	100					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	2.27624E+11	2.27624E+11	69.97	4.20E-13	
13	Residual	98	3.18819E+11	3253258833			
14	Total	99	5.46443E+11				
15							
16		Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
17	Intercept	150844.7	31283.1	4.822	0.000	88764.4	212925.1
18	Bedrooms	70878.0	8473.5	8.365	0.000	54062.7	87693.4

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.7478					
5	R Square	0.5591					
6	Adjusted R Square	0.5547					
7	Standard Error	49579.9					
8	Observations	100					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	3.0554E+11	3.05543E+11	124.30	3.96657E-19	
13	Residual	98	2.4090E+11	2458165361			
14	Total	99	5.4644E+11				
15							
16		Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
17	Intercept	180132.8	21042.9	8.560	0.000	138373.9	221891.7
18	H-Size	1284.1	115.2	11.149	0.000	1055.5	1512.6

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.7409					
5	R Square	0.5489					
6	Adjusted R Square	0.5443					
7	Standard Error	50153.2					
8	Observations	100					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	2.9994E+11	2.9994E+11	119.24	1.23E-18	
13	Residual	98	2.46503E+11	2515341627			
14	Total	99	5.46443E+11				
15							
16		Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
17	Intercept	177880.6	21673.8	8.207	0.000	134869.7	220891.4
18	Lot-Size	419.6	38.4	10.920	0.000	343.4	495.9

The *t*-tests in the multiple regression model imply that no independent variable is a factor in determining the selling price. The three simple linear regression models contradict this conclusion. They tell us that the number of bedrooms, the house size and the lot size are *all* linearly related to the price. How do we account for this contradiction? The answer is that the three independent variables are correlated with each other and consequently the estimated results suffer from multicollinearity. It is reasonable to believe that larger houses have more bedrooms and are situated on larger lots, and that smaller houses have fewer bedrooms and are located on smaller lots. To confirm this belief, we calculated the correlation among the three independent variables. In addition, the results of the *t*-test for correlation between price and the three independent variables show that all three correlations are significant at the 5% level.

Correlation coefficients

	A	B	C	D	E
1		Price	Bedrooms	H-Size	Lot-Size
2	Price	1			
3	Bedrooms	0.6454	1		
4	H-Size	0.7478	0.8465	1	
5	Lot-Size	0.7409	0.8374	0.9936	1

	A	B	C	D
1	t-Test of Correlation Coefficient (Price and Bedrooms)			
2				
3	Sample correlation	0.6454	t Stat	8.36
4	Sample size	100	P(T<=t) one-tail	2.10E-13
5	Alpha	0.05	t Critical one-tail	1.66
6			P(T<=t) two-tail	4.21E-13
7			t Critical two-tail	1.98

	A	B	C	D
1	t-Test of Correlation Coefficient (Price and Bedrooms)			
2				
3	Sample correlation	0.7478	t Stat	11.15
4	Sample size	100	P(T<=t) one-tail	1.97E-19
5	Alpha	0.05	t Critical one-tail	1.66
6			P(T<=t) two-tail	3.94E-19
7			t Critical two-tail	1.98

	A	B	C	D
1	t-Test of Correlation Coefficient (Price and Bedrooms)			
2				
3	Sample correlation	0.7409	t Stat	10.92
4	Sample size	100	P(T<=t) one-tail	6.15E-19
5	Alpha	0.05	t Critical one-tail	1.66
6			P(T<=t) two-tail	1.23E-18
7			t Critical two-tail	1.98

The coefficient of correlation between number of bedrooms and house size is 0.846; the correlation between number of bedrooms and lot size is 0.837; the correlation between house size and lot size is 0.994. The level of correlation is high here and so is the degree of multicollinearity. Therefore, in the multiple regression model, multicollinearity affects the *t*-tests so that they indicated that none of the independent variables is linearly related to price when, in fact, all are.

Another problem caused by multicollinearity is the interpretation of the coefficients. We interpret the coefficients as measuring the change in the dependent variable when the corresponding independent variable increases by one unit while all the other independent variables are held constant. This interpretation may be impossible when the independent variables are highly correlated, because when the independent variable increases by one unit, some or all of the other independent variables will change. Hence, the assumption that



while one independent variable increases all other independent variables are held constant is not possible. In the multiple regression model in this example, the coefficient of BEDROOMS is 4612. Without multicollinearity we would interpret this coefficient to mean that for each additional bedroom the average price increases by \$4612, provided that the other variables are held constant. However, as BEDROOMS is correlated with H-SIZE and LOT-SIZE, it is impossible to increase BEDROOMS by 1 and hold the other variables constant.

Example 16.8 raises two important questions for the statistics practitioner. First, how do we recognise the problem when it occurs, and second, how do we avoid or correct it?

Multicollinearity exists in virtually all multiple regression models. In fact, finding two completely uncorrelated variables is rare. The problem becomes serious, however, only when two or more independent variables are highly correlated. Unfortunately, we do not have a critical value that indicates when the correlation between two independent variables is large enough to cause problems. To complicate the issue, multicollinearity also occurs when a combination of several independent variables is correlated with another independent variable or with a combination of other independent variables. Consequently, even with access to all of the correlation coefficients, determining when the multicollinearity problem has reached the serious stage may be extremely difficult.

Minimising the effect of multicollinearity is often easier than recognising it. The statistics practitioner must try to include independent variables that are independent of each other. For example, the real estate agent wanted to include house size, the number of bedrooms and the lot size, three variables that are clearly related. Rather than developing a model that uses all such variables, the statistics practitioner may choose to include only house size, plus several other variables that measure other aspects of a house's value.

Another alternative is to use a stepwise regression package. *Forward stepwise regression* brings independent variables into the equation one at a time. Only if an independent variable improves the fitness of the model is it included. If two variables are strongly correlated, the inclusion of one of them in the model makes the second one unnecessary. *Backward stepwise regression* starts with all the independent variables included in the equation and removes variables if they are not strongly related to the dependent variable. Because the stepwise technique excludes redundant variables, it minimises multicollinearity.

16.3c Remedyng violations of required conditions

The most commonly used method to remedy non-normality or heteroscedasticity is to transform the dependent variable. There are three points to note about this procedure. First, the actual form of the transformation depends on which condition is unsatisfied and on the specific nature of the violation. Because there are many different ways to violate the required conditions of the statistical techniques, the list of transformations given here is unavoidably incomplete. Second, these transformations can be useful in improving the model. That is, if the linear model appears to be quite poor, we often can improve the fitness of the model by transforming y . Third, many computer software systems allow us to make transformations quite easily. You might want to experiment to see the effect these transformations have on your statistical results.

Here is a brief list of the most commonly used transformations:

- 1 *Log transformation*: $y' = \log y$ (provided $y > 0$). The log transformation is used when (a) the variance of the error variable increases as y increases, or (b) the distribution of the error variable is positively skewed.
- 2 *Squared transformation*: $y' = y^2$. Use this transformation when (a) the variance is proportional to the expected value of y , or (b) the distribution of the error variable is negatively skewed.

- 3 *Square-root transformation:* $y' = \sqrt{y}$ (provided that $y > 0$). The square-root transformation is helpful when the variance is proportional to the expected value of y .
- 4 *Reciprocal transformation:* $y' = 1/y$. When the variance appears to significantly increase when y increases beyond some critical value, the reciprocal transformation is recommended.

The following example will illustrate how we diagnose a violation of the required condition, its consequences, and how we remedy the problem.

EXAMPLE 16.9

LO5

The effect of time limits on quiz marks

XM16-09 A statistics lecturer wanted to know whether time limits on quizzes affected the marks on the quiz. Accordingly, he took a random sample of business statistics students and split them into five groups of 20 students each. All students took a quiz that involved simple manual calculations. Each group was given a different time limit. Group 1 was limited to 40 minutes, group 2 to 45 minutes, group 3 to 50 minutes, group 4 to 55 minutes, and group 5 to 60 minutes. The quizzes were marked (out of 40) and recorded. (Column 1 stores the time limits, and column 2 stores the marks.) Conduct a complete regression analysis, including diagnostics.

Solution

The following regression model was postulated.

$$\text{MARK} = \beta_0 + \beta_1(\text{TIME}) + \varepsilon$$

The Excel output is shown below.

Using the computer

Excel output for Example 16.9

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.8625					
5	R Square	0.7440					
6	Adjusted R Square	0.7414					
7	Standard Error	2.3046					
8	Observations	100					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	1512.5	1512.5	284.7743	0.0000	
13	Residual	98	520.5	5.3112			
14	Total	99	2033				
15							
16		Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
17	Intercept	-2.2000	1.6458	-1.3367	0.1844	-5.4661	1.0661
18	Time	0.5500	0.0326	16.8753	0.0000	0.4853	0.6147

The regression equation is

$$\text{MARK} = -2.2 + 0.55(\text{TIME})$$

The standard error of estimate, the coefficient of determination and the t -test of β_1 (and the F -test) all indicate a relatively good model. The residuals and the predicted values were calculated. The histogram of the residuals and the plot of the residuals versus the predicted values of y were produced by Excel and are exhibited in

Figures 16.5 and 16.6.

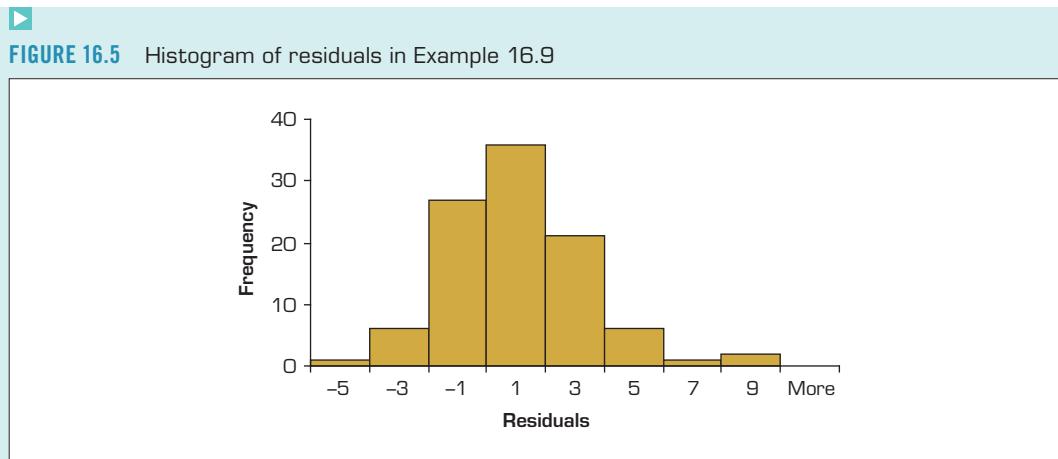
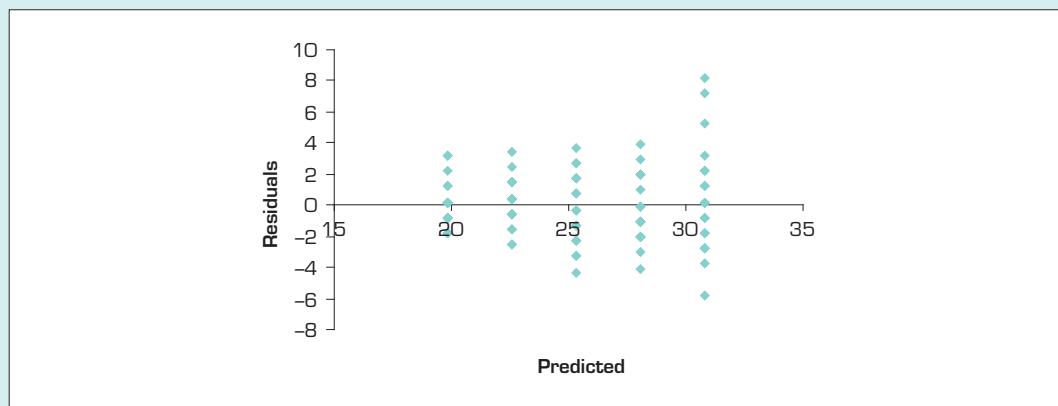


FIGURE 16.5 Histogram of residuals in Example 16.9



The error variable appears to be normal. However, the variance of the errors is clearly not constant; it increases as the predicted marks increase. The remedy we will apply is to transform the dependent variable. We will attempt the following two transformations.

- 1 $y' = \log_e(y)$. We will label the new variable LOGMARK.
- 2 $y' = 1/y$. We will label the new variable 1/MARK.

Once again, we use our software package to estimate the regression equation. The printout appears below.

1 Transformed data $y' = \log_e(y)$

Excel output for Example 16.9 (LOGMARK)

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.8783					
5	R Square	0.7714					
6	Adjusted R Square	0.7691					
7	Standard Error	0.0844					
8	Observations	100					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	2.3579	2.3579	330.7181	0.0000	
13	Residual	98	0.6987	0.0071			
14	Total	99	3.0566				
15							
16		Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
17	Intercept	2.1296	0.0603	35.3163	0.0000	2.0099	2.2492
18	Time	0.0217	0.0012	18.1857	0.0000	0.0193	0.0241

The histogram of the residuals (**Figure 16.7**) indicates that the error variable may be normal. The plot of the residuals versus the predicted values of the dependent variable, $\log_e(y)$ (**Figure 16.8**), shows some change in the error variance. However, the change is smaller than in the original model. Thus, the transformation has decreased the degree of heteroscedasticity.

FIGURE 16.7 Histogram of residuals in Example 16.9 (LOGMARK)

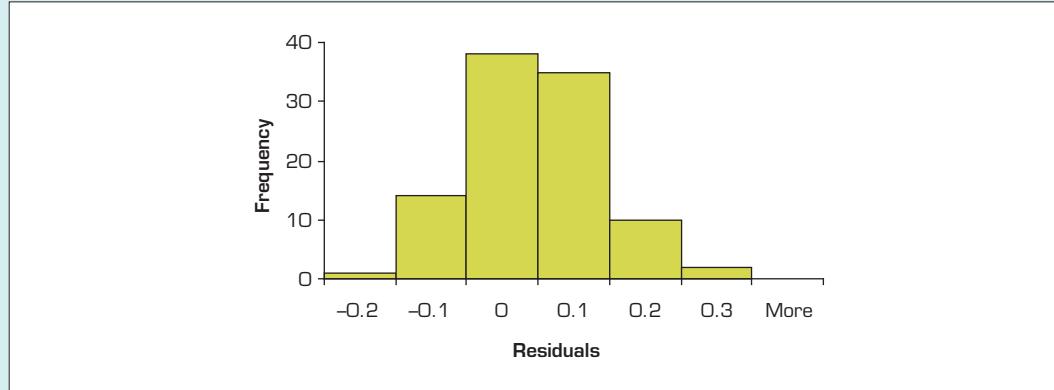
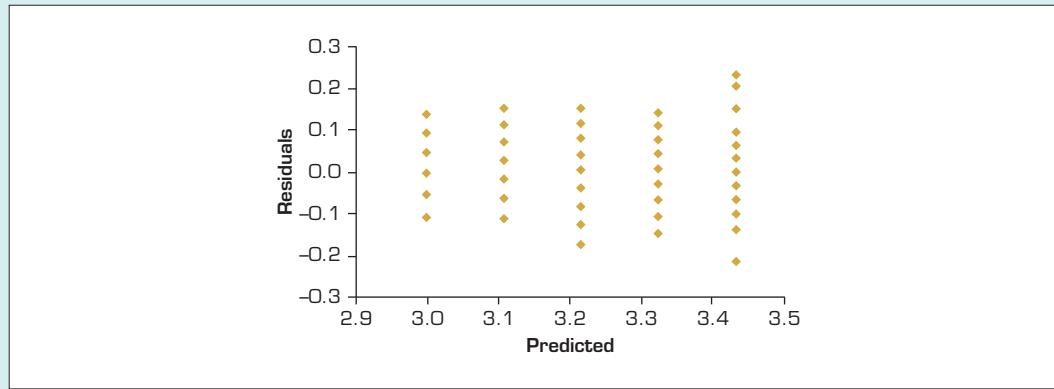


FIGURE 16.8 Plot of residuals versus predicted values in Example 16.9 (LOGMARK)



2 Transformed data $y' = 1/y$

Excel output for Example 16.9 (1/MARK)

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.88198					
5	R Square	0.77788					
6	Adjusted R Square	0.77562					
7	Standard Error	0.00335					
8	Observations	100					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	0.0038	0.0038	343.2104	0.0000	
13	Residual	98	0.0011	0.0000			
14	Total	99	0.0049				
15							
16		Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
17	Intercept	0.0846	0.0024	35.4007	0.0000	0.0798	0.0893
18	Time	-0.0009	0.0000	-18.5259	0.0000	-0.0010	-0.0008



As can be seen from the histogram of the residuals (**Figure 16.9**) and the plot of the residuals versus the predicted values of the dependent variable, $1/y$ (**Figure 16.10**), the problem of heteroscedasticity has been resolved. However, the error variable does not appear to be normal.

FIGURE 16.9 Histogram of residuals in Example 16.9 (1/MARK)

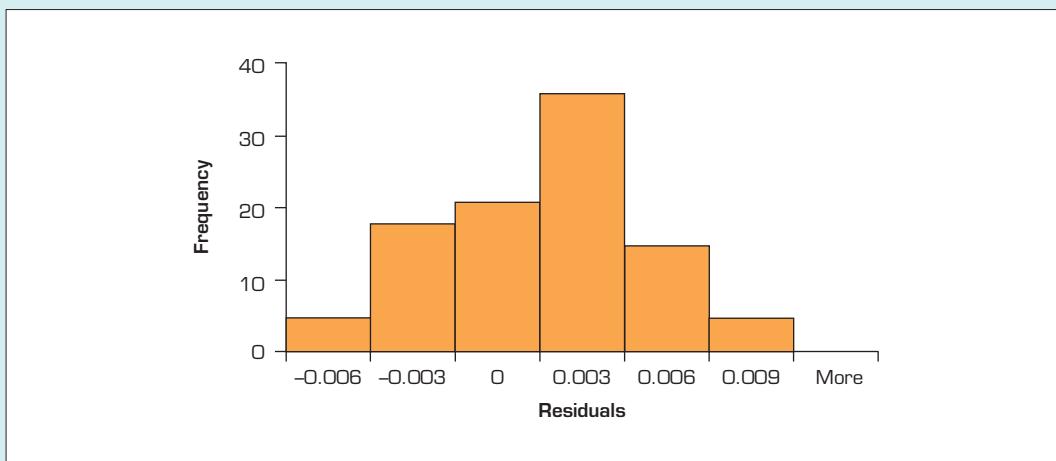
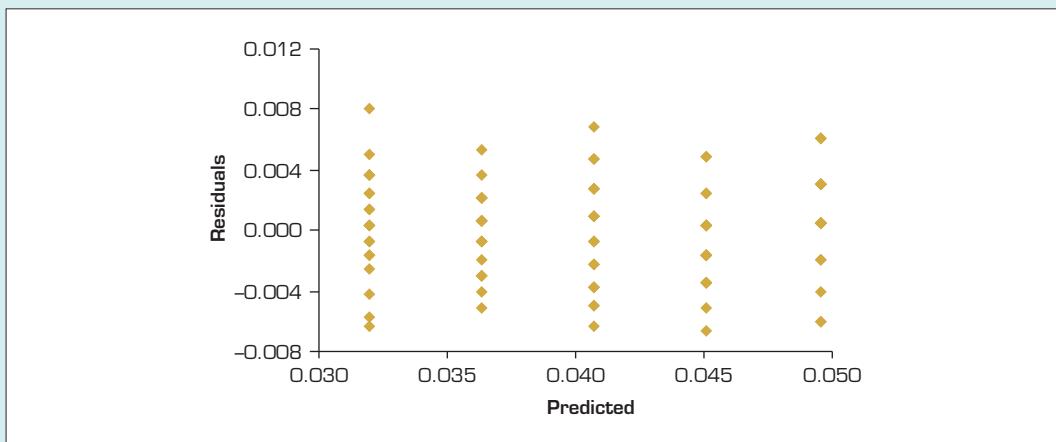


FIGURE 16.10 Plot of residuals versus predicted values in Example 16.9 (1/MARK)



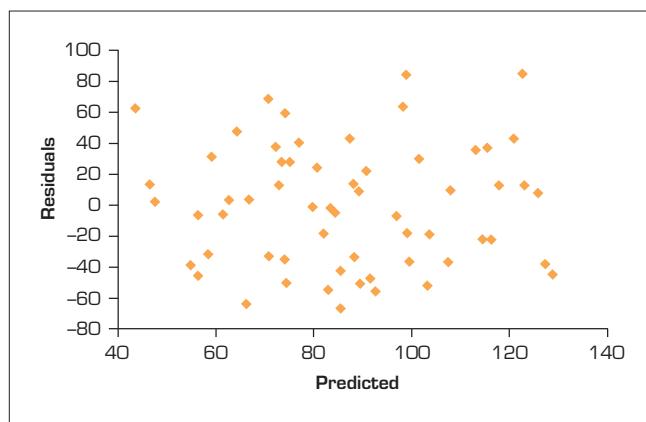
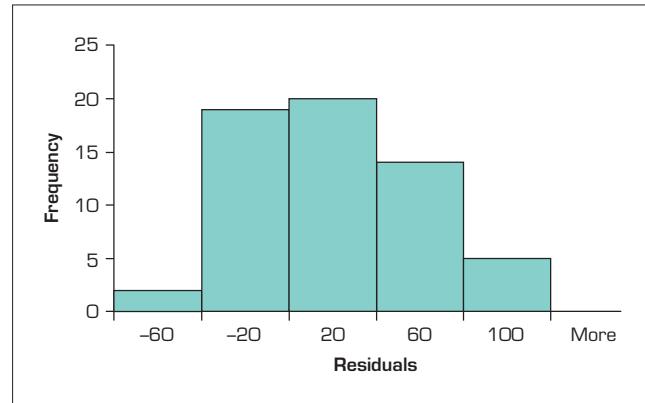
Thus, the logarithmic transformation is judged to be superior. By remedying a violation of the required condition, we have improved the fit. As you can see, both transformed dependent variable models have larger coefficients of determination and F -statistics. (Note that we cannot use the standard error of estimate to make the comparison because the dependent variables are different.)

In practice, statistics practitioners often experiment with different transformations to determine which one works best. Ideally, we look for transformations in which the required conditions are well satisfied and whose fit is best.

EXERCISES

Learning the techniques

- 16.16** Refer to Exercise 16.7. The residuals and predicted values for the regression equation were determined. The histogram of the residuals and the graph of the residuals versus the predicted values are shown below.
- Does it appear that the normality requirement is violated? Explain.
 - Is the error variable variance constant? Explain.

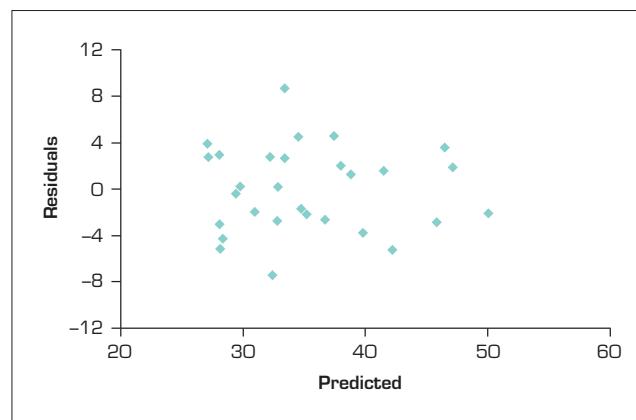
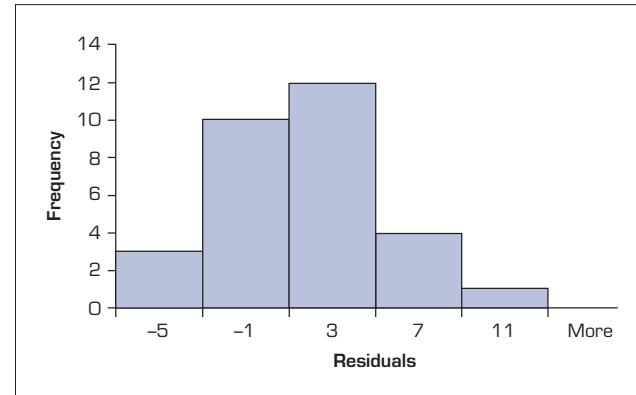


- 16.17** Refer to Exercise 16.7. The correlations for each pair of independent variables are shown below.
- What do these correlations tell you about the independent variables?
 - What do these statistics tell you about the *t*-tests of the coefficients in the multiple regression model?

Pair of independent variables	Correlation
LOT-SIZE and TREES	0.286
LOT-SIZE and DISTANCE	-0.189
TREES and DISTANCE	0.079

- 16.18** Refer to Exercise 16.8. The histogram of the residuals and the graph of the residuals and the predicted values are shown below.

- Does it appear that the normality requirement is violated? Explain.
- Is the error variable variance constant? Explain.



- 16.19 XR16-19** Each of the following pairs of columnar values represents an actual value of y and a predicted value of \hat{y} (based on a regression model). Graph the predicted values of y versus the residuals. In each case, determine from the graph whether the requirement that σ_e^2 is constant is satisfied.

a

	y	\hat{y}		y	\hat{y}
1	155	143		9	125
2	112	108		10	161
3	163	180		11	189
4	130	133		12	102
5	143	146		13	142
6	182	193		14	149
7	160	140		15	180
8	104	101			158

b

	y	\hat{y}		y	\hat{y}
1	10	7	8	23	22
2	22	21	9	11	14
3	29	29	10	27	27
4	15	13	11	19	17
5	24	25	12	26	27
6	13	16	13	20	22
7	17	19	14	14	11

c

	y	\hat{y}		y	\hat{y}
1	46	48	10	56	54
2	40	43	11	62	65
3	53	54	12	44	46
4	60	63	13	49	47
5	52	49	14	61	57
6	59	56	15	42	45
7	45	41	16	57	62
8	55	53	17	50	51
9	47	44			

16.20 XR16-20 Each of the following sets of data represents the actual and predicted values of a time series.

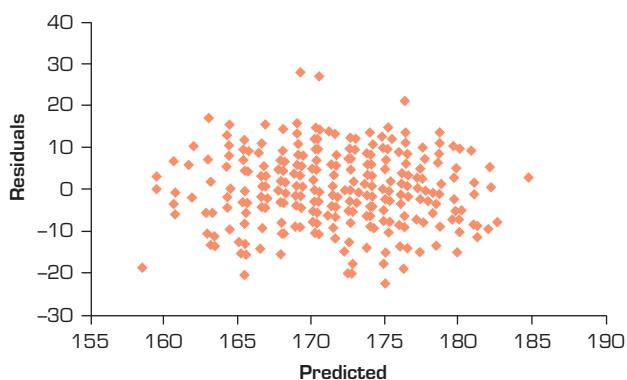
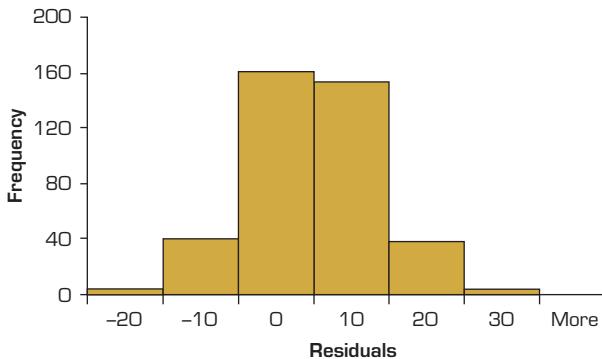
- a Graph the residuals versus the time periods to determine whether the requirement that the values of ϵ are independent of one another is satisfied.
- b Graph the predicted values of y versus the residuals to determine whether the requirement that σ_ϵ^2 is constant is satisfied.

Period	y	\hat{y}	Period	y	\hat{y}
1	60	58	7	52	55
2	71	68	8	65	68
3	53	57	9	72	67
4	59	60	10	54	53
5	75	71	11	61	65
6	50	48	12	66	68

Period	y	\hat{y}
1	325	303
2	265	257
3	350	326
4	375	361
5	290	272
6	280	304
7	360	371
8	250	270
9	300	309
10	330	352

16.21 Refer to Exercise 16.10. The histogram of the residuals and the graph of the residuals and the predicted values are shown below.

- a Does it appear that the normality requirement is violated? Explain.



- b Is the error variable variance constant? Explain.

Applying the techniques

16.22 XR16-22 Self-correcting exercise. The least squares regression line is $\hat{y} = 145.786 + 0.4777x$. The values of \hat{y}_i and e_i were calculated, with the following results.

i	\hat{y}_i	e_i	i	\hat{y}_i	e_i
1	211.09	-9.09	9	284.50	9.50
2	219.19	-7.19	10	293.56	9.44
3	229.20	-8.20	11	308.33	4.67
4	239.21	-6.21	12	324.54	-1.54
5	245.41	-0.41	13	337.41	-2.41
6	252.56	6.44	14	356.00	-6.00
7	262.57	10.43	15	377.93	-9.93
8	274.49	10.51			

- a Draw a histogram of the residuals to determine whether the error variable is normally distributed.
- b Plot the residuals versus the predicted value of y to see whether the requirement that σ_ϵ^2 is a constant is satisfied.
- c Plot the residuals versus time to examine the required condition that the errors be independent.

- 16.23** The correlation matrix showing the correlations between y and x_1 , y and x_2 , and x_1 and x_2 is shown next.

	y	x_1
x_1	0.924	
x_2	0.975	0.963

- a Does it appear that collinearity is a problem?
- b What are the likely consequences of collinearity?

- 16.24** Refer to Exercise 16.8. The correlation between ASSGNMNT and MID-SEM is 0.104.

- a What does this correlation tell you about the independent variables?
- b What does it say about the t -tests of β_1 and β_2 in the multiple regression model?

- 16.25** Refer to Exercise 16.10. The correlation between FATHER and MOTHER is 0.251.

- a What does this correlation tell you about the independent variables?
- b What does it say about the t -tests of β_1 and β_2 in the multiple regression model?

Computer applications

- 16.26 XR16-26** The observations of variables y , x_1 and x_2 are recorded in columns 1, 2 and 3 respectively.

- a Conduct a regression analysis of these data.
- b Calculate the residuals and standardised residuals. Identify any observations that should be checked.
- c Draw a histogram of the residuals. Is it likely that the normality requirement is violated?
- d Plot the residuals versus the predicted values of y . Is the variance of the error variable constant?
- e If heteroscedasticity exists, propose several possible remedies. Attempt each and report your findings.

- 16.27 XR16-27** The observations of variables y , x_1 and x_2 are recorded in columns 1, 2 and 3 respectively.

- a Conduct a regression analysis of these data.
- b Calculate the residuals and standardised residuals. Identify the observations that should be checked for accuracy.
- c Draw a histogram of the residuals. Is it likely that the normality requirement is violated?
- d Plot the residuals versus the predicted values of y . Is the variance of the error variable constant?

- e If heteroscedasticity exists, propose several possible remedies. Attempt each and report your findings.

- 16.28** Determine whether there are violations of the required conditions in the regression model used in Exercise 15.81. Which, if any, observations should be checked to ensure that they were correctly recorded?

- 16.29** Refer to Exercise 15.83. Conduct an analysis of the residuals to determine whether any of the required conditions are violated. Identify any observations that should be checked for accuracy.

- 16.30** Refer to Exercise 16.12.

- a Use whatever techniques you deem necessary to check the normality requirement and for heteroscedasticity.
- b Is multicollinearity a problem? Explain.
- c Identify all observations that should be checked.

- 16.31** Refer to Exercise 16.13.

- a Use whatever techniques you deem necessary to check the normality requirement and for heteroscedasticity.
- b Is multicollinearity a problem? Explain.

- 16.32 XR16-32** Supermarkets frequently price products such as bread and milk to attract customers to the store. The manager of a dairy that supplies milk to a supermarket wanted to know how sales of milk are affected by different prices. Consequently, she recorded the weekly sales of milk at one supermarket, the price of a litre of her company's brand (price A), and the price of a litre of her competitor's brand (price B). The data for the past 52 weeks are stored in columns 1 to 3 respectively.

- a Develop a regression model, and use a software package to produce the statistics.
- b Perform a complete diagnostic analysis to determine whether the required conditions are satisfied.
- c If one or more conditions are unsatisfied, attempt to remedy the problem.
- d Assess how well the model fits the data.
- e Interpret and test each of the coefficients.
- f Is multicollinearity a problem that affects your answer in part (e)? Explain.

16.4 Regression diagnostics – III (time series)

In Chapter 15, we pointed out that, in general, we check to see if the errors are independent when the data constitute a time series – data gathered sequentially over a series of time periods. In Section 15.7, we described the graphical procedure for determining whether the required condition that the errors are independent is violated. We plot the residuals versus the time periods and look for patterns. In this section, we augment that procedure with the *Durbin–Watson test*.

16.4a Durbin–Watson test

The Durbin–Watson test allows the statistics practitioner to determine whether there is evidence of first-order autocorrelation – a condition in which a relationship exists between consecutive residuals e_t and e_{t-1} of the form $e_t = \rho e_{t-1} + v_t$, $t = 2, 3, \dots, n$, where t is the time period and $\rho = \text{Corr}(e_t, e_{t-1})$. The Durbin–Watson statistic is defined as

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=2}^n e_t^2}$$

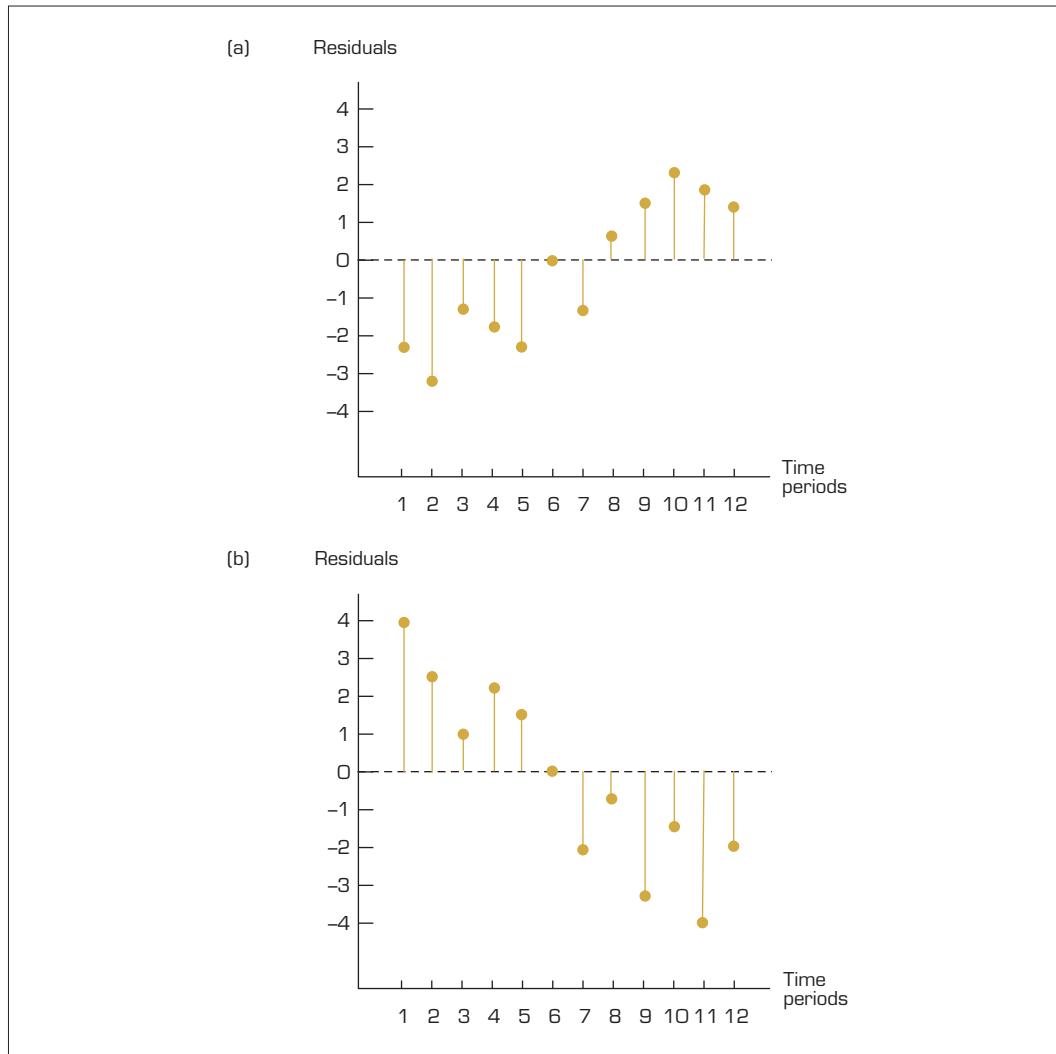
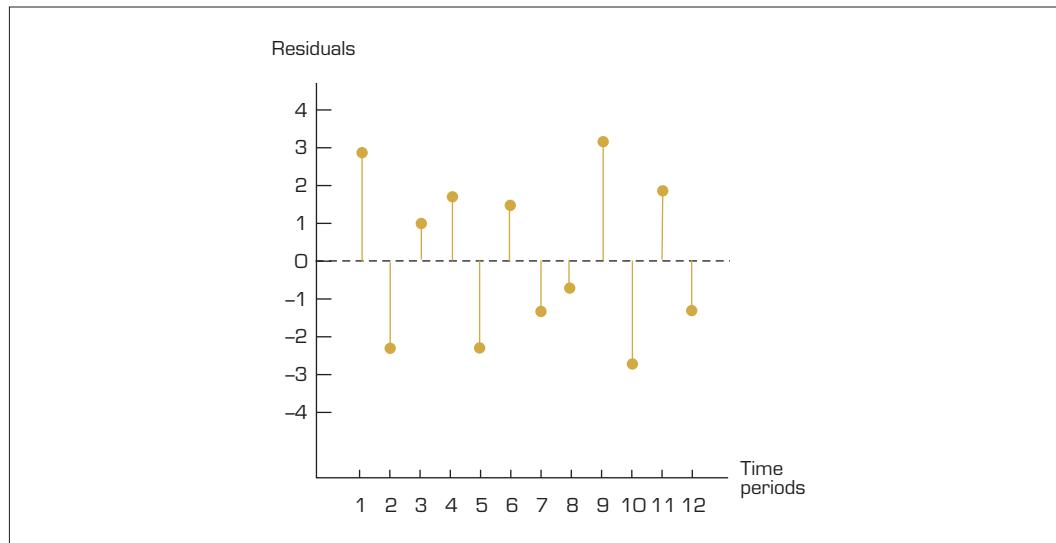
It can be shown that $d \approx 2(1 - \rho)$. As the range of the values of ρ is between -1 and $+1$, the range of the values of d is

$$0 \leq d \leq 4$$

where $d = 0$ (equivalent to $\rho = +1$) indicates perfect positive first-order autocorrelation; $d = 2$ (equivalent to $\rho = 0$) indicates no first-order autocorrelation; and $d = 4$ (equivalent to $\rho = -1$) indicates perfect negative first-order autocorrelation. Furthermore, values of d between 0 and 2 ($0 < d < 2$) indicate a positive first-order autocorrelation and values of d between 2 and 4 ($2 < d < 4$) imply a negative first-order autocorrelation. Positive first-order autocorrelation is a common occurrence in business and economic time series. It occurs when consecutive residuals tend to be similar. In that case, $(e_t - e_{t-1})^2$ will be small, producing a small value for d . Negative first-order autocorrelation occurs when consecutive residuals differ widely. For example, if positive and negative residuals generally alternate, $(e_t - e_{t-1})^2$ will be large and, as a result, d will be greater than 2 . **Figures 16.11(a) and 16.11(b)** depict positive autocorrelation, and **Figure 16.12** illustrates negative autocorrelation. Notice that in **Figure 16.11(a)**, as an absolute value, the first residual is a small number; the second residual, also a small number, is somewhat larger; and that trend continues. In **Figure 16.11(b)**, the first residual is large, and, in general, succeeding residuals decrease. In both figures, consecutive residuals are similar. In **Figure 16.12**, the first residual is a positive number, which is followed by a negative residual. The remaining residuals follow this pattern (with some exceptions). Consecutive residuals are quite different.

Tables 8(a) and (b) in Appendix B are designed to test for positive first-order autocorrelation, $H_0: d = 2$ (or $\rho = 0$) against $H_A: d < 2$ (or $\rho > 0$), by providing critical values of d_L and d_U for a variety of values of n and k and for $\alpha = 0.01$ and 0.05 .

The decision is made in the following way. If $d < d_L$, we conclude that there is enough evidence to show that positive first-order autocorrelation exists. If $d > d_U$, we conclude that there is not enough evidence to show that positive first-order autocorrelation exists. And if $d_L \leq d \leq d_U$, the test is inconclusive. The recommended course of action when the test is inconclusive is to continue testing with more data until a conclusive decision can be made.

FIGURE 16.11 Positive first-order autocorrelation**FIGURE 16.12** Negative first-order autocorrelation

For example, to test for positive first-order autocorrelation with $n = 20$, $k = 3$ and $\alpha = 0.05$, we test for the following hypotheses:

H_0 : There is no first-order autocorrelation (i.e. $d = 0$).

H_A : There is positive first-order autocorrelation (i.e. $d < 2$).

The decision is made as follows:

If $d < d_L = 1.00$, reject the null hypothesis in favour of the alternative hypothesis.

If $d > d_U = 1.68$, do not reject the null hypothesis.

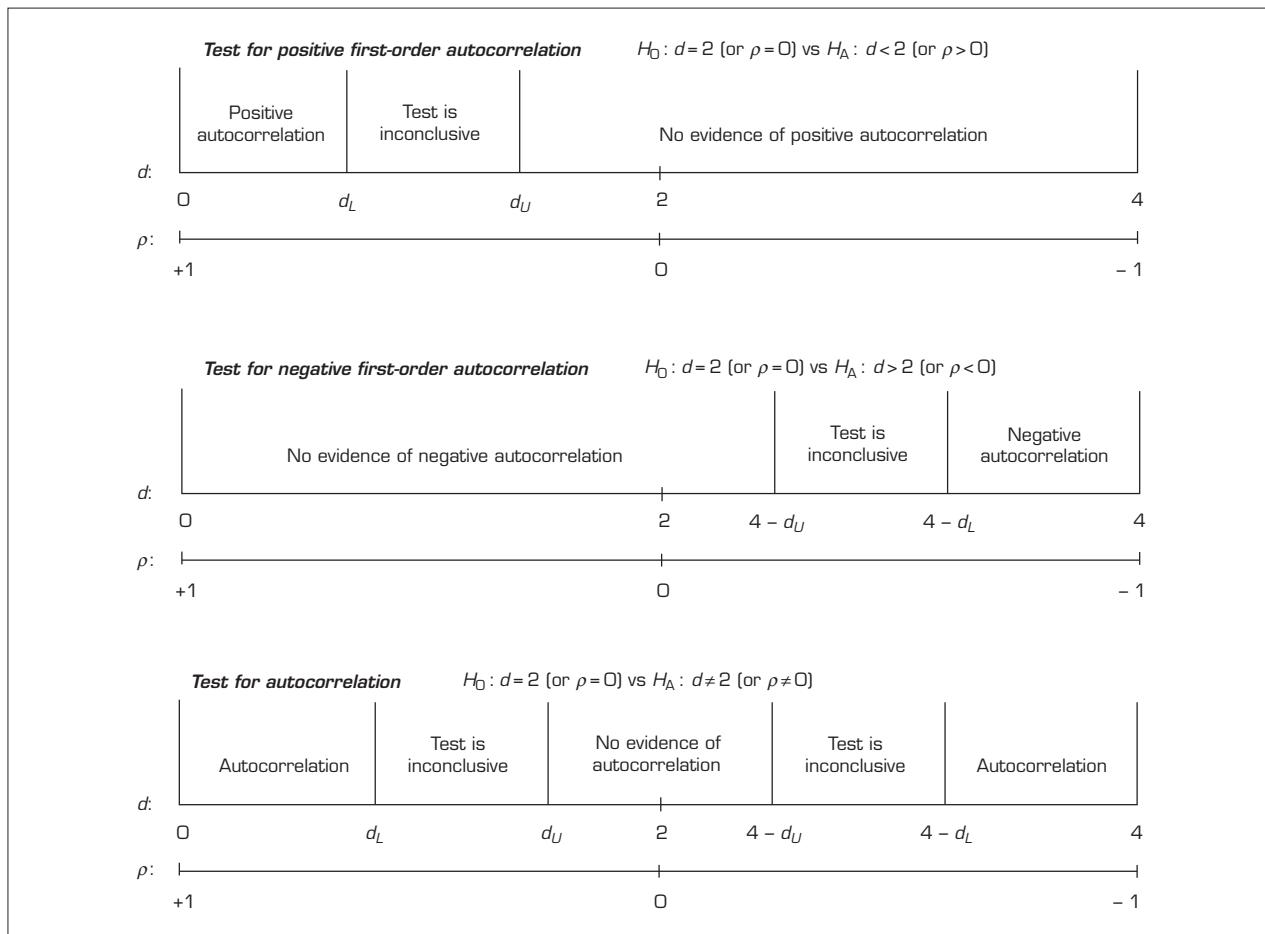
If $1.00 \leq d \leq 1.68$, the test is inconclusive.

To test for negative first-order autocorrelation, $H_0: d = 2$ (or $\rho = 0$) against $H_A: d > 2$ (or $\rho < 0$), we change the critical values. The appropriate critical values for this test are $4 - d_U$ and $4 - d_L$. If $d > 4 - d_L$, we conclude that negative first-order autocorrelation exists. If $d < 4 - d_U$, we conclude that there is not enough evidence to show that negative first-order autocorrelation exists. If $4 - d_U \leq d \leq 4 - d_L$, the test is inconclusive.

We can also test simply for first-order autocorrelation by combining the two one-tail tests. If $d < d_L$ or $d > 4 - d_L$, we conclude that autocorrelation exists. If $d_U \leq d \leq 4 - d_U$, we conclude that there is no evidence of autocorrelation. If $d_L \leq d < d_U$ or $4 - d_U \leq d \leq 4 - d_L$, the test is inconclusive. The significance level will be 2α (where α is the one-tail significance level).

Figure 16.13 describes the range of values of d and the conclusion for each interval.

FIGURE 16.13 Durbin–Watson test



For time-series data, we add the Durbin–Watson test to our list of regression diagnostics. That is, we determine whether the error variable is normally distributed with constant variance (as we did in Section 16.3), we identify outliers and (if our software allows it) influential observations that should be verified, and we conduct the Durbin–Watson test.

EXAMPLE 16.10

L01

Christmas week surfboard sales

XM16-10 Christmas week is a critical period for most surfboard sales and hire businesses. Because many students and adults are free from other obligations, they are able to spend several days indulging in their favourite pastime, surfing. A large proportion of gross revenue is earned during this period. A surfboard sales and hire business wanted to determine the effect that weather and other factors had on its sales. The manager of a Surfers Paradise (Gold Coast) surfboard sales and hire business recorded data on the number of surfboards sold during the Christmas week (y), the total number of tourist bus arrivals (x_1), and the average temperature in degrees centigrade (x_2) for the past 20 years. Develop a multiple regression model, and diagnose any violations of the required conditions.

Solution

The model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Excel output for Example 16.10

	A	B	C	D	E	F	G
1	Regression of variable Sales:						
2	Goodness of fit statistics [Sales]:						
3	Observations	20					
4	Sum of weights	20					
5	DF	17					
6	R ²	0.122					
7	Adjusted R ²	0.019					
8	MSE	292.853					
9	RMSE	17.113					
10	DW	0.609					
11							
12	Analysis of variance [Sales]:						
13	Source	df	SS	MS	F	Pr > F	
14	Model	2	694.05	347.03	1.185	0.330	
15	Error	17	4978.50	292.85			
16	Corrected Total	19	5672.55				
17							
18	Model parameters [Sales]:						
19	Source	Value	Standard Error	t	Pr > t	Lower bound (95%)	Upper bound (95%)
20	Intercept	112.251	66.287	1.693	0.109	-27.602	252.105
21	Arrivals	0.752	0.516	1.457	0.163	-0.336	1.839
22	Temperature	-0.913	1.970	-0.463	0.649	-5.069	3.244

The estimated model is

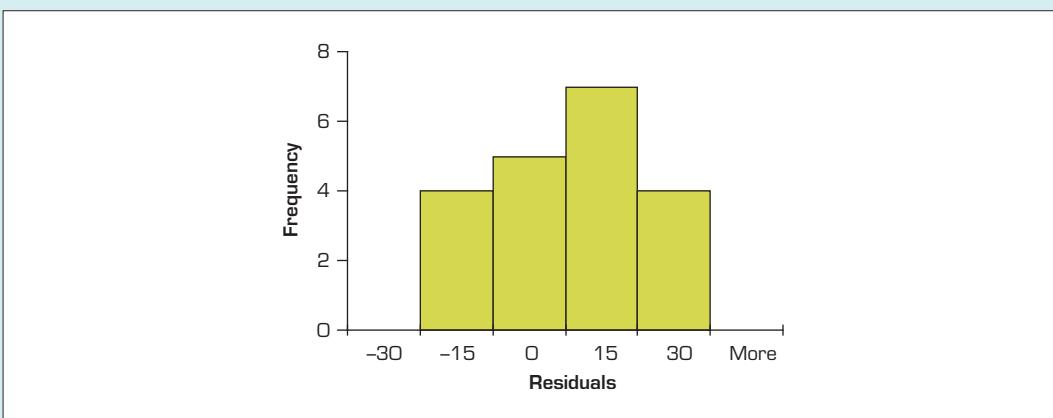
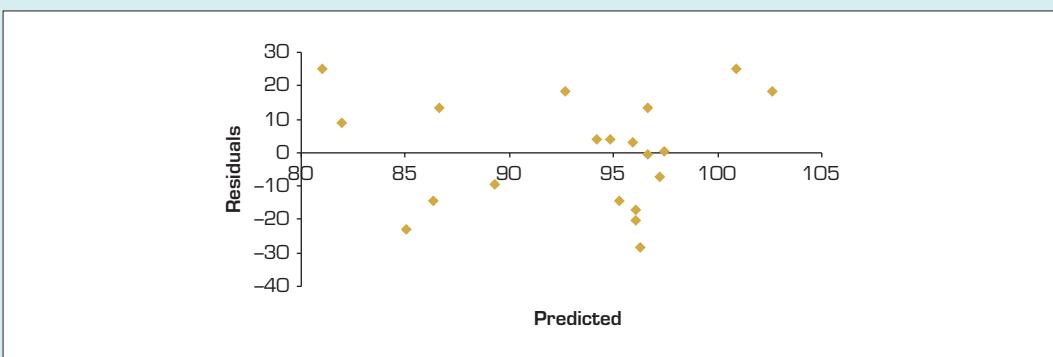
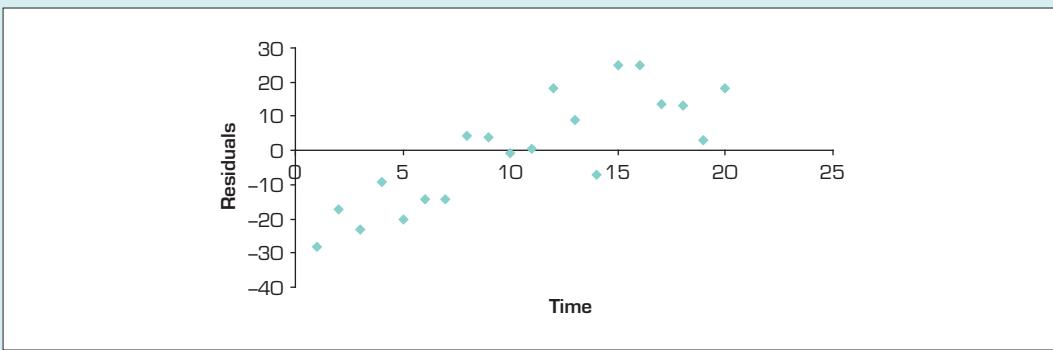
$$\hat{y} = 112.3 + 0.752x_1 - 0.913x_2$$

or

$$\text{Sales} = 112.3 + 0.752(\text{Arrivals}) - 0.913(\text{Temperature})$$

Interpreting the results

As you can see, the coefficient of determination is small (adjusted $R^2 = 0.02$) and the p -value of the F -test is 0.33, both of which indicate that the fitness and the utility of the model are poor. We use Excel to draw the histogram (**Figure 16.14**) of the residuals and plot the predicted values of y versus the residuals in **Figure 16.15**. Because the observations constitute a time series, we also use Excel to plot the residuals versus the time periods (years) in **Figure 16.16**. The histogram of the residuals in **Figure 16.14** reveals that the error variable may be normally distributed.

FIGURE 16.14 Histogram of residuals in Example 16.10**FIGURE 16.15** Plot of residuals versus predicted values of y in Example 16.10**FIGURE 16.16** Plot of residuals versus time periods in Example 16.10

The graph of the residuals versus the predicted values of y in **Figure 16.15** seems to indicate that the variance of the error variable is constant. That is, there does not appear to be any evidence of heteroscedasticity. This graph of residuals versus the time periods in **Figure 16.16** reveals a serious problem. There is a strong (positive) linear relationship between consecutive values of the residuals, which indicates that the requirement that the errors are independent has been violated. To confirm this diagnosis, we obtain the Durbin–Watson statistic from the XLSTAT regression output. [Excel Data Analysis does not directly provide the DW statistic.]

Using XLSTAT

	B	C
10	DW	0.609

The Durbin–Watson statistic will be in the goodness-of-fit statistics part of the regression output.



COMMANDS

Run the regression using data in XM16-10. Click **Outputs** and check **Predictions and residuals** and **Studentized residuals**.

The critical values are determined by noting that $n = 20$ and $k = 2$ (there are two independent variables in the model). If we wish to test for positive first-order autocorrelation with $\alpha = 0.05$, we find in Table 8(a) in Appendix B that

$$d_L = 1.10 \text{ and } d_U = 1.54$$

The null and alternative hypotheses are

H_0 : There is no first-order autocorrelation ($\rho = 0$ or $d = 2$).

H_A : There is positive first-order autocorrelation ($\rho > 0$ or $d < 2$).

The rejection region is $d < d_L = 1.10$. As $d = 0.61$, we reject the null hypothesis and conclude that there is enough evidence to infer that positive first-order autocorrelation exists.

Autocorrelation usually indicates that the model needs to include an independent variable that has a time-ordered effect on the dependent variable. The simplest such independent variable represents the time periods. To illustrate, we included a third independent variable, time period (x_3), which records the number of years since the year the data were gathered.

Thus, the time variable x_3 takes values $\{1, 2, \dots, 20\}$. Therefore, the extended new model can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

The estimation results are shown in the output below. The estimated sample regression equation is

$$\hat{y} = 90.49 + 0.708x_1 - 0.960x_2 + 2.286x_3$$

or

$$\text{Sales} = 90.49 + 0.708(\text{Arrivals}) - 0.960(\text{Temperature}) + 2.286(\text{Time})$$

Excel output for Example 16.10 (time variable included)

	B	C	D	E	F	G	H
1	<i>Regression of variable Sales:</i>						
2	Goodness of fit statistics [Sales]:						
3	Observations	20					
4	Sum of weights	20					
5	DF	16					
6	R ²	0.734					
7	Adjusted R ²	0.685					
8	MSE	94.146					
9	RMSE	9.703					
10	DW	1.885					
11							
12	<i>Analysis of variance [Sales]:</i>						
13	Source	df	SS	MS	F	Pr > F	
14	Model	3	4166.22	1388.74	14.75	< 0.0001	
15	Error	16	1506.33	94.15			
16	Corrected Total	19	5672.55				
17							
18	<i>Model parameters [Sales]:</i>						
19	Source	Value	Standard Error	t	Pr > t	Lower bound (95%)	Upper bound (95%)
20	Intercept	90.493	37.755	2.397	0.029	10.457	170.529
21	Arrivals	0.708	0.292	2.420	0.028	0.088	1.328
22	Temperature	-0.960	1.117	-0.860	0.403	-3.328	1.407
23	Time	2.286	0.376	6.073	< 0.0001	1.488	3.084





As we did before, we calculate the residuals and conduct regression diagnostics using Excel. The results are shown in **Figures 16.17–16.19**. The histogram (**Figure 16.17**) reveals that the error variable may be normally distributed. The graph of the residuals versus the predicted values of y (**Figure 16.18**) seems to indicate that the variance of the error variable is constant. That is, there does not appear to be any evidence of heteroscedasticity.

FIGURE 16.17 Histogram of residuals in Example 16.10 (time variable included)

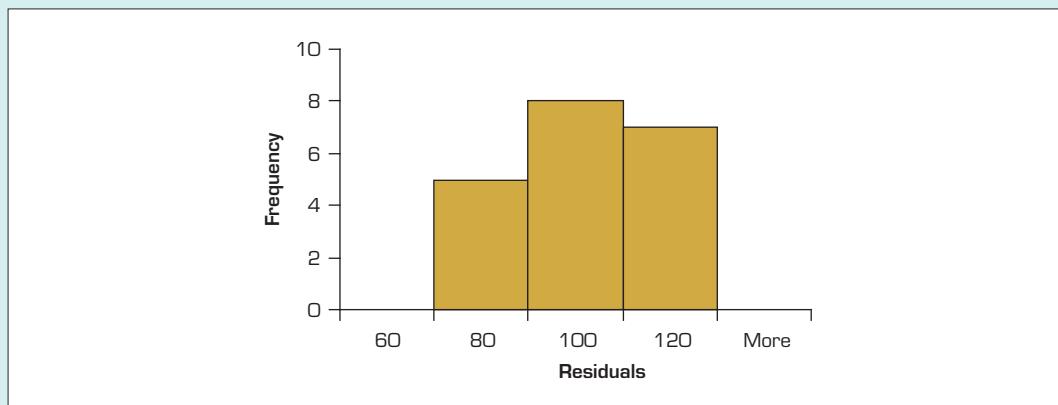


FIGURE 16.18 Plot of residuals versus predicted values of y in Example 16.10 (time variable included)

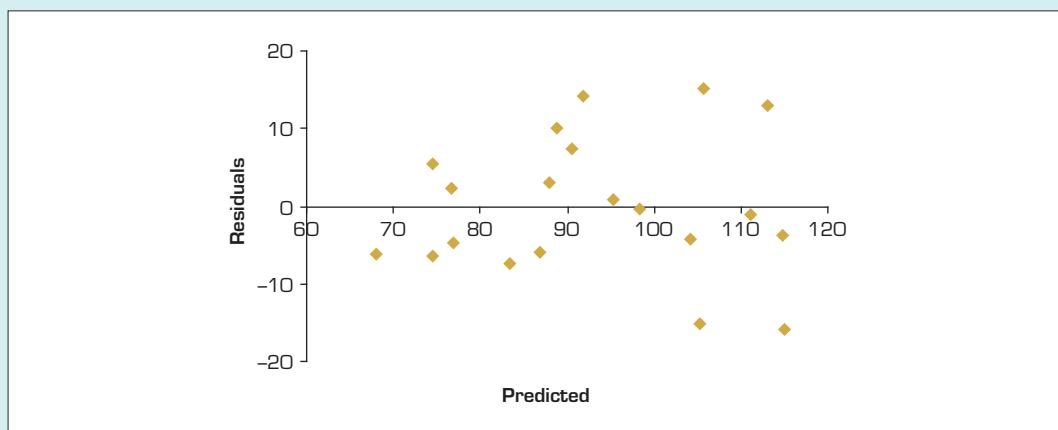
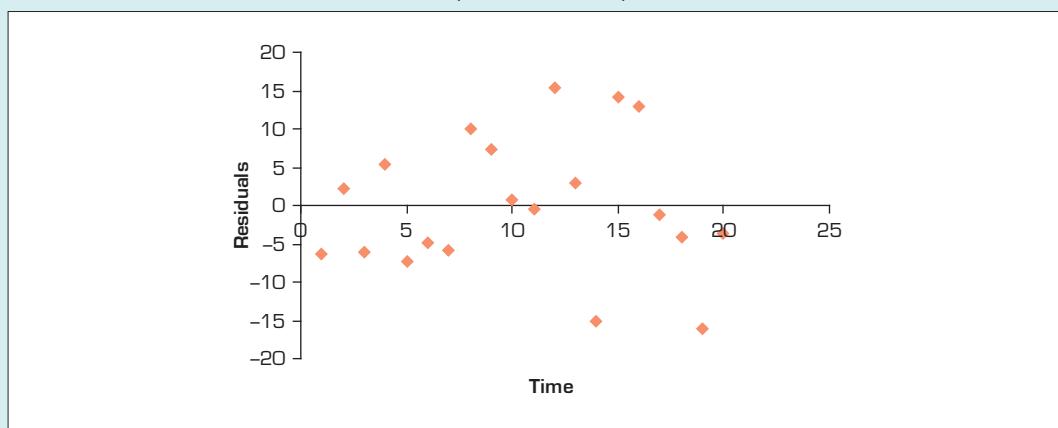


FIGURE 16.19 Plot of residuals versus time periods in Example 16.10 (time variable included)



The graph of residuals versus the time periods (**Figure 16.19**) reveals that there is no sign of autocorrelation. To confirm our diagnosis, we conducted the Durbin–Watson test.



	B	C
10	DW	1.885

From Table 8(a) in Appendix B, we find the critical values of the Durbin–Watson test. With $k = 3$ and $n = 20$, we find

$$d_L = 1.00 \text{ and } d_U = 1.68$$

As $d = 1.88 > 1.68 = d_U$, we conclude that there is not enough evidence to infer the presence of positive first-order autocorrelation. Notice that the model is improved dramatically. The F -test tells us that the model is valid. The t -tests tell us that both the number of tourist bus arrivals and time are significantly linearly related to the number of surfboard sales or hire. Higher temperature has a negative but insignificant impact on surfboard sales or hire.

16.4b Developing an understanding of statistical concepts

Notice that the addition of the variable TIME explained a large proportion of the variation in the number of surfboards sold. That is, the surfboard business experienced a relatively steady increase in sales over the past 20 years. Once this variable was included in the model, the variable ARRIVALS became significant because it was able to explain some of the remaining variation in surfboard sales. Without TIME, the variables ARRIVALS and TEMPERATURE were unable to explain a significant proportion of the variation in sales. The graph of the residuals versus the time periods and the Durbin–Watson test enabled us to identify the problem and correct it. In overcoming the autocorrelation problem, we improved the model so that we identified ARRIVALS as an important variable in determining surfboard sales. This result is quite common. Correcting a violation of a required condition will frequently improve the model.

EXERCISES

Learning the techniques

- 16.33** Given the following information, perform the Durbin–Watson test to determine whether first-order autocorrelation exists.

$$n = 25 \quad k = 5 \quad \alpha = 0.01 \quad d = 3.75$$

- 16.34** Test the following hypotheses with $\alpha = 0.05$.

H_0 : There is no first-order autocorrelation.

H_A : There is positive first-order autocorrelation.

$$n = 50 \quad k = 2 \quad d = 1.38$$

- 16.35** Test the following hypotheses with $\alpha = 0.01$.

H_0 : There is no first-order autocorrelation.

H_A : There is first-order autocorrelation.

$$n = 90 \quad k = 5 \quad d = 1.20$$

- 16.36** Test the following hypotheses with $\alpha = 0.05$.

H_0 : There is no first-order autocorrelation.

H_A : There is negative first-order autocorrelation.

$$n = 33 \quad k = 4 \quad d = 2.25$$

Applying the techniques

- 16.37** **XR16-37 Self-correcting exercise.** The values of \hat{y}_i and e_i are as follows:

Predicted value \hat{y}_i	Residual e_i
6.0795	-0.0795
9.6682	2.3318
10.5626	0.4374
7.0667	0.9333
11.8169	1.1831
7.7867	1.2133
7.3339	-1.3339
6.7067	1.2933
5.1852	-1.1852
6.7996	3.2004
9.3082	-1.3082
7.7867	1.2133
11.4569	-3.4569
12.1769	-0.1769
6.2652	-4.2652

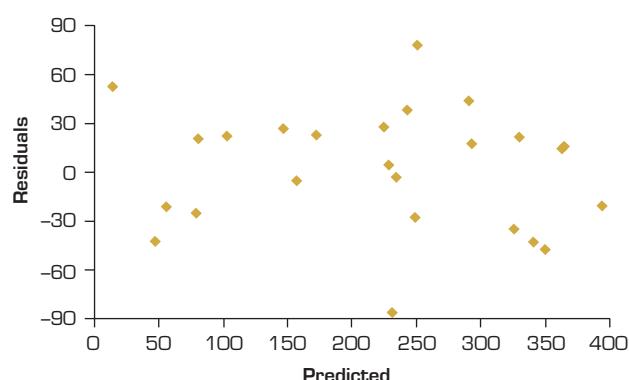
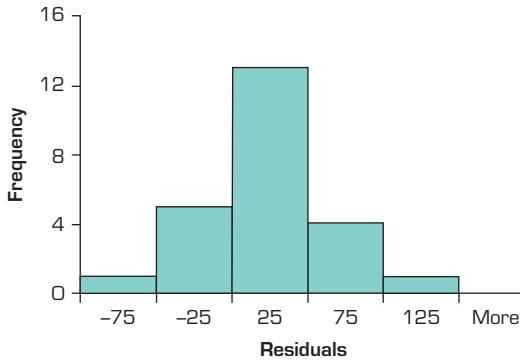
- a** Plot the residuals versus the predicted values. Does it appear that σ_e^2 is a constant?
- b** Plot the residuals versus time to examine the required condition that the errors be independent.
- c** The Durbin–Watson statistic is $d = 2.15$, $k = 2$ and $\alpha = 0.05$. What does this tell you about the possible existence of first-order autocorrelation?
- 16.38** Refer to Exercise 16.9. The Durbin–Watson statistic is $d = 1.76$ and $\alpha = 0.05$. What does this statistic tell you about the regression model?
- ### Computer applications
- 16.39 XR16-39** Observations of variables y , x_1 and x_2 were taken over 100 consecutive time periods. The data are recorded in the first three columns.
- a** Conduct a regression analysis of these data.
- b** Calculate the residuals and standardised residuals. Identify observations that should be checked.
- c** Draw the histogram of the residuals. Does it appear that the normality requirement is satisfied?
- d** Plot the residuals versus the predicted values of y . Is the error variance constant?
- e** Plot the residuals versus the time periods. Describe the graph.
- f** Perform the Durbin–Watson test. Is there evidence of autocorrelation?
- g** If autocorrelation was detected in part (f), propose an alternative regression model to remedy the problem. Use the computer to generate the statistics associated with this model.
- h** Redo parts (a) to (f) using the alternative model proposed in part (g). Compare the two models.
- 16.40 XR16-40** Weekly sales of a company's product (y) and those of its main competitor (x) were recorded for one year. These data are recorded in chronological order in columns 1 (company's sales) and 2 (competitor's sales).
- a** Conduct a regression analysis of these data.
- b** Calculate the residuals and standardised residuals. Identify observations that should be checked.
- c** Draw the histogram of the residuals. Does it appear that the normality requirement is satisfied?
- 16.41 XR16-41** Observations of variables y , x_1 , x_2 and x_3 were taken over 80 consecutive time periods. The data are stored in the first four columns.
- a** Conduct a regression analysis of these data.
- b** Calculate the residuals and standardised residuals. Identify observations that should be checked.
- c** Draw the histogram of the residuals. Does it appear that the normality requirement is satisfied?
- d** Plot the residuals versus the predicted values of y . Is the error variance constant?
- e** Plot the residuals versus the time periods. Does there appear to be autocorrelation?
- f** Perform the Durbin–Watson test. Is there evidence of autocorrelation?
- g** If autocorrelation was detected in part (f), propose an alternative regression model to remedy the problem. Use the computer to generate the statistics associated with this model.
- h** Redo parts (a) to (f) using the alternative model proposed in part (g). Compare the two models.
- 16.42** Refer to Exercise 15.76. Perform a complete regression diagnostic analysis of the simple regression model used in that exercise. That is, determine whether the error variable is normal with constant variance and whether the errors are independent. Identify any observations that should be checked for accuracy.

- 16.43** Refer to Exercise 16.9. The correlations for each pair of independent variables are shown below.

Pair of independent variables	Correlation
PERMITS and MORTGAGE	0.005
PERMITS and A-VACNY	-0.150
PERMITS and O-VACNY	-0.103
MORTGAGE and A-VACNY	-0.040
MORTGAGE and O-VACNY	-0.033
A-VACNY and O-VACNY	0.065

- a What do these correlations tell you about the independent variables?
- b Is it likely that the *t*-tests of the coefficients are meaningful? Explain.

- 16.44** Refer to Exercise 16.9. The histogram of the residuals and the graph of the residuals and the predicted values are shown below.



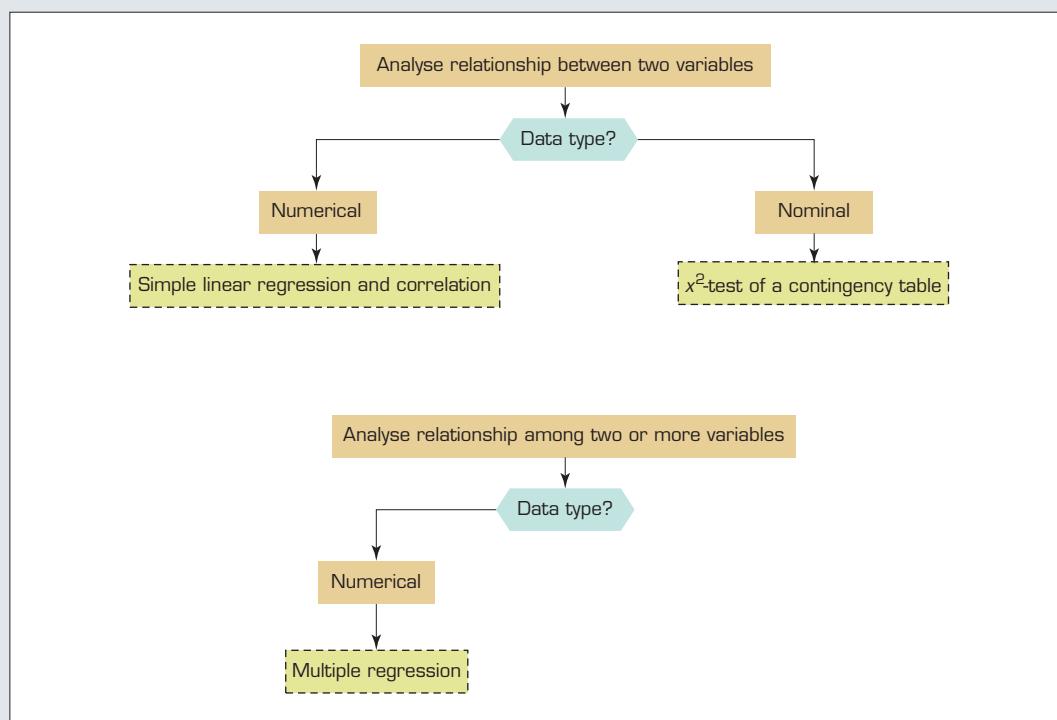
- a Does it appear that the normality requirement is violated? Explain.
- b Is the error variable variance constant? Explain.

Study Tools

CHAPTER SUMMARY

The *multiple regression model* extends the simple linear regression model introduced in Chapter 15. The statistical concepts and techniques are similar to those presented in simple linear regression. We assess the model in three ways: *standard error of estimate*, the *coefficient of determination* (and the coefficient of determination adjusted for degrees of freedom), and the *F-test* of the *analysis of variance*. We can use the *t*-tests of the coefficients to determine whether each of the independent variables is linearly related to the dependent variable. As we did in Chapter 15, we showed how to diagnose violations of the required conditions and to identify other problems. Transformations were shown to be the best way of dealing with *non-normality* and *heteroscedasticity*. We introduced *multicollinearity* and demonstrated its effect and its remedy. Finally, we presented the *Durbin–Watson test* to detect *first-order autocorrelation*.

The flowchart below is designed to help students identify the correct statistical technique to use.



The data files for Examples, Exercises and Cases are provided on the companion website (accessible through <https://login.cengagebrain.com/>).

SYMBOLS

Symbol	Pronounced	Represents
β_i	Beta-sub- <i>i</i> or beta- <i>i</i>	Coefficient of <i>i</i> th independent variable
$\hat{\beta}_i$	Beta-hat-sub- <i>i</i> or beta-hat- <i>i</i>	Sample estimate of coefficient β_i

SUMMARY OF FORMULAS

Standard error of estimate	$s_e = \sqrt{\frac{SSE}{n - k - 1}}$
Test statistic for β_i	$t = \frac{\beta_i - \hat{\beta}_i}{s_{\beta_i}}$
Coefficient of determination	$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{SSE}{(n-1)s_y^2}$
Adjusted coefficient of determination	Adjusted $R^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$
Mean square for error	$MSE = \sqrt{\frac{SSE}{n - k - 1}}$
Mean square for regression	$MSR = \frac{SSR}{k}$
F statistic	$F = \frac{MSR}{MSE}$
Durbin-Watson statistic	$d = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}$

SUPPLEMENTARY EXERCISES

16.45 XR16-45 An auctioneer of antique and semi-antique Persian rugs kept records of his weekly auctions in order to determine the relationships among price, age of rug, number of people attending the auction, and number of times the winning bidder had previously attended his auctions. He felt that with this information, he could plan his auctions better, serve his steady customers better and make a higher profit overall for himself. Part of the data is shown in the table.

Use the Excel package to estimate the model,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Price (\$) <i>y</i>	Age of rug <i>x</i> ₁	Audience size <i>x</i> ₂	Previous attendance <i>x</i> ₃
1080	80	40	1
2540	150	80	12
.	.	.	.
.	.	.	.
2140	115	95	5

- a Do the signs of the coefficients conform to what you expected?
- b Do the results allow us to conclude at the 5% significance level that price is linearly related to each of age, audience size and previous attendance?
- c What proportion of the variation in *y* is explained by the independent variables?
- d Test the utility of the overall regression model, with $\alpha = 0.05$.
- e What price would you forecast for a 100-year-old rug, given an audience size of 120 that had, on average, attended three of the auctioneer's auctions before?
- f The correlation matrix is given below:

	<i>y</i>	<i>x</i> ₁	<i>x</i> ₂
<i>x</i> ₁	0.920		
<i>x</i> ₂	0.890	0.803	
<i>x</i> ₃	0.382	0.326	0.206

What (if anything) do the correlations tell you about your original answers in parts (a) to (e)?

- g** Check to determine whether the error variable is normally distributed.
- h** Check to determine whether σ_e^2 is a constant.
- i** Check whether the errors are independent.
- j** How do your answers in parts (g) to (i) affect your answers in parts (a) to (e)?

16.46 XR16-46 The managing director of a real estate company wanted to know why certain branches of the company outperformed others. He felt that the key factors in determining total annual sales (in \$millions), y , were the advertising budget (in \$000s), x_1 , and the number of sales agents, x_2 . To analyse the situation, he took a sample of 15 offices and ran the data through a statistical software system. Part of the output are shown below.

The regression equation is

$$\hat{y} = -19.5 + 0.158x_1 + 0.962x_2$$

Variable	Coefficient	Standard error	t-ratio
Constant	-19.47	15.84	-1.23
x_1	0.158	0.056	2.82
x_2	0.962	0.778	1.24
$s_e = 7.362$	$R^2 = 52.4\%$	Adjusted $R^2 = 44.5\%$	

Analysis of variance			
Source	DF	SS	MS
Regression	2	716.58	358.29
Residual	12	650.35	54.20
Total	14	1366.93	

- a** Interpret the coefficients.
- b** Test to determine whether there is a linear relationship between each independent variable and the dependent variable, with $\alpha = 0.05$.
- c** Test the overall utility of the model.
- d** Interpret the value of R^2 .

The predicted values and the residuals are as follows:

Office	Predicted value \hat{y}_i	Residuals e_i
1	34.41	-2.41
2	44.10	2.90
3	22.99	-4.99
4	27.77	-2.77
5	49.84	-0.84
6	39.06	1.94
7	36.82	15.18
8	32.18	5.82
9	38.66	-2.66
10	25.22	3.78
11	40.02	2.98
12	31.12	-3.12
13	40.33	-16.33
14	39.49	-3.49
15	36.99	4.01

The correlation matrix is

	y	x_1
x_1	0.681	
x_2	0.457	0.329

- e** Is collinearity a problem?
- f** Does it appear that the error variable is normal?
- g** Does it appear that σ_e^2 is a constant?
- h** Does it appear that the errors are independent?
- i** Do any of your answers to parts (e) to (h) cause you to doubt your answers to parts (a) to (d)?

Suppose that a third independent variable was included – the average number of years of experience in the real estate business x_3 for each office. These data are as follows:

Office	Average years of experience x_3	Office	Average years of experience x_3
1	12	9	12
2	15	10	8
3	8	11	17
4	12	12	9
5	16	13	11
6	14	14	10
7	13	15	13
8	10		

The computer output is shown below.

The regression equation is

$$\hat{y} = -8.2 + 0.0905x_1 - 0.071x_2 + 1.93x_3$$

Variable	Coefficient	Standard error	t-ratio
Constant	-8.17	16.21	-0.50
x_1	0.09054	0.06601	1.37
x_2	-0.0714	0.9500	-0.08
x_3	1.927	1.145	1.68
$s_e = 6.857$	$R^2 = 62.2\%$	Adjusted $R^2 = 51.8\%$	

Analysis of variance			
Source	DF	SS	MS
Regression	3	849.78	283.26
Error	11	517.16	47.01
Total	14	1366.93	

- j What differences do you observe between this output and the original output? How do you account for any differences?

- k The correlation matrix is

	y	x_1	x_2
x_1	0.681		
x_2	0.457	0.329	
x_3	0.743	0.647	0.679

Does this explain some of the differences? Why or why not?

16.47 XR16-47 The agronomist referred to in Exercise 15.85 believed that the amount of rainfall as well as the amount of fertiliser used would affect the crop yield. She redid the experiment in the following way. Thirty greenhouses were rented. In each, the amount of fertiliser and the amount of water were varied. At the end of the growing season, the amount of corn was recorded (column 1 = crop yield in kilograms; column 2 = amount of fertiliser applied in kilograms; column 3 = amount of water in litres per week).

- a Determine the sample regression line and interpret the coefficients.
- b Do these data allow us to infer at the 5% significance level that there is a linear relationship between the amount of fertiliser and the crop yield?
- c Do these data allow us to infer at the 5% significance level that there is a linear relationship between the amount of water and the crop yield?

- d What can you say about the multiple regression fitness of the model?
- e Predict the crop yield when 300 kg of fertiliser and 1000 litres of water are applied.
- f Are the required conditions satisfied? Explain.
- g Is multicollinearity a problem in this model? Explain.

16.48 XR16-48 The Director of the Department of Education in Queensland was analysing the average mathematics test scores in the schools under his control. He noticed that there were dramatic differences in scores among the schools. In an attempt to improve the scores of all the schools, he attempted to determine the factors that account for the differences. Accordingly, he took a random sample of 40 schools across the state and, for each, determined the mean test score last year, the percentage of teachers in each school who have at least one university degree in mathematics, the mean age, and the mean annual income of the mathematics teachers. These data are recorded in columns 1 to 4 respectively.

- a Conduct a regression analysis to develop the equation.
- b Is the model useful in explaining the variation among schools? Explain.
- c Are the required conditions satisfied? Explain.
- d Is multicollinearity a problem? Explain.
- e Interpret and test the coefficients (with $\alpha = 0.05$).
- f Predict the test score at a school where 50% of the mathematics teachers have mathematics degrees, their mean age is 43 and their mean annual income is \$78500.

16.49 XR16-49 Life insurance companies are keenly interested in predicting how long their customers will live, because their premiums and profitability depend on such numbers. An actuary for one insurance company gathered data from 100 recently deceased male customers. He recorded the age at death of the customer plus the ages at death of his mother and father, the mean ages at death of his grandmothers and the mean ages at death of his grandfathers. These data are recorded in columns 1 to 5 respectively.

- a Perform a multiple regression analysis on these data.
- b Is the model likely to be useful in predicting men's longevity?

- c Are the required conditions satisfied?
- d Is multicollinearity a problem here?
- e Interpret and test the coefficients.
- f Predict the longevity of a man whose parents lived to the age of 70, whose grandmothers averaged 80 years and whose grandfathers averaged 75.
- g Estimate the mean longevity of men whose mothers lived to 75, whose fathers lived to 65, whose grandmothers averaged 85 years and whose grandfathers averaged 75.

16.50 XR16-50 University students often complain that universities reward lecturers for research but not for teaching, and argue that lecturers react to this situation by devoting more time and energy to the publication of their findings and less time and energy to classroom activities. Lecturers counter that research and teaching go hand in hand; more research makes better teachers. A student organisation at one university decided to investigate the issue. They randomly selected 50 lecturers who are employed by a multicampus university. The students recorded the salaries of the lecturers, their average teaching evaluations (on a 10-point scale) and the total number of journal articles published in their careers. These data are recorded in columns 1 to 3 respectively. Perform a complete analysis (produce the regression equation, assess it and diagnose it) and report your findings.

16.51 XR16-51 One of the critical factors that determine the success of a catalogue store chain is the availability of products that consumers want to buy. If a product is sold out, future sales to that customer are less likely. Because of this, stores are regularly resupplied by delivery trucks operating from a central warehouse. In an analysis of the chain's operations, the general manager wanted to determine the factors that affected how long it took to unload delivery trucks. A random sample of 50 deliveries to one store was observed. The times (in minutes) to unload the truck, the total number of boxes and the total weight (in hundreds of kilograms) of the boxes were recorded.

- a Determine the multiple regression equation.
- b How well does the model fit the data? Explain.
- c Are the required conditions satisfied?
- d Is multicollinearity a problem?

- e Interpret and test the coefficients. What does this analysis tell you?
- f Produce a prediction for the amount of time needed to unload a truck with 100 boxes of total weight 5000 kg.
- g Produce an estimate of the average amount of time needed to unload trucks with 100 boxes of total weight 5000 kg.

16.52 XR16-52 The owner of a drive-in theatre is very concerned about the ongoing operation of the drive-in, due to competition from his competitors – the in-house cinemas opening up in most major shopping complexes. The owner thinks that the number of people coming to his drive-in is mostly dependent on the number of times he has advertised discount tickets in the local newspaper and the number of times his competitors advertised discount tickets in the local newspaper. In order to analyse this belief, the owner asked his son, a third-year business student, to analyse the data he has collected and recorded for the three variables: the number of tickets sold at his drive-in (TICKETS); the number of times he has advertised discount tickets in the local newspaper (OWNADV); and the number of times the competitors have advertised discount tickets in the local newspaper (COMPADV) for the past 20 years. Develop a multiple regression model, and diagnose any violations of the required conditions.

16.53 XR16-53 In Exercise 15.47, an economist examined the relationship between office rents and the city's office vacancy rate. The model appears to be quite poor. It was decided to add another variable that measures the state of the economy. The city's unemployment rate was chosen for this purpose.

- a Determine the regression equation.
- b Determine the coefficient of determination and describe what this value means.
- c Test the model's validity in explaining office rent.
- d Determine which of the two independent variables is linearly related to rents.
- e Determine whether the error is normally distributed with a constant variance.
- f Determine whether there is evidence of autocorrelation.
- g Predict with 95% confidence the office rent in a city whose vacancy rate is 10% and whose unemployment rate is 7%.

Case Studies

CASE 16.1 Are lotteries a tax on the poor and uneducated?

C16-01 Lotteries have become important sources of revenue for governments. Many people have criticised lotteries, however, referring to them as a tax on the poor and uneducated. In an examination of the issue, a random sample of 100 adults was asked how much they spend on lottery tickets. Each was also interviewed about various socioeconomic variables. The purpose of this study was to test the following beliefs.

- 1 Relatively uneducated people spend more on lotteries than do relatively educated people.
- 2 Older people buy more lottery tickets than do younger people.
- 3 People with more children spend more on lotteries than do people with fewer children.
- 4 Relatively poor people spend a greater proportion of their income on lotteries than do relatively rich people.

The following data were stored in columns 1 to 5 respectively:

- Amount spent on lottery tickets as a percentage of total household income
- Number of years of education
- Age
- Number of children
- Personal income (in thousands of dollars)

What conclusions can you draw?

CASE 16.2 Demand for beer in Australia

C16-02 For many reasons, the consumption of alcohol continues to attract a lot of attention in the media. On the one hand, the alcohol industry is always interested to know how the market share of its product changes and is looking for ways to increase the demand for alcohol. On the other hand, the government, the health profession and social workers are searching for ways to reduce alcohol consumption by introducing control mechanisms such as increasing alcohol taxes, banning alcohol advertising and using random breath testing.

Data are presented for the per capita consumption of beer, the prices of beer, wine and spirits, and the per capita income for Australia during the period 1962–2017. Using all the variables in logarithmic form, analyse the demand for beer in Australia.

CASE 16.3 Book sales vs free examination copies revisited

C16-03 After performing the simple regression analysis (Case 15.5), the publishing editor was very impressed with the results. The coefficient of determination was reasonably good, indicating a linear relationship. This suggested that the number of free copies is an indicator of sales revenues. However, he decided to improve the model by including additional variables. He decided to include the total sales revenues for the 20 books in the previous year. These data are stored in the following way.

Column 1: Sales revenues from textbooks this year

Column 2: Number of free copies

Column 3: Sales revenues from textbooks last year

Include the additional variable in the model, and discuss your findings.

CASE 16.4 Average hourly earnings in New Zealand

C16-04 A leader of the Workers Union in New Zealand would like to study the movement in the average hourly earnings of New Zealand workers. He collected and recorded data on average earnings, labour cost and rate of inflation. Set up a suitable regression model to investigate the impact of labour cost and inflation on the hourly earnings of an average New Zealand worker.

CASE 16.5 Testing a more effective device to keep arteries open

C16-05 A stent is a metal mesh cylinder that holds a coronary artery open after a blockage has been removed. However, in many patients, the stents, which are made of bare metal, become blocked as well. One cause of the recurrence of blockages is the body's rejection of the foreign object. In a study published in the *New England Journal of Medicine* (January 2004), a new stent was tested. The new stents are polymer based and, after insertion, slowly release a drug (paclitaxel) to prevent rejection. A sample of 1314 patients who were receiving a stent in a single, previously untreated coronary artery blockage was recruited. A total of 652 were randomly assigned to receive a bare-metal stent, and 662 to receive an identical-looking, polymer drug-releasing stent. The results were recorded in the following way:

Column 1: Patient identification number

Column 2: Stent type (1 = bare metal, 2 = polymer-based)

Column 3: Reference-vessel diameter (the diameter of the artery that is blocked, in millimetres)

Column 4: Lesion length (the length of the blockage, in millimetres)

Reference-vessel diameters and lesion lengths were measured before the stents were inserted. The following outcomes were recorded 12 months after the stents had been inserted:

Column 5: Inadequate blood flow (2 = yes; 1 = no)

Column 6: Blockage that needed to be reopened (2 = yes; 1 = no)

Column 7: Death from cardiac causes (2 = yes; 1 = no)

Column 8: Death from blockage (2 = yes; 1 = no)

- a Test to determine whether the two groups of patients were not different before the stents were inserted.
- b Determine whether there is enough evidence to infer that the polymer-based stent is superior to the bare-metal stent.
- c As a laboratory researcher in the pharmaceutical company, write a report that describes this experiment and the results.

Appendix 16.A

F-distribution

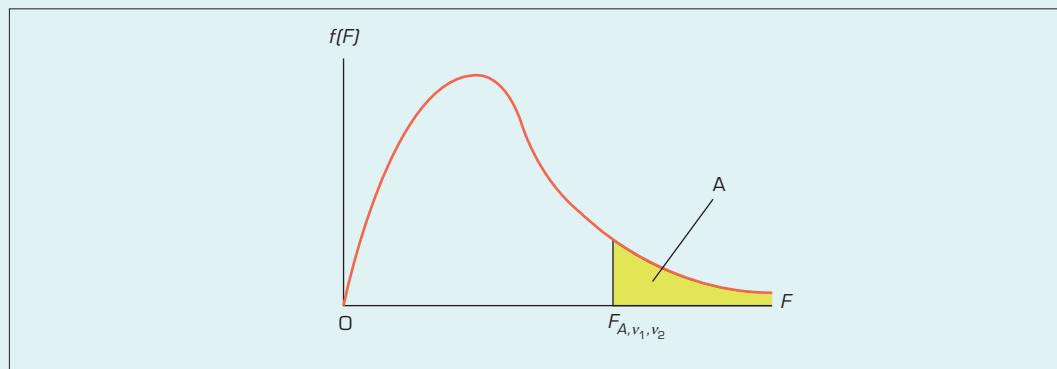
Statisticians have shown that the ratio of two independent chi-squared variables divided by their degree of freedom follows an **F-distribution**. The degrees of freedom of the F distribution are identical to the degrees of freedom of the two chi-squared distributions. Variables that are F distributed range from 0 to ∞ . The approximate shape of the distribution is depicted in **Figure A16.1**.

A16.1. The exact shape is determined by two numbers of degrees of freedom. Because the statistic is a ratio, one number of degrees of freedom is labelled as the *numerator degrees of freedom*, denoted v_1 (Greek letter *nu*), and the other is labelled as the *denominator degrees of freedom*, denoted v_2 .

F-distribution

A continuous distribution used in statistical inference.

FIGURE A16.1 F-distribution



16.Aa Determining values of F manually

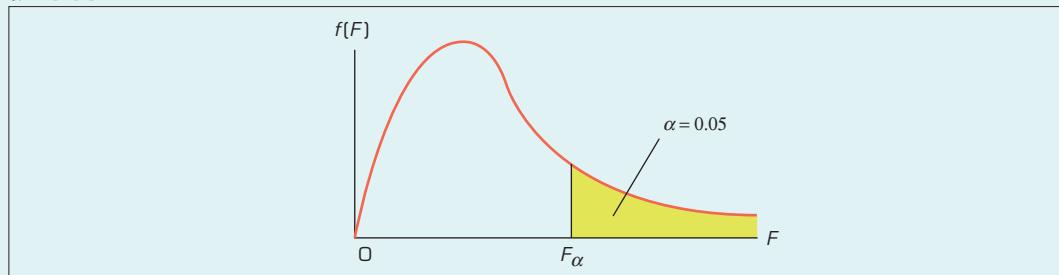
Tables 6(a)–6(d) in Appendix B provide the critical values for the F-distribution. It lists values of F_{A,v_1,v_2} , where F_{A,v_1,v_2} is the value of F with v_1 and v_2 degrees of freedom such that the area to its right under the F -distribution is A . That is

$$P(F > F_{A,v_1,v_2}) = A$$

Part of Table 6(a) in Appendix B (for $A = 0.05$) is reproduced here as **Table A16.1**. To determine any critical value, find the numerator degrees of freedom v_1 across the top row and the denominator degrees of freedom v_2 down the first column. The intersection of that row and that column shows the critical value. To illustrate, suppose that we want to find $F_{0.05,5,7}$.

Tables 6(a)–6(d) in Appendix B provide the critical values for four values of A : 0.05, 0.025, 0.01 and 0.005. (**Table A16.1** lists some of the critical values for $A = 0.05$.) The numerator number of degrees of freedom is 5, which we find across the top row, and the denominator number of degrees of freedom is 7, which we locate in the first column. The intersection is 3.97. Thus, $F_{0.05,5,7} = 3.97$. (See **Table A16.1**.)

TABLE A16.1 Reproduction of part of Table 6(a) in Appendix B: Critical values of the F -distribution
 $\alpha = 0.05$



v_1	Numerator degrees of freedom								
	1	2	3	4	5	6	7	8	9
v_2	161.40	199.50	215.70	224.60	230.20	234.00	236.80	238.90	240.50
1	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
2	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
3	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
4	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
5	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
6	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
7	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
8	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
9	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02

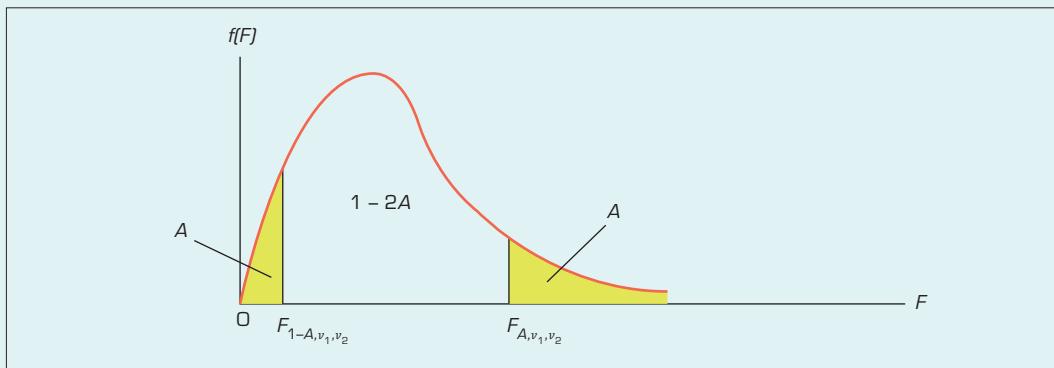
Note that the order in which the degrees of freedom appear is important. To find $F_{0.05,7,5}$ (numerator degrees of freedom = 7 and denominator degrees of freedom = 5) we locate 7 across the top row and 5 down the first column. The number in the intersection is $F_{0.05,7,5} = 4.88$.

Tables 6(a)–6(d) of Appendix B provide only values of F_A , which are the values of F that are in the right tail of the F -distribution. The left-tail values, which we label F_{1-A} (as we did to denote the left-tail values of the chi-squared distribution), are not listed. **Figure A16.2** illustrates this notation. The left-hand tail values are not provided because we can easily determine the values of F_{1-A} from the values of F_A . Mathematicians have derived the following formula.

$$F_{1-A, v_1, v_2} = \frac{1}{F_{A, v_2, v_1}}$$

For example,

$$F_{0.95,4,8} = \frac{1}{F_{0.05,8,4}} = \frac{1}{6.04} = 0.166$$

FIGURE A16.2 F_A and F_{1-A} 

16.Ab Determining F probabilities and F values using the computer

To calculate the probability to the right of any F value, proceed as follows:

COMMANDS

- 1 Click **FORMULAS, fx, All** from the categories dropdown menu and select the **FDIST** function.
Click **OK**.
- 2 Type the value of $x(X)$, numerator degrees of freedom (**Deg_freedom1**) and denominator degrees of freedom (**Deg_freedom2**).
Alternatively, type
= FDIST([X], [Numerator degrees of freedom], [Denominator degrees of freedom])

For example, $\text{FDIST}(3.97, 5, 7) = 0.05$.

To determine a value of an F random variable, follow these instructions:

COMMANDS

- 1 Click **FORMULAS, fx, All** from the categories dropdown menu and select the **FINV** function.
Click **OK**.
- 2 Type the probability to the right of the value (**Probability**) the numerator degrees of freedom (**Deg_freedom1**) and the denominator degrees of freedom (**Deg_freedom2**).
Alternatively, type
= FINV([Probability], [Numerator degrees of freedom], [Denominator degrees of freedom])

For example, $\text{FINV}(0.05, 5, 7) = 3.97$.

PART THREE

Applications

CHAPTER 17 Time series analysis and forecasting

CHAPTER 18 Index numbers

Part 1 of this book dealt with the foundation for statistical inference, which was developed further in Part 2. In this part of the book we discuss several techniques that in some ways are different from the methods presented earlier. The development of the techniques of statistical inference followed a pattern first seen in Chapter 10. In the parametric methods (Chapters 10–16) we started by identifying a parameter of interest, that parameter's best estimator and the estimator's sampling distribution. The sampling distribution was then used to develop the confidence interval estimator and test statistic.

In the two chapters of Part 3, that pattern is not followed. Chapter 17 discusses forecasting time series, which shares the goals of Chapters 15–16, but the technique and circumstances are different. In Chapter 18, we present index numbers, which are used to measure how variables change over time. Index numbers are used primarily as descriptive measurements with little or no inference.

Time series analysis and forecasting

Learning objectives

This chapter deals with the basic components of a time series, time series decomposition and simple forecasting.

At the completion of this chapter, you should be able to:

- L01** identify the four possible components of a time series
- L02** use the smoothing technique to remove the random variation and identify the remaining components
- L03** use the linear, logarithmic and polynomial regression models to analyse the trend
- L04** measure the cyclical effect using the percentage of trend method
- L05** measure the seasonal effect by computing the seasonal indexes
- L06** calculate MAD and SSFE to determine which forecasting model works best
- L07** use exponential smoothing to forecast a time series
- L08** use regression models to forecast a time series.

CHAPTER OUTLINE

- Introduction
- 17.1 Components of a time series**
- 17.2 Smoothing techniques**
- 17.3 Trend analysis**
- 17.4 Measuring the cyclical effect**
- 17.5 Measuring the seasonal effect**
- 17.6 Introduction to forecasting**
- 17.7 Time series forecasting with exponential smoothing**
- 17.8 Time series forecasting with regression**

SPOTLIGHT ON STATISTICS

Retail turnover of Australian food services sector

Cafes, restaurants and the take-away food service play a major role in the Australian economy. This service sector employs a significant number of full-time and part-time employees. Many people depend on the food services sector for their daily meal needs. Using Australian food services retail turnover data for the period January 2008 to December 2018, the director of a small-business government ministry would like to model the time series, forecast the retail turnover of the food service sector for the 12 months of 2019 and assess the quality of the forecasts using the corresponding actual data for 2019. The data are stored in file **CH17\XM17-00**.

Source: Australian Bureau of Statistics, ABS cat. no. 8501.0, *Retail Trade Australia*, December 2019, Canberra, Australia.

For the solution, please see pages 787–9.



Source: Shutterstock.com/asar nasib

Introduction

Any variable that is measured over time in sequential order is called a **time series**. Our objective in this chapter is to analyse time series in order to detect patterns that will enable us to forecast the future value of the time series. There is an almost unlimited number of such applications in management and economics. Some examples are listed.

- 1 Governments want to know future values of interest rates, unemployment rates and percentage increases in the cost of living.
- 2 Housing industry economists must forecast mortgage interest rates, demand for housing and the cost of building materials.
- 3 Many companies attempt to predict the demand for their product and their share of the market.
- 4 Universities often try to forecast the number of students who will be applying for admission into their degree programs.

Forecasting is a common practice among managers and government decision makers. This chapter will focus on time series forecasting, which uses historical time series data to predict future values of variables such as sales or rates of unemployment. This entire chapter is an application tool both for economists and for managers in all functional areas of business because forecasting is such a vital factor in decision making in these areas.

For example, the starting point for aggregate production planning by operations managers is to forecast demand for the company's products. These forecasts will make use of economists' forecasts of macroeconomic variables (such as gross domestic product, disposable income and housing starts) as well as the marketing managers' internal forecasts of the future needs of their customers. Not only are these sales forecasts critical to production planning, but they are also the key to accurate pro-forma (i.e. forecast) financial statements, which are produced by the accounting and financial managers to assist in their planning for future financial needs, such as borrowing. Likewise, the human resources department will find such forecasts of a company's growth prospects to be invaluable for their planning for future manpower requirements.

There are many different forecasting techniques. Some are based on developing a model that attempts to analyse the relationship between a dependent variable and one or more independent variables. We presented some of these methods in the chapters on regression analysis (Chapters 15 and 16). The forecasting methods to be discussed in this chapter are all based on time series analysis. The first step is to analyse the components of a time series, which we discuss in Section 17.1. In Sections 17.2 to 17.5, we deal with methods for detecting which components exist and measuring them. Once we have this information, we can develop forecasting tools. We will only scratch the surface of this topic. Our objective is to expose you to the concepts of forecasting and to introduce some of the simpler techniques. The level of this text precludes us from investigating more complicated methods.

time series

A variable measured over time in sequential order.

17.1 Components of a time series

A time series may consist of four different components.

Components of a time series

- 1 Long-term trend (T)
- 2 Cyclical variation (C)
- 3 Seasonal variation (S)
- 4 Random variation (R)

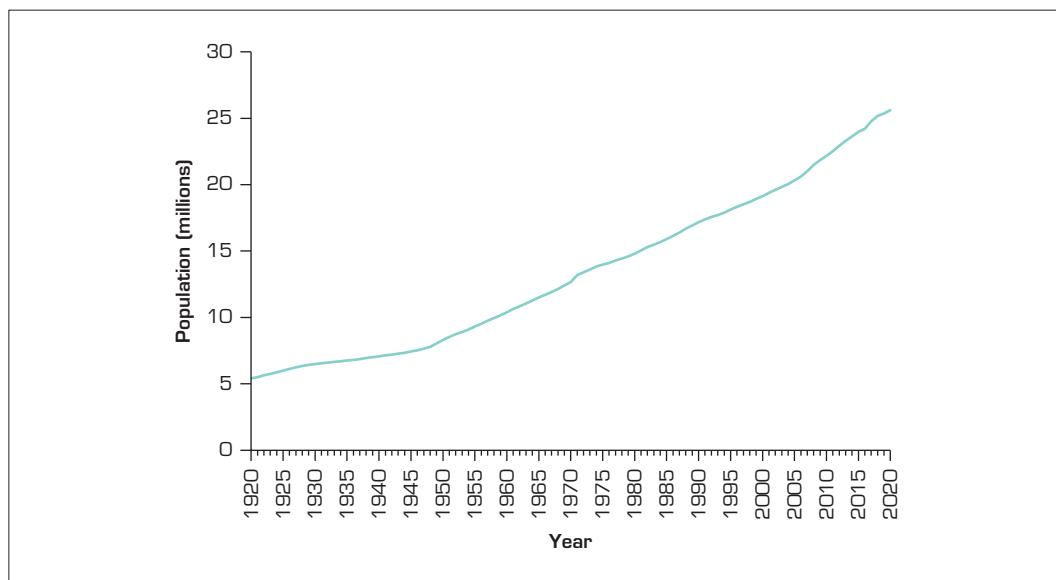
17.1a Long-term trend

trend

A long-term pattern in a time series.

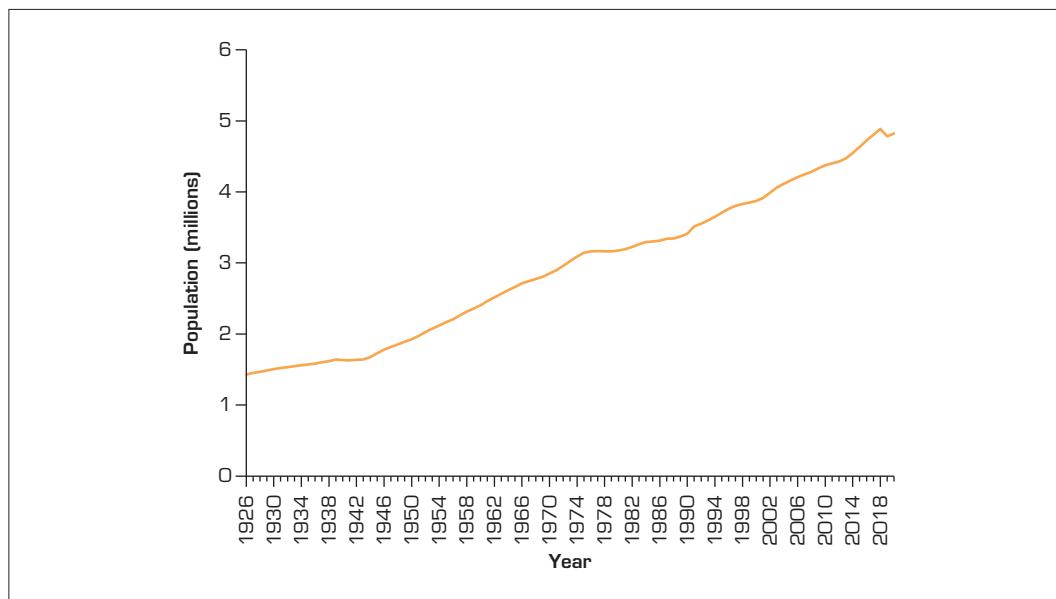
A **trend** (also known as a *secular trend*) is a long-term relatively smooth pattern or direction that the series exhibits. Its duration is more than one year. For example, the population of Australia during the past 100 years has exhibited a trend of relatively steady growth from 5.4 million in 1920 to 25.6 million in 2020 (see **Figure 17.1a**). The population of New Zealand has also had a steady growth from 1.4 million in 1926 to 4.8 million in 2020 (see **Figure 17.1b**).

FIGURE 17.1A Population of Australia, 1920–2020



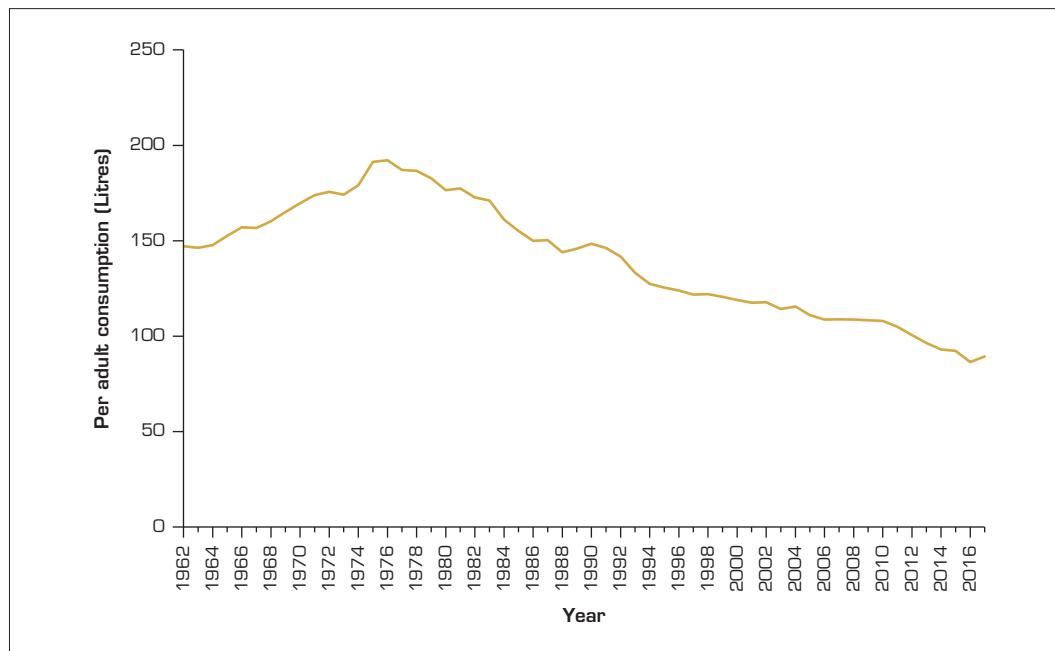
Source: Australian Bureau of Statistics, *Australian Demographic Statistics*, various issues.

FIGURE 17.1B Population of New Zealand, 1926–2020



Source: *Statistics New Zealand*, http://archive.stats.govt.nz/browse_for_stats/population/estimates_and_projections/historical-population-tables.aspx.

The trend of a time series is not always linear. For example, **Figure 17.2** describes Australian beer consumption per person from 1965 to 2017. As you can see, consumption grew between 1965 and 1974, and has been decreasing since then.

FIGURE 17.2 Per capita beer consumption, Australia, 1962–2017

Source: Australian Bureau of Statistics, various issues of *Apparent Consumption of Selected Foodstuffs, Australia*, and *Apparent Consumption of Alcohol*, cat. no. 4307.0.55.0.

17.1b Cyclical variation

A **cycle** is a wave-like pattern about a long-term trend that is generally apparent over a number of years. By definition, it has a duration of more than one year. Examples of cycles include the well-known business cycles that record periods of economic recession and inflation, long-term product-demand cycles, and cycles in the monetary and financial sectors.

Figure 17.3 displays a series of regular cycles. Unfortunately, in practice, cycles are seldom regular and often appear together with other components. The total number of building approvals in New South Wales, Australia, between 1983 and 2019 is depicted in Figure 17.4. There appear to be a number of short irregular cycles in this time series.

cycle

A wave-like trend in time series data resulting in a cyclical effect.

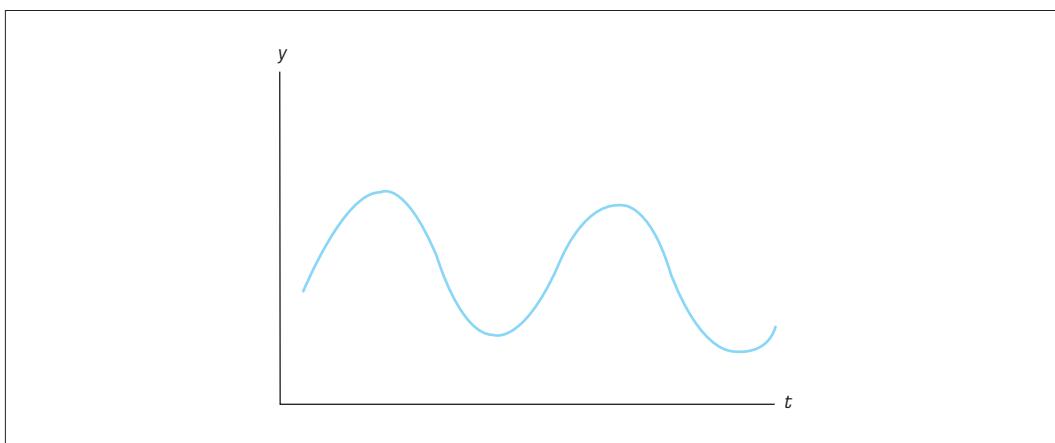
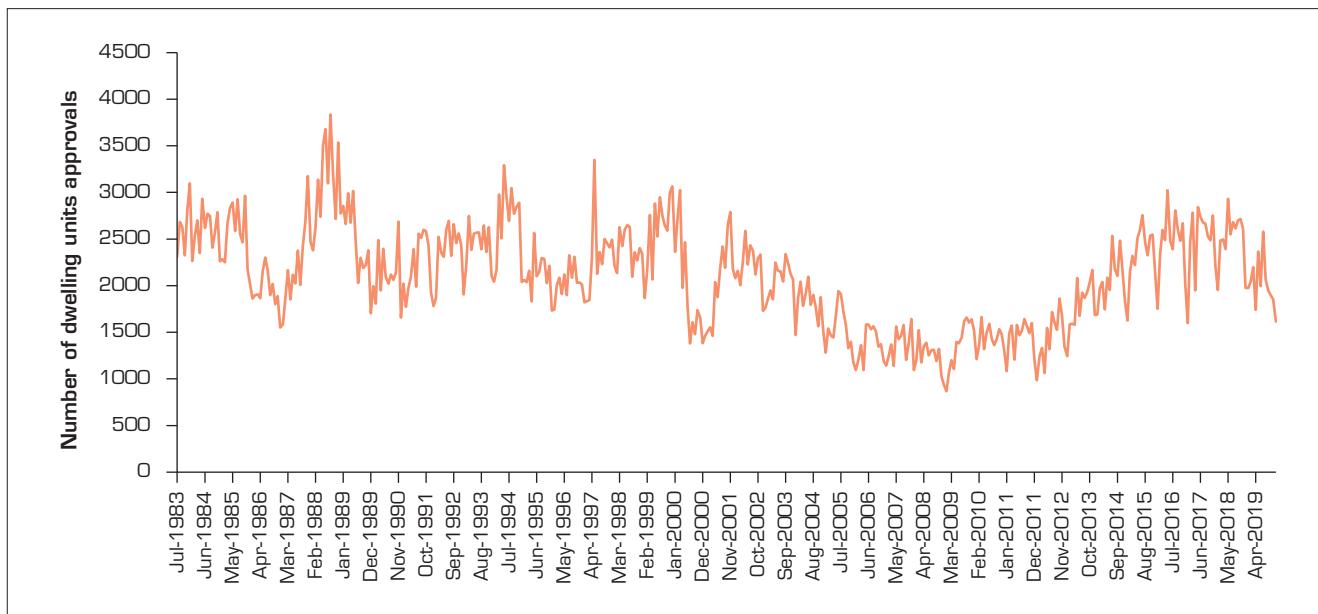
FIGURE 17.3 Cyclical variation in a time series

FIGURE 17.4 Number of monthly dwelling (house) units approvals, New South Wales, 1983–2019

Source: Australian Bureau of Statistics, ABS Cat. no. 8731.0 - Building Approvals, Australia, Dec 2019

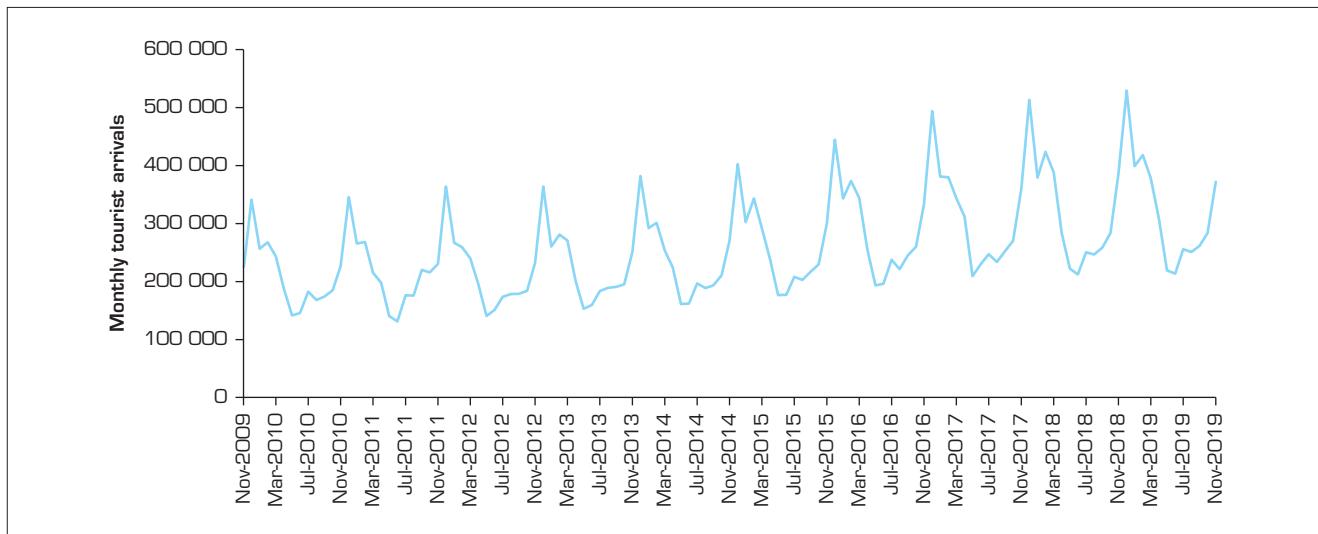
17.1c Seasonal variation

seasonal variations

Short-term seasonal cycles in time series data.

Seasonal variations are like cycles, but they occur over short, repetitive calendar periods and, by definition, have durations of less than one year. The term *seasonal variation* may refer to the four traditional seasons, or to systematic patterns that occur during the period of one week or even over the course of one day. Stock market prices, for example, often show highs and lows at particular times of the day.

An illustration of seasonal variation is provided in **Figure 17.5**, which graphs monthly tourist arrivals in New Zealand. It is obvious from the graph that there is a general increasing trend in tourist arrivals in New Zealand over the years and arrivals are higher during December than in the other months of the year.

FIGURE 17.5 Number of monthly tourist arrivals, New Zealand, Nov 2009–Nov 2019

Source: Statistics New Zealand, Monthly overseas visitor arrivals, December 2009–19, <https://www.stats.govt.nz/topics/tourism>.

17.1d Random variation

Random variation comprises the irregular changes in a time series that are not caused by any other components, and tends to hide the existence of the other, more predictable components. Because random variation exists in almost all time series, one of the functions of this chapter is to present ways to remove the random variation, thereby allowing us to describe and measure the other components and, ultimately, to make accurate forecasts. If you examine **Figures 17.1, 17.2, 17.4** and **17.5**, you will detect some degree of random variation; because of it we would not be able to predict the time series with 100% confidence, even if we had precise information about the other components. Your study of previous chapters in this book would have taught you that this is not something new. Statistics practitioners must always live with uncertainty.

random variation

Irregular changes in a time series.

17.1e Time series models

The time series model is generally expressed either as an additive model in which the value of the time series at time t is specified as

$$y_t = T_t + C_t + S_t + R_t$$

or as a multiplicative model, in which the value of the time series at time t is specified as

$$y_t = T_t \times C_t \times S_t \times R_t$$

Recall that the four components of the time series model are long-term trend (T), cyclical effect (C), seasonal effect (S) and random variation (R).

Both models may be equally acceptable; however, it is frequently easier to understand the techniques associated with time series analysis if we refer to the multiplicative model.

In the next four sections, we present ways of determining which components are present in a time series.

17.2 Smoothing techniques

If we can determine which components actually exist in a time series, we can develop a better forecast. Unfortunately, the existence of the random variation component often hides the other components. One of the simplest ways of removing the random fluctuation is to smooth the time series. In this section, we describe two methods of doing this: *moving averages* and *exponential smoothing*.

17.2a Moving averages

A **moving average** for a time period is the simple arithmetic average of the values in that time period and those close to it. For example, to calculate the three-period moving average for any time period, we would sum the value of the time series in that time period, the value in the previous time period, and the value in the following time period and divide by three. We calculate the three-period moving average for all time periods except the first and the last. To calculate the five-period moving average, we average the value in that time period, the values in the two previous time periods and the values in the two following time periods. We can choose any number of periods with which to calculate the moving averages.

moving average

The arithmetic average of a point in a time series with nearby points.

EXAMPLE 17.1

LO2

Patterns of cigarette sales I

XM17-01 As part of an effort to forecast future sales, the manager of a wholesale cigarette company in New Zealand recorded the quarterly cigarette sales (in millions) over 16 quarters. These sales are shown in the following table and recorded. Calculate the 3-quarter and 5-quarter moving averages, and then graph the quarterly cigarette sales and the moving averages.

Quarterly cigarette sales

Year	Quarter	Time period	Cigarette sales (in \$millions)
1	1	1	377
	2	2	574
	3	3	582
	4	4	903
2	1	5	356
	2	6	664
	3	7	583
	4	8	835
3	1	9	404
	2	10	626
	3	11	576
	4	12	838
4	1	13	388
	2	14	570
	3	15	575
	4	16	1017

Solution**Calculating manually**

To calculate the first 3-quarter moving averages, we group the cigarette sales in periods 1, 2 and 3 and then we average them. Thus, the first moving average is

$$\frac{377 + 574 + 582}{3} = \frac{1533}{3} = 511$$

The second moving average is calculated by dropping the sales of the first period (377), adding the sales of the fourth period (903) and then calculating the new average. Thus, the second moving average is

$$\frac{574 + 582 + 903}{3} = \frac{2059}{3} = 686.3$$

The process continues as shown in the following table. Similar calculations are used to produce the 5-quarter moving averages (also shown in the same table).

Notice that we place the moving averages in the centre of the group of values that are being averaged. It is for this reason that we prefer to use an odd number of periods in calculating moving averages. Later in this section we will discuss how to deal with an even number of periods.



Time period	Cigarette sales (in \$millions)	3-quarter moving average	5-quarter moving average
1	377		
2	574	511.0	
3	582	686.3	558.4
4	903	613.7	615.8
5	356	641.0	617.6
6	664	534.3	668.2
7	583	694.0	568.4
8	835	607.3	622.4
9	404	621.7	604.8
10	626	535.3	655.8
11	576	680.0	566.4
12	838	600.7	599.6
13	388	598.7	589.4
14	570	511.0	677.6
15	575	720.7	
16	1017		

Using the computer

Using Excel Data Analysis

Excel output for Example 17.1



We show the output for the three-period moving average only. Notice that Excel places the moving averages in the third period of each group of three rather than in the centre. The graph accompanying the table is drawn in the same way. Later in this chapter, when we present forecasting, we will discuss why Excel lists the moving averages in the way it does.

COMMANDS

- Type the data in one column or open the data file (**XM17-01**).
- Click **DATA, Data Analysis** and **Moving Average**. Click **OK**.
- Specify the **Input Range (D1:D17)**. Click **Labels** in First Row (if appropriate).
- Specify the number of periods (**3**) and the **Output Range (E2)**.
- Specify **Chart Output** if you want to graph the time series. Click **OK**.



Interpreting the results

To see how the moving averages remove some of the random variation, examine **Figures 17.6** and **17.7**.

Figure 17.6 depicts the quarterly cigarette sales in New Zealand. It is difficult to discern any of the time series components, because of the large amount of random variation. Now consider the 3-quarter moving average in **Figure 17.7**. You should be able to detect a seasonal pattern that exhibits peaks in the fourth quarter of each year (periods 4, 8, 12 and 16) and valleys in the first quarter of each year (periods 5, 9 and 13) without the effect of the random variation. There is also a small but discernible long-term trend of increasing sales.

FIGURE 17.6 Quarterly cigarette sales

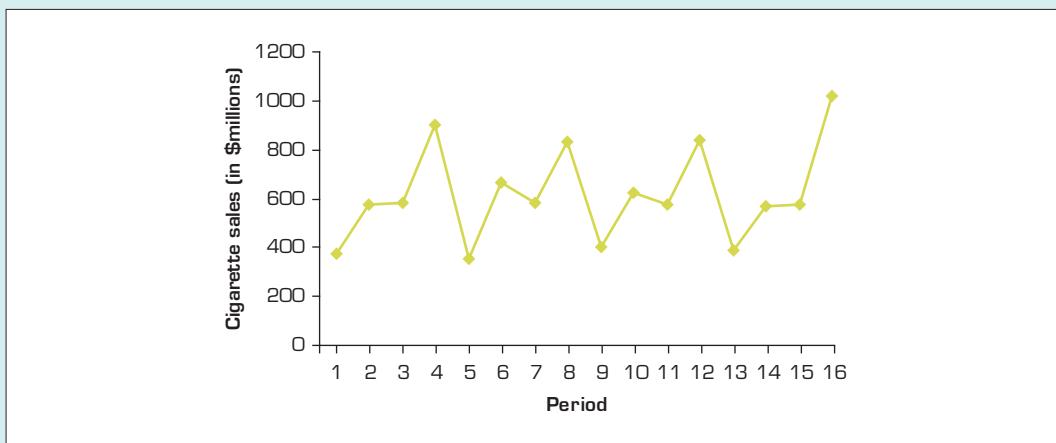
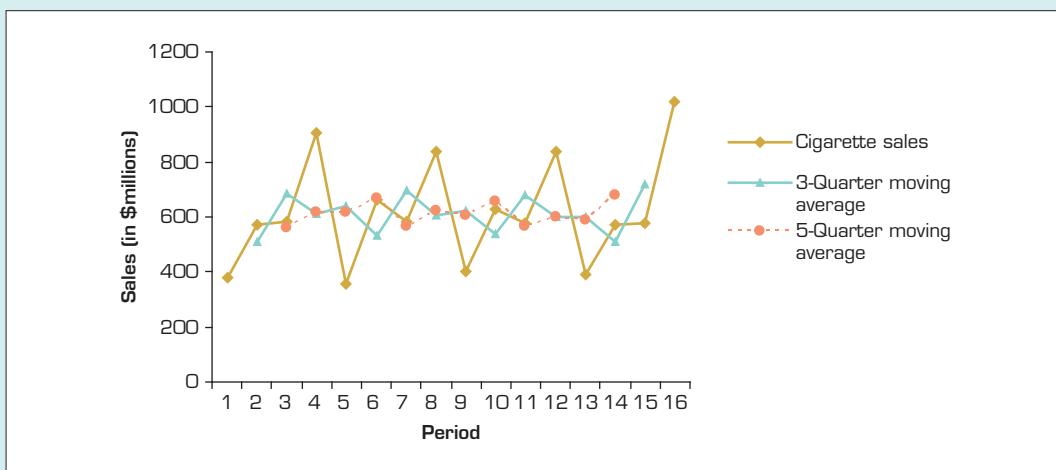


FIGURE 17.7 Quarterly cigarette sales and the 3-quarter and 5-quarter moving averages



Notice also in **Figure 17.7** that the 5-quarter moving average produces more smoothing than the 3-quarter moving average. In general, the longer the time period over which we average, the smoother the series becomes. Unfortunately, in this case we have smoothed too much, since the seasonal pattern is no longer apparent in the 5-quarter moving average. All that we can see is the long-term trend. It is important to realise that our objective is to smooth the time series sufficiently to remove the random variation and to reveal the other components (trend, cycle and/or season) present. With too little smoothing, the random variation disguises the real pattern. With too much smoothing, however, some or all of the other effects may be eliminated along with the random variation.

17.2b Centred moving averages

If we use an even number of periods in calculating the moving averages, we have to work out where to place the moving averages on a graph or a table. For example, suppose that we calculate the four-period moving average of the following data:

Period	Time series
1	15
2	27
3	20
4	14
5	25
6	11

The first moving average is

$$\frac{15+27+20+14}{4} = \frac{76}{4} = 19.$$

However, since this value represents time periods 1, 2, 3 and 4, we must place this value between periods 2 and 3. The next moving average is

$$\frac{27+20+14+25}{4} = \frac{86}{4} = 21.5$$

and it must be placed between periods 3 and 4. The moving average that falls between periods 4 and 5 is

$$\frac{20+14+25+11}{4} = \frac{70}{4} = 17.5$$

Having the moving average fall between the time periods causes various problems, including the difficulty of graphing. Centring the moving averages corrects the problem. This is performed by calculating the two-period moving average of the four-period moving averages. Thus, the **centred moving average** for period 3 is

$$\frac{19.0+21.5}{2} = 20.25$$

Similarly, the centred moving average for period 4 is

$$\frac{21.5+17.5}{2} = 19.5$$

centred moving average

A technique for centring the moving averages when the number of time periods used to calculate the averages is an even number.

Period	Time series, y_t	Four-period moving average	Four-period centred moving average
1	15	–	–
2	27	–	–
		19.0	
3	20		20.25
		21.5	
4	14		19.50
		17.5	
5	25	–	–
6	11	–	–

The following table summarises our results.

Period	Time series, y_t	Four-period centred moving average
1	15	–
2	27	–
3	20	20.25
4	14	19.50
5	25	
6	11	

Because of the extra calculation involved in centring a moving average, we prefer to use an odd number of periods. However, in some situations we are required to use an even number of periods. Such cases are discussed in Section 17.6. Excel does not centre the moving averages.

17.2c Exponential smoothing

Two drawbacks are associated with the moving averages method of smoothing a time series. First, we do not have moving averages for the first and last sets of time periods. If the time series has few observations, the missing values may constitute an important loss of information. Second, the moving average ‘forgets’ most of the previous time series values. For example, in the 5-quarter moving average described in Example 17.1, the average for period 4 reflects periods 2, 3, 4, 5 and 6 but is not affected by period 1. Similarly, the moving average for period 5 forgets periods 1 and 2. Both of these problems are addressed by **exponential smoothing**.

exponential smoothing

A technique for smoothing a time series in a way that includes all data prior to time t in the smoothed time series in period t .

The exponentially smoothed time series is defined next.

Exponentially smoothed time series

$$ES_1 = y_1 \quad \text{and} \quad ES_t = \omega y_t + (1-\omega)ES_{t-1} \quad \text{for } t \geq 2$$

where

ES_t = exponentially smoothed time series at time t

y_t = time series at time t

ES_{t-1} = exponentially smoothed time series at time $t - 1$

ω = smoothing constant, where $0 \leq \omega \leq 1$

We begin by setting

$$ES_1 = y_1$$

Then

$$\begin{aligned} ES_2 &= \omega y_2 + (1-\omega)ES_1 \\ &= \omega y_2 + (1-\omega)y_1 \\ ES_3 &= \omega y_3 + (1-\omega)ES_2 \\ &= \omega y_3 + (1-\omega)[\omega y_2 + (1-\omega)y_1] \\ &= \omega y_3 + \omega(1-\omega)y_2 + (1-\omega)^2 y_1 \end{aligned}$$

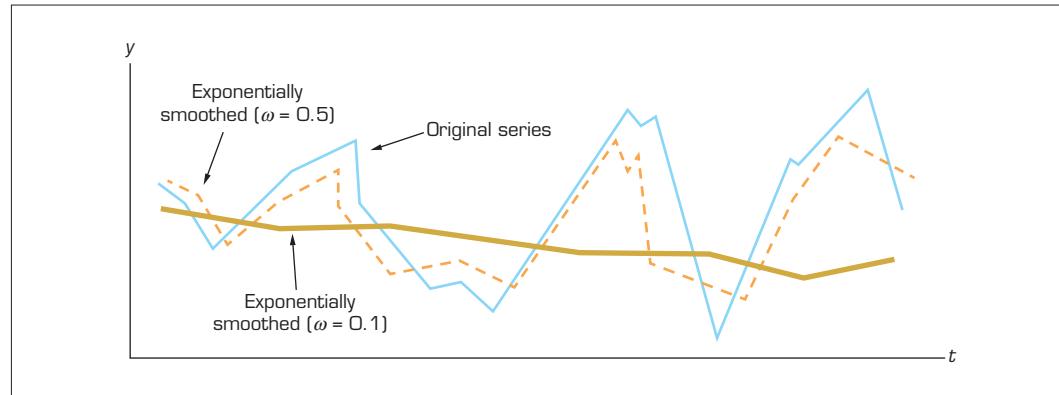
and so on. In general, we have

$$ES_t = \omega y_t + \omega(1-\omega)y_{t-1} + \omega(1-\omega)^2 y_{t-2} + \dots + (1-\omega)^{t-1} y_1$$

This formula states that the smoothed time series in period t depends on all the previous observations of the time series.

The smoothing constant ω is chosen on the basis of how much smoothing is required. A small value of ω produces a great deal of smoothing. A large value of ω results in very little smoothing. **Figure 17.8** depicts a time series and two exponentially smoothed series with $\omega = 0.1$ and $\omega = 0.5$.

FIGURE 17.8 Original time series and two exponentially smoothed series



EXAMPLE 17.2

LO2

Patterns of cigarette sales II

Apply the exponential smoothing technique with $\omega = 0.2$ and $\omega = 0.7$ to the data in Example 17.1, and graph the results.

Solution

Calculating manually

The exponentially smoothed values are calculated from the formula

$$ES_1 = y_1 \quad \text{and} \quad ES_t = \omega y_t + (1 - \omega)ES_{t-1}, \quad t = 2, 3, 4, \dots, 16$$

The results with $\omega = 0.2$ and $\omega = 0.7$ are shown in the following table.

Period	Quarter	Cigarette sales (in \$millions)	Exponentially smoothed sales ES_t with $\omega = 0.2$	Exponentially smoothed sales ES_t with $\omega = 0.7$
1	1	377	377.0	377.0
	2	574	0.2(574) + 0.8(377) = 416.4	0.7(574) + 0.3(377) = 514.9
	3	582	0.2(582) + 0.8(416.4) = 449.5	0.7(582) + 0.3(514.9) = 561.9
	4	903	0.2(903) + 0.8(449.5) = 540.2	0.7(903) + 0.3(561.9) = 800.7
2	1	356	503.4	489.4
	2	664	535.5	611.6
	3	583	545.0	591.6
	4	835	603.0	762.0
3	1	404	563.2	511.4
	2	626	575.8	591.6
	3	576	575.8	580.7
	4	838	628.2	760.8
4	1	388	580.2	499.8
	2	570	578.2	549.0
	3	575	577.5	567.2
	4	1017	0.2(1017) + 0.8(577.5) = 665.4	0.7(1017) + 0.3(567.2) = 882.1



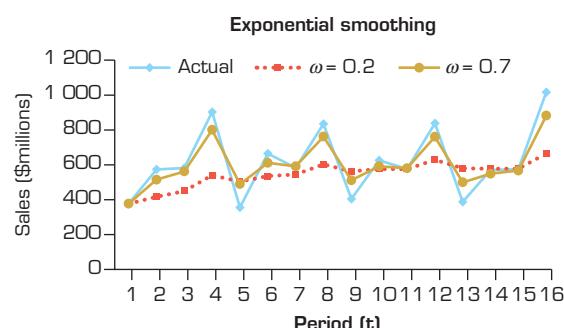


Using the computer

We show the output for $\omega = 0.2$ and $\omega = 0.7$. The printout for exponential smoothing is similar to that of moving averages.

Excel output for Example 17.2

	A	B	C	D	E	F						
1	Year	Quarter	Period (t)	Cigarette sales [in millions]	Cigarette sales [$\omega = 0.2$]	Cigarette sales [$\omega = 0.7$]						
2	1	1	1	377	#NA	#NA						
3		2	2	574	377.0	377.0						
4		3	3	582	416.4	514.9						
5		4	4	903	449.5	561.9						
6	2	1	5	356	540.2	800.7						
7		2	6	664	503.4	489.4						
8		3	7	583	535.5	611.4						
9		4	8	835	545.0	591.6						
10	3	1	9	404	603.0	762.0						
11		2	10	626	563.2	511.4						
12		3	11	576	575.8	591.6						
13		4	12	838	575.8	580.7						
14	4	1	13	388	628.2	760.8						
15		2	14	570	580.2	499.8						
16		3	15	575	578.2	549.0						
17		4	16	1017	577.5	567.2						



We show the output for $\omega = 0.2$ (damping factor $1 - \omega = 0.8$) and $\omega = 0.7$. This output is similar to the one produced to show moving averages. Note that in Excel printouts the smoothed values appear one period later than the way we calculated them above.

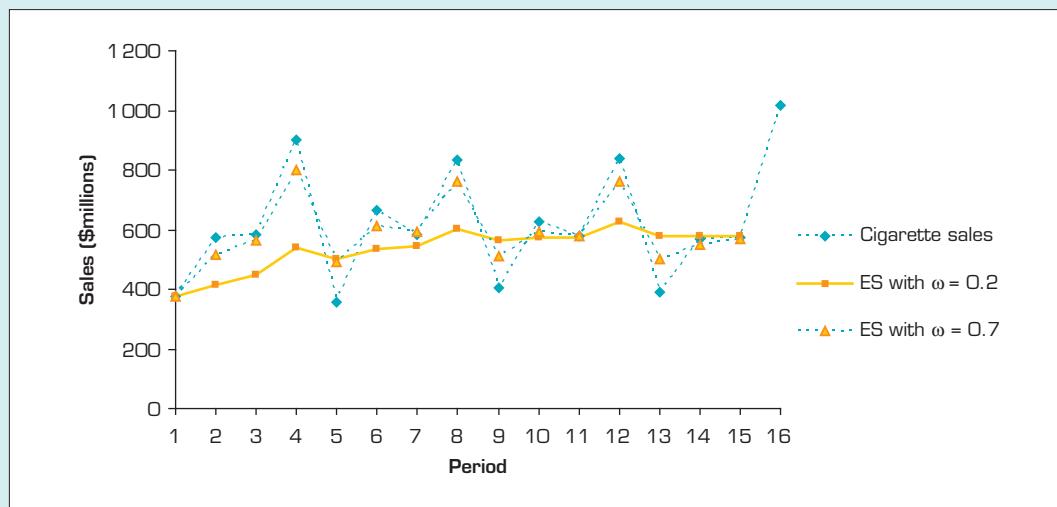
COMMANDS

- 1 Type the data in one column or open the data file (**XM17-01**).
- 2 Click **DATA**, **Data Analysis** and **Exponential Smoothing**. Click **OK**.
- 3 Specify the input range (**D1:D17**).
- 4 Hit tab, and type the value of the **Damping factor**, which is $(1 - \omega)$ (**0.8**). Click **Labels** (if appropriate).
- 5 Specify the output range (**F2**).
- 6 Specify **Chart Output** if you want a graph. Click **OK**.

Interpreting the results

Figure 17.9 depicts the graph of the original time series and the exponentially smoothed series. As you can see, $\omega = 0.7$ results in very little smoothing, while $\omega = 0.2$ results in perhaps too much smoothing. In both smoothed time series, it is difficult to discern the seasonal pattern that we detected by using moving averages. A different value of ω (perhaps $\omega = 0.5$) would be likely to produce more satisfactory results.



FIGURE 17.9 Quarterly cigarette sales and exponentially smoothed sales with $\omega = 0.2$ and $\omega = 0.7$ 

Moving averages and exponential smoothing are relatively crude methods of removing random variations in order to discover the existence of other components. In the next three sections, we will attempt to measure the components more precisely. So far, we have used exponential smoothing only to smooth a time series in order to better detect the components of the series. Using exponential smoothing for forecasting will be discussed in Section 17.7.

EXERCISES

Learning the techniques

17.1 XR17-01 Consider the following time series $\{y_t\}$:

Period t	y_t	Period t	y_t
1	16	7	24
2	22	8	29
3	19	9	21
4	24	10	23
5	30	11	19
6	26	12	15

- a Calculate the 3-period moving averages.
- b Calculate the 5-period moving averages.
- c For parts (a) and (b) graph the time series and the two moving averages.

17.2 XR17-02 Consider the following time series $\{y_t\}$:

Period t	y_t	Period t	y_t
1	48	7	43
2	41	8	52
3	37	9	60
4	32	10	48
5	36	11	41
6	31	12	30

- a Calculate the 3-period moving averages.

- b Calculate the 5-period moving averages.

- c For parts (a) and (b) graph the time series and the two moving averages.

17.3 XR17-03 Calculate the 4-period centred moving averages for the following time series $\{y_t\}$:

Period t	y_t	Period t	y_t
1	44	5	66
2	42	6	62
3	49	7	63
4	56	8	49

17.4 a XR17-04 Apply exponential smoothing with $\omega = 0.1$ to help detect the components of the following time series $\{y_t\}$:

Period t	y_t	Period t	y_t
1	12	6	16
2	18	7	25
3	16	8	21
4	24	9	23
5	17	10	14

- b** Repeat part (a) with $\omega = 0.8$.
- c** For parts (a) and (b) draw the time series and plot the two sets of exponentially smoothed values. Does there appear to be a trend component in the time series?

- 17.5 XR17-05** Apply exponential smoothing with $\omega = 0.1$, to help detect the components of the following time series $\{y_t\}$:

Period t	y_t	Period t	y_t
1	38	6	48
2	43	7	50
3	42	8	49
4	45	9	46
5	46	10	45

- 17.6 a** Repeat Exercise 17.5 with $\omega = 0.8$.
- b** Draw the time series and the two sets of exponentially smoothed values in Exercise 17.5 and part (a). Does there appear to be a trend component in the time series?

Applying the techniques

- 17.7 XR17-07 Self-correcting exercise.** The following daily sales were recorded in a medium-size merchandising firm.

Day	Week			
	1	2	3	4
Monday	43	51	40	64
Tuesday	45	41	57	58
Wednesday	22	37	30	33
Thursday	25	22	33	38
Friday	31	25	37	25

- a** Plot the series on a graph.
- b** Calculate the three-day moving averages, and superimpose them on the graph.
- c** Does there appear to be a seasonal (weekly) pattern?
- d** Calculate the five-day moving averages and superimpose them on the same graph. Does this help you to answer part (c)?

- 17.8 a XR17-08** The quarterly sales of a department store chain were recorded for the years 2016–19. These data are shown in the table below.

Quarter	Week			
	2016	2017	2018	2019
1	18	33	25	41
2	22	20	36	33
3	27	38	44	52
4	31	26	29	45

- i** Graph the time series.
- ii** Calculate the 4-quarter centred moving averages and superimpose them on the time series graph.
- iii** What can you conclude from your time series smoothing?
- b** Repeat part (a), using exponential smoothing with $\omega = 0.4$.
- c** Repeat part (a), using exponential smoothing with $\omega = 0.8$.

Computer applications

The following exercises require the use of a computer and software.

- 17.9 XR17-09** The quarterly numbers of tourist arrivals to Australia for the years 2009–19 are recorded. (Source: Australian Bureau of Statistics, *Overseas Arrivals and Departures*, ABS cat. no. 3401.0, Australia, March 2020.)

- a** Calculate the 4-quarter centred moving averages.
- b** Plot the time series and the moving averages on the same graph.
- c** Does there appear to be a quarterly seasonal pattern?
- d** Does there appear to be a trend?

- 17.10** Apply the exponential smoothing technique with $\omega = 0.2$ and $\omega = 0.7$ to the data in Exercise 17.9.

- a** Plot the time series and the exponentially smoothed values on the same graph.
- b** Does there appear to be a quarterly seasonal pattern?
- c** Does there appear to be a trend?

17.3 Trend analysis

In the last section, we described how smoothing a time series can give us a clearer picture of which components are present. In order to forecast, however, we often need more precise measurements of the time series components. In this section we will discuss methods that allow us to describe trend. In subsequent sections, we will consider how to measure cyclical and seasonal effects.

As we noted earlier, a trend can be linear or non-linear and, indeed, can take on a whole host of other functional forms, some of which we will discuss. The easiest way of isolating the long-term trend is by regression analysis, in which the independent variable is time.

If we believe that the long-term trend is essentially linear, we will use the linear trend model.

17.3a Linear trend model

Linear model for long-term trend

$$y_t = \beta_0 + \beta_1 t + \epsilon$$

where t is the time period.

If we believe that the trend is non-linear, we can use one of the polynomial models. As described in that chapter, if the time series is non-linear with one change in slope, the second-order polynomial (which is also called quadratic) model may be best.

17.3b Polynomial trend model

Quadratic model for long-term trend

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \epsilon$$

Cubic model for long-term trend

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \epsilon$$

The quadratic or cubic model may apply, for example, to a new product that has experienced a rapid early growth rate followed by the inevitable levelling off. Examples 17.3 and 17.4 illustrate how and when these models are used.

EXAMPLE 17.3

LO3

Measuring the trend in Australian population data

XM17-03 The Australian Bureau of Statistics publishes Australian population numbers on a regular basis. The annual data for the past 101 years from 1920 to 2020 (part of which is shown in the following table) are recorded. An analyst believes that there is a strong linear trend component in Australian population data over this period. Use regression analysis to measure the trend.

Year	Period t	Population (millions)
1920	1	5.41
1921	2	5.51
1922	3	5.64
...
2017	98	24.77
2018	99	25.17
2019	100	25.36
2020	101	25.61

Source: Australian Bureau of Statistics, *Australian Historical Population Statistics*, 2016, ABS cat. no. 3105.0.65.001, and *Australian Demographic Statistics*, ABS cat. no. 3101.0, Jun 2019.

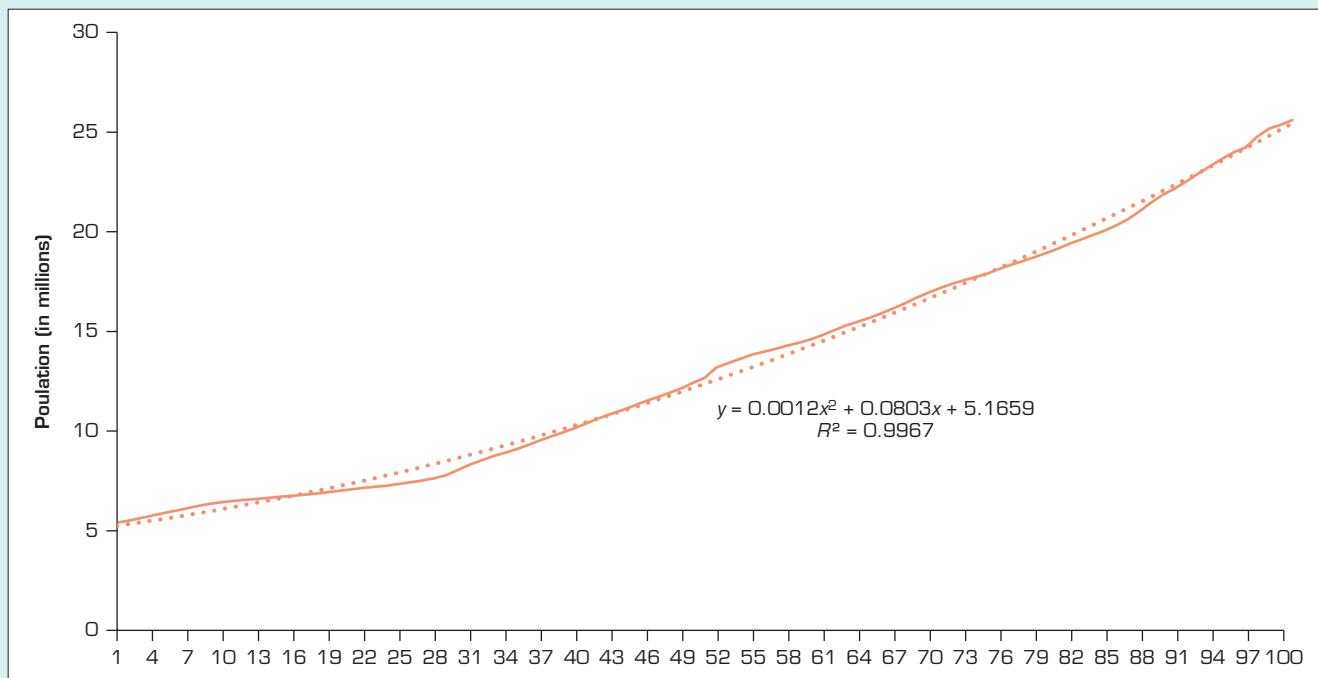
Solution

Although this technique can be performed manually, realistic applications use the computer exclusively.

The time series of the Australian population depicted in **Figure 17.10** shows the presence of a strong non-linear trend. Therefore, we propose a quadratic trend model:

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon$$

FIGURE 17.10 Time series and trend line for Example 17.3



We can use the regression techniques introduced in Chapters 15–16 for estimation. To do so, we must store the values of t and t^2 for the quadratic model. It is easier (though not necessary) to change the times from years 1920 to 2020 to time periods 1 to 101. When that was done, we estimated the model, resulting in the following output.



Using the computer

Excel regression output for Example 17.3

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.9983					
5	R Square	0.9967					
6	Adjusted R Square	0.9966					
7	Standard Error	0.3488					
8	Observations	101					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	2	3584.018	1792.009	14731.47	3.198E-122	
13	Residual	98	11.921	0.122			
14	Total	100	3595.939				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	5.1659	0.1062	48.6385	0.0000	4.9551	5.3767
18	t	0.0803	0.0048	16.7165	0.0000	0.0708	0.0899
19	t ²	0.0012	0.0000	26.1115	0.0000	0.0011	0.0013

The fitted trend equation is $\hat{y}_t = 5.1659 + 0.0803t + 0.0012t^2$.

Interpreting the results

The value of $R^2 = 0.997$ or 99.7% indicates an excellent fit of the quadratic trend. However, in general, because of the possible presence of cyclical and seasonal effects and because of random variation, we do not usually expect a very good fit of a quadratic trend for most time series. The regression trend line that we just estimated is superimposed on the graph of the time series in **Figure 17.10** and shows a clear upwards trend.

One of the purposes of isolating the trend, as we suggested earlier, is to use it for forecasting. For example, we could use it for forecasting one year in advance, through 2021 ($t = 102$). From our quadratic trend equation, we get

$$\begin{aligned}\hat{y}_{t=102} &= 5.1659 + 0.0803t + 0.0012t^2 \\ &= 5.1659 + 0.0803(102) + 0.0012(102)^2 \\ &= 25.76 \text{ million}\end{aligned}$$

This value, however, represents the forecast based only on trend. If we believe that a cyclical pattern also exists, we should incorporate that into the forecast as well.

EXAMPLE 17.4

LO3

Measuring the trend in Australian per capita beer consumption

XM17-04 Australians are among the world's greatest beer drinkers. In the last decade, however, because of road safety measures and health concerns, many Australians have decreased their consumption of beer. The effect has been serious for brewers. To help analyse the problem, the annual per capita consumption of beer in Australia has been recorded for the period 1962–2018. Some of these data are shown in the following table. Apply regression analysis to determine the long-term trend.

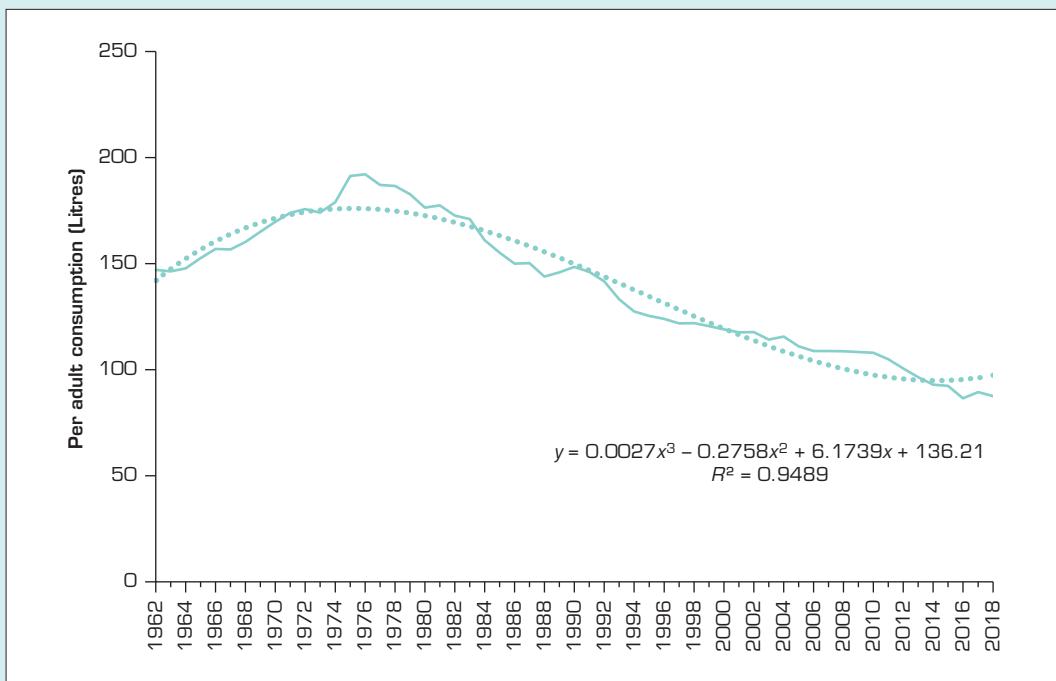
Year	Period t	Beer consumption (litres/capita)	Year	Period t	Beer consumption (litres/capita)
1962	1	147.13	1991	30	146.19
1963	2	146.36
1964	3	147.77	2014	53	93.02
1965	4	152.56	2015	54	92.38
1966	5	157.01	2016	55	86.49
...	2017	56	89.43
1990	29	148.45	2018	57	87.56

Source: Australian Bureau of Statistics, various issues of *Apparent Consumption of Selected Foodstuffs*, cat. no. 4307.0.55.001, Australia.

Solution

The time series of per capita beer consumption is depicted in Figure 17.11.

FIGURE 17.11 Time series and cubic trend for Example 17.4



An examination of the data displayed in **Figure 17.11** reveals that between 1962 and 1975 (periods 1 and 14), Australian beer consumption rose steadily but has declined somewhat erratically since then. The pattern suggests that a cubic model might be best. The model

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \varepsilon$$

was estimated with the following results. (Time periods were coded so that $t = 1$ represents 1962 and $t = 57$ represents 2018.) Before performing a regression to estimate this cubic model, remember to set up a column of values of t , a column for values of t^2 and a column for values of t^3 .



Using the computer

Excel output for Example 17.4

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.974					
5	R Square	0.949					
6	Adjusted R Square	0.946					
7	Standard Error	7.134					
8	Observations	57					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	3	50093.5	16697.8	328.08	0.000	
13	Residual	53	2697.4	50.9			
14	Total	56	52791.0				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	136.214	4.043	33.694	0.000	128.105	144.322
18	t	6.174	0.598	10.317	0.000	4.974	7.374
19	t^2	-0.276	0.024	-11.558	0.000	-0.324	-0.228
20	t^3	0.003	0.000	10.083	0.000	0.002	0.003

The fitted trend line is

$$\hat{y}_t = 136.21 + 6.174t - 0.276t^2 + 0.003t^3 \quad \text{Adjusted } R^2 = 0.95, F = 328.08$$

Interpreting the results

All of the statistics (Adjusted $R^2 = 0.95$; $F = 328.08$; p -value = 0) indicate that the cubic model fits the data quite well. Thus, the trend is measured by

$$\hat{y}_t = 136.21 + 6.174t - 0.276t^2 + 0.003t^3$$

Figure 17.11 depicts the time series and the fitted cubic trend equation.

EXERCISES

Learning the techniques

17.11 XR17-11 Consider the following time series data $\{y_t\}$.

Period t	y_t	Period t	y_t
1	0.5	5	4.1
2	0.6	6	6.9
3	1.3	7	10.8
4	2.7	8	19.2

- a Plot the time series. Which model (linear or polynomial) would fit better?
- b Use the regression technique to calculate the linear trend line and the quadratic trend line. Which line fits better?

17.12 a XR17-12 Plot the following time series $\{y_t\}$ to determine which model appears to fit better.

Period t	y_t	Period t	y_t
1	55	6	39
2	57	7	41
3	53	8	33
4	49	9	28
5	47	10	20

- b For the data in part (a), use the regression technique to calculate the linear trend line and the quadratic trend line. Which line fits better?

Applying the techniques

The following exercises require the use of a computer and software.

- 17.13 XR17-13** The following table shows the enrolment numbers $\{y_t\}$ of Year 12 students at a high school in Queensland for the last 11 years.

Period t	y_t	Period t	y_t
1	185	7	242
2	198	8	243
3	213	9	250
4	225	10	248
5	235	11	253
6	240		

- a Plot the time series.
- b Which of the trend models appears to fit better? Explain.
- c Use regression analysis to calculate the linear and quadratic trend lines.
- d Which line fits the time series better? Explain.
- e Forecast the student enrolments for the next two time periods based on the linear and quadratic trends you estimated above.

Computer applications

- 17.14 XR17-14** The Australian Bureau of Statistics (ABS) publishes the Gross Domestic Product (GDP) for Australia on a regular basis. The annual GDP data (in

billions of dollars) for the past 59 years from 1960 to 2018 were recorded. A business analyst believes that the trend over this period is basically linear.

- a Plot the time series.
- b Which trend model is likely to fit better? Explain.
- c Use regression analysis to measure the trend.

- 17.15 XR17-15** Exports are an important component of the exchange rate and, domestically, are an important indicator of employment and profitability in certain industries. The value of Australian exports has increased in the 46-year period (1973–2018). The export data (in millions of dollars) are recorded. (Source: Australian Bureau of Statistics, *Balance of Payments and International Investment*, Australia, cat. no. 5302.0.)

- a Plot the time series.
- b Which trend model is likely to fit better? Explain.
- c Estimate a linear trend line.

$$y_t = \beta_0 + \beta_1 t + \varepsilon \quad (t = 1, 2, \dots, 46)$$

Plot the estimated line on the graph.

- d Estimate a quadratic trend line.

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon \quad (t = 1, 2, \dots, 46)$$

Plot the estimated line on the graph.

- e Which of the two trend models in parts (c) and (d) appears to provide the better fit? Forecast the Australian exports based on both models for the years 2019, 2020 and 2021.

17.4 Measuring the cyclical effect

The fundamental difference between cyclical and seasonal variations lies in the length of the time period under consideration. In addition, however, seasonal effects are considered to be predictable, whereas cyclical effects (except in the case of certain well-known economic and business cycles) are often viewed as being unpredictable – varying in both duration and amplitude and not necessarily even being repetitive. Nevertheless, cycles need to be isolated, and the measure we use to identify cyclical variation is the **percentage of trend**.

The percentage of trend is calculated in the following way:

- 1 Determine the trend line (by regression).
- 2 For each time period, calculate the trend value \hat{y}_t .
- 3 The percentage of trend is $(y_t / \hat{y}_t) \times 100$.

percentage of trend

The amount of trend produced by a given effect.

EXAMPLE 17.5

LO4

Measuring the cyclical effect on New Zealand TELECOM stock prices

XM17-05 The following table shows the daily closing stock prices for New Zealand TELECOM for part of the period 7 February 2019 to 5 February 2020.¹ Assuming a linear trend, calculate the percentage of trend for each day.

Date	Period, t	Stock price (NZ\$)
7-Feb-19	1	4.030
8-Feb-19	2	4.050
11-Feb-19	3	4.025
12-Feb-19	4	4.055
13-Feb-19	5	4.095
14-Feb-19	6	4.080
...
...
29-Jan-20	248	4.550
30-Jan-20	249	4.625
31-Jan-20	250	4.660
3-Feb-20	251	4.620
4-Feb-20	252	4.605
5-Feb-20	253	4.680

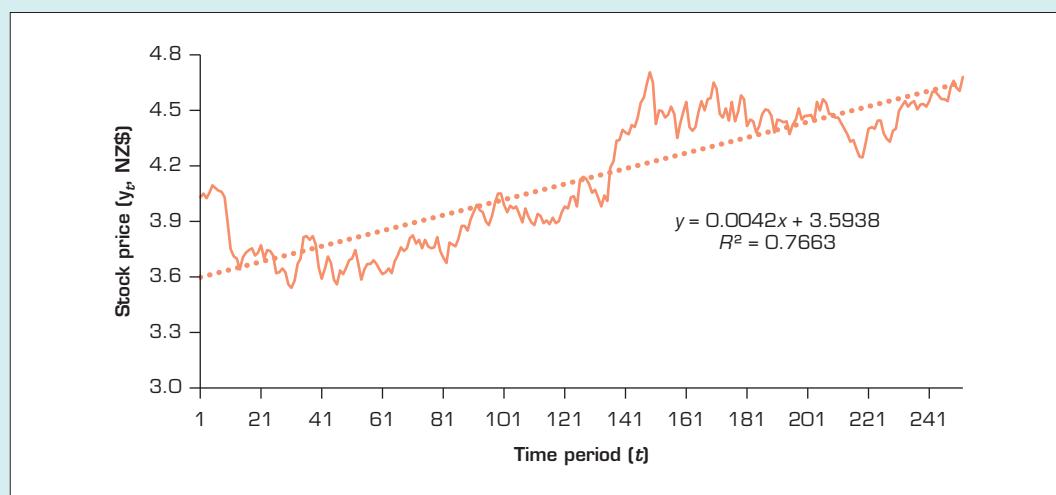
Source: <https://nz.finance.yahoo.com/quote/CMTL/history?p=CMTL&guccounter=1>.

Using the computer**Using Excel Data Analysis**

The time series of share prices depicted in **Figure 17.12** shows the presence of a linear trend. From the Excel output, we observe that the trend line is

$$\hat{y}_t = 3.595 - 0.0042t$$

FIGURE 17.12 Time series and trend line for Example 17.5



Source: Yahoo! New Zealand Business & Finance, <http://nz.finance.yahoo.com/q/hp?s=TEL.NZ>

¹ Weekends and New Zealand public holidays excluded.



For each value of t ($t = 1, 2, \dots, 253$), the predicted values, \hat{y}_t , (also available from the output) and the percentage of trend, y_t / \hat{y}_t , were determined using Excel.

Excel output for Example 17.5

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.876					
5	R Square	0.767					
6	Adjusted R Square	0.766					
7	Standard Error	0.169					
8	Observations	252					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	23.470	23.470	822.023	0.000	
13	Residual	250	7.138	0.029			
14	Total	251	30.608				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	3.5950	0.0213	168.699	0.000	3.5530	3.6370
18	t	0.0042	0.0001	28.671	0.000	0.0039	0.0045

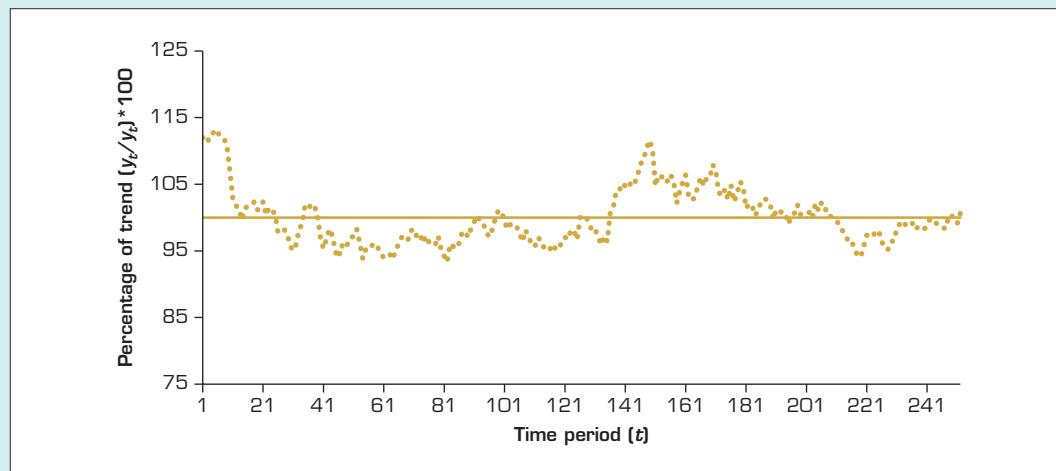
COMMANDS

- Type the data or open the data file (**XM17-05**). Store the month and year in the first column and the dependent variable in the second column. For the linear model, store the values of t in the second column. For the quadratic model, store values of t and t^2 in second and third columns.
- Click **DATA**, **Data Analysis** and **Regression**. Click **OK**.
- Specify the **Input Y Range** (time series) (**B1:B253**) and the **Input X Range** (t for the linear model and t and t^2 for the quadratic model) (**C1:C253**). Click **Labels** (if appropriate).
- Click **New Worksheet Ply**: for the regression output and type name of the new worksheet (**Regression**). To obtain the predicted y -values (as well as residuals), tick the box **Residuals** and click **OK**.
- In the column next to t (**column D**), type column label $y^$ (**in cell D1**) and copy the predicted values of y from the Regression output (**Regression worksheet B25:B276 to D2:D253**). Alternatively, using the regression equation calculated in steps 2, 3 and 4, calculate $y^$ (**by typing in cell D2 =Regression!\$B\$17+Regression!\$B\$18*C2 and copy the formula for cells D3–D253**).
- In the column next to $y^$ (**column E**), type the column label ' $y/y^$ ' (**in cell E1**) and calculate the percentage of trend (**Type =100*B2/D2 in cell E2 and copy the formula for cells E3–E253**).
- Plot the percentage of trend in a line chart.

Interpreting the results

Figure 17.12 describes the time series and the trend line. The percentage of trend represents the amount by which the actual share price lies above or below the line. **Figure 17.13** is another way of depicting these values. The trend line now appears as the 100% line.



FIGURE 17.13 Percentage of trend for Example 17.5

The problem we face in trying to interpret **Figure 17.13** is that of distinguishing between a random variation and a cyclical pattern. If there appears to be a random collection of percentage of trend values above and below the 100% line, we could conclude that its cause is random and not cyclical. However, if we see alternating groups of percentage of trend values above and below the 100% line and the patterns are regular, we could confidently identify the cyclical effect. In **Figure 17.13** there appears to be a cyclical pattern, although it is somewhat irregular. This example highlights the main problem of forecasting time series that possess a cyclical component; the cyclical effect is often quite clearly present but too irregular to forecast with any degree of accuracy. Forecasting methods for this type of problem are available, but they are too advanced for our use. We will be satisfied with simply identifying and measuring the cyclical component of time series.

EXERCISES

The following exercises require the use of a computer and software.

Learning the techniques

17.16 XR17-16 Consider the time series shown in the following table.

Period t	y_t	Period t	y_t
1	30	9	41
2	27	10	38
3	24	11	43
4	21	12	36
5	23	13	29
6	27	14	24
7	33	15	20
8	38	16	18

- a** Calculate the percentage of trend for each time period.

- b** Plot the percentage of trend.
c Describe the cyclical effect (if there is one).

17.17 XR17-17 For the time series shown in the following table:

Period t	y_t	Period t	y_t
1	6	7	20
2	11	8	22
3	21	9	18
4	17	10	17
5	27	11	12
6	23	12	15

- a** Plot the time series.
b Plot the trend line.
c Calculate the percentage of trend.
d Plot the percentage of trend.
e Describe the cyclical effect (if there is one).

Applying the techniques

- 17.18 XR17-18 Self-correcting exercise.** The monthly cash rates in Australia between 1990 and 2020 are recorded. (Source: © Reserve Bank of Australia. CC BY 4.0 International, <https://creativecommons.org/licenses/by/4.0/legalcode>)
- Plot the time series and graph the trend line.
 - Calculate the percentage of trend.
 - Plot the percentage of trend. Does there appear to be a cyclical effect?
- 17.19 XR17-19** As a preliminary step in forecasting future values, a large mail-order retail outlet has recorded the annual sales values (in millions) over a 21-year period.
- Plot the time series.
 - Plot the trend line.
 - Calculate the percentage of trend.
 - Does there appear to be a cyclical pattern? Describe it (if there is one).
- 17.20 XR17-20** One of the key statistics governments and businesses use these days to measure the pulse
- of the economy is the consumer sentiment index. A historical series of such an index, the monthly *Westpac–Melbourne Institute Consumer Sentiment Index*, is recorded for the period January 2010 to January 2020. (Source: *Westpac–Melbourne Institute Consumer Sentiment Index*, Melbourne Institute of Applied Economics and Social Research, www.economagic.com)
- Plot the time series and graph the trend line.
 - Calculate the percentage of trend.
 - Plot the percentage of trend. Does there appear to be a cyclical effect?
- 17.21 XR17-21** The Australian All Ordinaries share price index is one of the main economic indicators of Australian financial markets. The data for the monthly Australian All Ordinaries Index for January 2007–February 2020 are recorded. Assuming a linear trend, calculate the percentage of trend for each month. (Source: <https://au.finance.yahoo.com/q/hp?s=%AORD&a=07&b=3&c=1984&d=11&e=12&f=2015&g=m>)

17.5 Measuring the seasonal effect

seasonal indexes

Measures of the seasonal effects in time series data.

Seasonal variation may occur within a year or within an even shorter time interval, such as a month, a week or a day. In order to measure the seasonal effect, we construct **seasonal indexes**, which attempt to measure the degree to which the seasons differ from one another. One requirement for this method is that the time series is sufficiently long to allow us to observe several occurrences of each season. For example, if our seasons are the quarters of a year, we need to observe the time series for at least four years. Similarly, if the seasons are the days of the week, our time series should be observed for no less than a month.

17.5a Estimating the seasonal indexes

The seasonal indexes are calculated using the following steps:

- Step 1** Remove the effect of seasonal and random variations by calculating the moving averages. Set the number of periods equal to the number of types of season. For example, we calculate 12-month moving averages if the months of the year represent the seasons. A 5-day moving average is used if the seasons are the days of the working week. If the number of periods in the moving average is even, we calculate centred moving averages. The effect of moving averages is seen in the multiplicative model of time series.

$$y_t = T_t \times C_t \times S_t \times R_t$$

The moving averages remove S_t and R_t leaving

$$MA_t = T_t \times C_t$$

(Note: For an additive model, $y_t = T_t + C_t + S_t + R_t$ and $MA_t = T_t + C_t$)

- Step 2** Calculate the ratio of the time series over the moving averages. Thus, we have

$$\frac{y_t}{MA_t} = \frac{T_t \times C_t \times S_t \times R_t}{T_t \times C_t} = S_t \times R_t$$

The result is a measure of seasonal and random variation.

(Note: For an additive model, $y_t - MA_t = S_t + R_t$)

Step 3 For each type of season, calculate the average of the ratios in Step 2. This procedure removes most (but seldom all) of the random variation (R_i). The resulting average is a measure of the seasonal index (S_i).

Step 4 The seasonal indexes are the average ratios from Step 3 adjusted to ensure that the average seasonal index is one.

(Note: In the case of the additive model, we ensure that this average is equal to zero.)

EXAMPLE 17.6

LO5

Measuring seasonal variation in inbound tourist arrivals to Australia I

XM17-06 The tourism industry is to some extent subject to enormous seasonal variation. The ABS publishes various information on tourism-related variables. The quarterly short-term inbound tourist arrival numbers to Australia for the years 2014(1)–2019(4) are shown in the following table. Calculate the seasonal indexes for each quarter to measure the amount of seasonal variation.

Year	Quarter	Arrivals ('000)	Year	Quarter	Arrivals ('000)
2014	1	1800.9	2017	1	2283.8
	2	1459.4		2	1885.1
	3	1652.3		3	2112.9
	4	2009.7		4	2533.5
2015	1	1952.8	2018	1	2489.5
	2	1524.1		2	1935.9
	3	1774.4		3	2206.3
	4	2198.6		4	2614.1
2016	1	2173.6	2019	1	2494.6
	2	1706.9		2	2029.0
	3	1994.9		3	2274.6
	4	2393.8		4	2667.6

Source: Australian Bureau of Statistics, *Number of movements – Short-term visitors*, cat. no. 3401.0 Overseas Arrivals and Departures, Australia, March 2020.

Solution

Calculating manually

Step 1 As there are four quarters (seasons) per year, we will calculate a four-quarter centred moving average to remove the seasonal and random effects.

To calculate the four-quarter centred moving averages, we first determine the four-quarter moving averages and then calculate the two-period moving averages of these values. So, for example, the moving average that falls between quarters 2 and 3 is

$$\frac{1800.9 + 1459.4 + 1652.3 + 2009.7}{4} = \frac{6922.3}{4} = 1730.6$$

Similarly, the moving average that falls between quarters 3 and 4 is

$$\frac{1459.4 + 1652.3 + 2009.7 + 1952.8}{4} = \frac{7074.2}{4} = 1768.6$$

Therefore, the third-quarter centred moving average is

$$\frac{1730.6 + 1768.6}{2} = \frac{3499.2}{2} = 1749.6$$



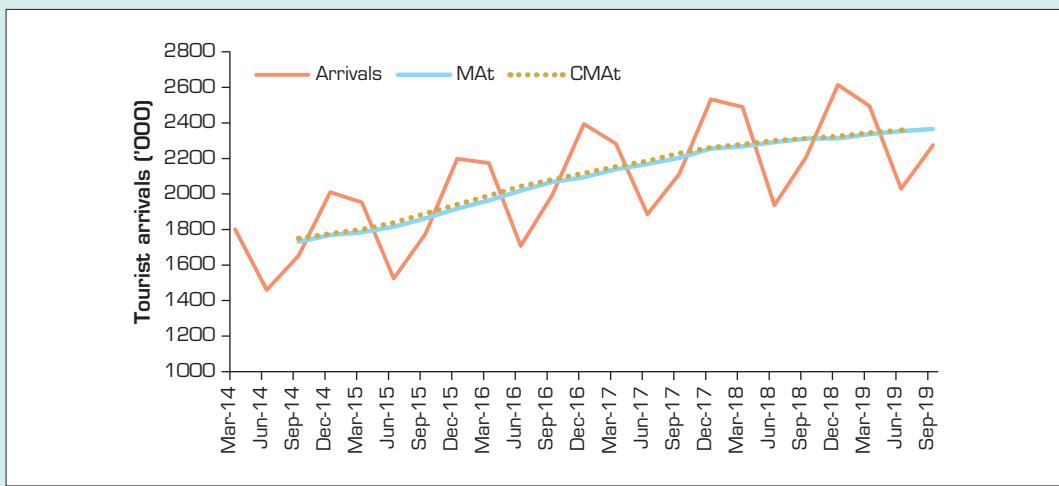
- Step 2** The next step is to find the number of guest arrivals (y_t) divided by the centred moving averages (MA_t) to obtain the value for $S_t \times R_t$.

The outcomes of the first two steps are shown in the last two columns of the following table.

Year	Quarter	Arrivals y_t ('000)	MA_t	CMA_t	Ratio (y_t / CMA_t)
2014	1	1800.9			
	2	1459.4			
	3	1652.3	1730.6	1749.6	0.944
	4	2009.7	1768.6	1776.6	1.131
2015	1	1952.8	1784.7	1800.0	1.085
	2	1524.1	1815.3	1838.9	0.829
	3	1774.4	1862.5	1890.1	0.939
	4	2198.6	1917.7	1940.5	1.133
2016	1	2173.6	1963.4	1990.9	1.092
	2	1706.9	2018.5	2042.9	0.836
	3	1994.9	2067.3	2081.1	0.959
	4	2393.8	2094.9	2117.1	1.131
2017	1	2283.8	2139.4	2154.2	1.060
	2	1885.1	2168.9	2186.4	0.862
	3	2112.9	2203.8	2229.5	0.948
	4	2533.5	2255.3	2261.6	1.120
2018	1	2489.5	2268.0	2279.6	1.092
	2	1935.9	2291.3	2301.4	0.841
	3	2206.3	2311.5	2312.1	0.954
	4	2614.1	2312.7	2324.4	1.125
2019	1	2494.6	2336.0	2344.5	1.064
	2	2029.0	2353.1	2359.8	0.860
	3	2274.6	2366.5		
	4	2667.6			

Figure 17.14 depicts the time series and the moving averages.

FIGURE 17.14 Time series and the moving averages for Example 17.6



Step 3 If we now group the ratios ($S_t \times R_t$) by quarter as in the table below, we can see the similarities within each type of quarter and the differences between different types of quarter. For example, the ratios for quarter 1 are 1.0849, 1.0917, 1.0602, 1.0921 and 1.0640 whereas for quarter 2 they are 0.8288, 0.8355, 0.8622, 0.8412 and 0.8598. By averaging these values for each quarter, we remove most of the random variation.

Step 4 Finally, we adjust or normalise the averages by dividing each average by the total 4.008 and multiplying by 4.000. (Note that the total here is very close to 4, but in general, this need not be the case.) The seasonal indexes are these adjusted averages. These final estimates are $S_1 = 1.0784$, $S_2 = 0.8454$, $S_3 = 0.9486$ and $S_4 = 1.1277$. The following table summarises steps 3 and 4.

Estimates of $S_t \times R_t = y_t / CMA_t$

Year	Quarter				Total
	1	2	3	4	
2014	-	-	0.9444	1.1312	
2015	1.0849	0.8288	0.9388	1.1330	
2016	1.0917	0.8355	0.9586	1.1307	
2017	1.0602	0.8622	0.9477	1.1202	
2018	1.0921	0.8412	0.9542	1.1247	
2019	1.0640	0.8598	-	-	
Average	1.0786	0.8455	0.9487	1.1279	4.0008
Seasonal index (S_t)	1.0784	0.8454	0.9486	1.1277	4.0000

Interpreting the results

The seasonal indexes tell us that, on average, the tourist arrivals to Australia in the second and third quarters are below the annual average, while the tourist arrivals in the first and fourth quarters are above the annual average. That is, we expect the number of tourist arrivals to Australia in the second quarter (June) to be 15.46% [= 100 × (1 – 0.8454)] and the third quarter (September) to be 5.14% [= 100 × (1 – 0.9486)] below the annual level. The numbers for the first (March) and fourth (Dec) quarters are expected to be above the annual level by 7.84% and 12.77% respectively.

17.5b An alternative method to estimate $S_t \times R_t$

One of the drawbacks of the moving averages method is the large number of calculations necessary to solve even a relatively small problem. However, if the time series contains no discernible cyclical component, we can use regression trend analysis instead of moving averages in Step 1. When the time series has no cyclical effect, we can represent the model as

$$y_t = T_t \times S_t \times R_t$$

As the regression line ($\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 t$) represents trend T_t , it follows that

$$\frac{y_t}{T_t} = \frac{y_t}{\hat{y}_t} = S_t \times R_t$$

We can use y_t / \hat{y}_t as an estimate for $S_t \times R_t$ and then average these values to remove the random variation to obtain S_t as we did using moving averages. The alternative method can now be summarised in the following 4 steps.

Step 1 Remove the effect of seasonal and random variation by regression analysis; that is, compute the sample regression line:

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 t$$

Step 2 For each time period, compute the ratio:

$$\frac{y_t}{\hat{y}_t}$$

The ratio removes most of the trend variation.

Step 3 For each type of season, compute the average of the ratios in step 2. This procedure removes most (but seldom all) of the random variation, leaving a measure of seasonality.

Step 4 Adjust the averages in step 3 so that the average of all the seasons is 1 (if necessary).

EXAMPLE 17.7

LO5

Measuring seasonal variation in inbound tourist arrivals to Australia II

For the data in Example 17.6, use the trend estimates to determine the seasonal indexes.

Solution

As the time series tourist arrivals to Australia in our example seems to contain no cyclical pattern, we would expect the indexes based on moving averages to be quite similar to the indexes determined by regression analysis.

We performed a regression analysis with y = number of tourist arrivals to Australia and t = time period, 1, 2, ..., 24. The estimated regression equation is

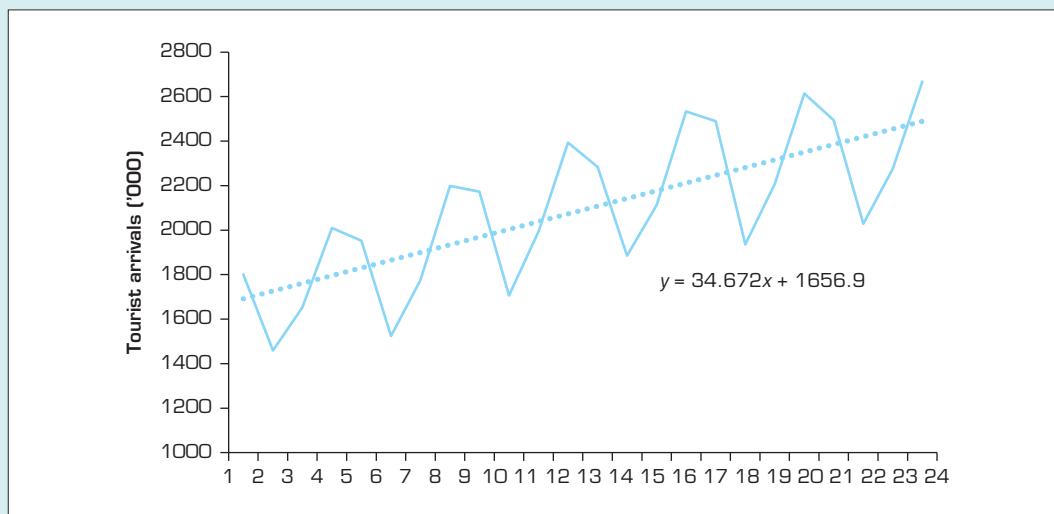
$$\hat{y}_t = 1656.94 + 34.67t$$

Even if the fitness of this equation is good, bear in mind that large deviations between the line and the points are inevitable when either seasonal or cyclical components exist.

Next, for each time period, we computed \hat{y}_t and the ratio y_t / \hat{y}_t . The results are presented in the following table. Figure 17.15 plots the time series and the regression trend line.

Year	Quarter (t)	Arrivals ('000) (y_t)	$\hat{y}_t = 1656.94 + 34.67t$	y_t / \hat{y}_t
2014	1	1800.9	1691.6	1.065
	2	1459.4	1726.3	0.845
	3	1652.3	1761.0	0.938
	4	2009.7	1795.6	1.119
2015	5	1952.8	1830.3	1.067
	6	1524.1	1865.0	0.817
	7	1774.4	1899.6	0.934
	8	2198.6	1934.3	1.137
...	
...	
2018	17	2489.5	2246.4	1.108
	18	1935.9	2281.0	0.849
	19	2206.3	2315.7	0.953
	20	2614.1	2350.4	1.112
2019	21	2494.6	2385.1	1.046
	22	2029.0	2419.7	0.839
	23	2274.6	2454.4	0.927
	24	2667.6	2489.1	1.072



FIGURE 17.15 Time series and trend for Example 17.7

The seasonal indexes are calculated as shown in the following table.

Estimates of $S_t \times R_t = y_t/T_t$

Year	Quarter				Total
	1	2	3	4	
2014	1.0646	0.8454	0.9383	1.1192	
2015	1.0669	0.8172	0.9341	1.1366	
2016	1.1039	0.8519	0.9787	1.1547	
2017	1.0836	0.8799	0.9705	1.1455	
2018	1.1082	0.8487	0.9528	1.1122	
2019	1.0459	0.8385	0.9267	1.0717	
Average	1.0789	0.8469	0.9502	1.1233	3.9993
Seasonal index (S_t)	1.0790	0.8471	0.9503	1.1235	4.0000

As you can see, the two sets of seasonal indexes obtained using moving averages, as well as using a trend line, are very close to each other. Although we applied the two methods to the same data, the actual step 1 technique in a practical application would depend on whether or not we had identified a cyclical pattern in the time series.

At this point, the seasonal indexes only measure the extent to which the seasons vary from one another. In Section 17.8, however, we will show you how the seasonal indexes can play a critical role in forecasting.

17.5c Deseasonalising a time series

One application of seasonal indexes is to remove the seasonal effects of a time series. The process is called **deseasonalising** a time series. The resulting series is called a seasonally adjusted time series. Often this allows the statistician to compare more easily the time series across seasons. For example, the unemployment rate varies according to the season. During the December quarter (October–December), unemployment usually rises, and it falls in the March quarter (January–March). To determine whether unemployment has increased or decreased from the previous month, we frequently ‘deseasonalise’ the data.

deseasonalising

A method of removing seasonal effects from a time series, resulting in a seasonally adjusted series.

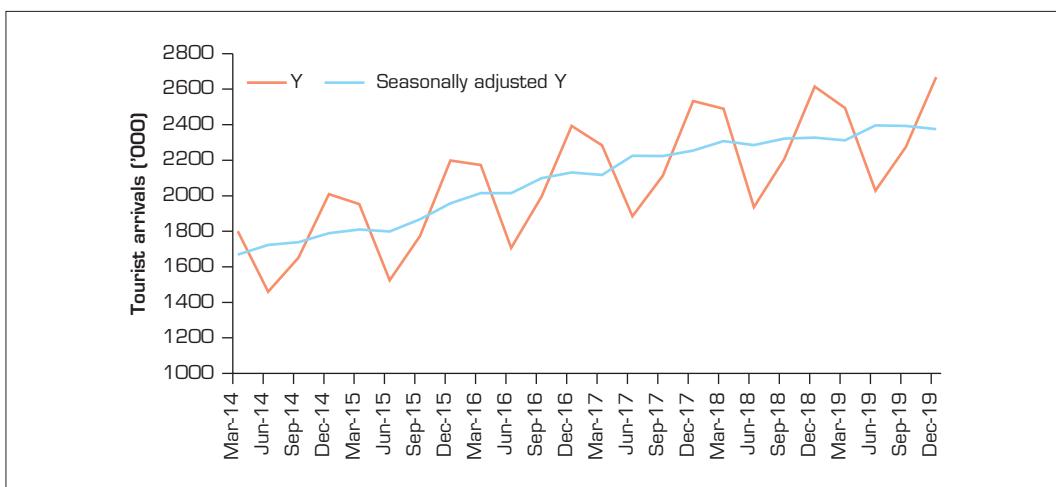
The process is quite simple. For a multiplicative model, after calculating the seasonal indexes, we *divide the original time series by the seasonal indexes*. To illustrate, we have deseasonalised the number of tourist arrivals in Example 17.6 by using the seasonal indexes for each quarter calculated in Example 17.6 using the trend analysis method. The seasonally adjusted rates are shown in the last column of Table 17.1. This is also depicted in Figure 17.16.

By removing the seasonality, we can see there has been a ‘real’ increase in the number of tourist arrivals (compare the arrivals and seasonally adjusted arrivals columns of the table) over the six-year period. This enables the statistics practitioner to examine the factors that produced the increase in numbers.

TABLE 17.1 Number of arrivals and seasonally adjusted arrivals

Year	Quarter	Arrivals ('000) (y_t)	Seasonal index (s_i)	Seasonally adjusted arrivals ('000) (y_t/s_i)
2014	1	1800.9	1.0790	1669.0
	2	1459.4	0.8471	1722.8
	3	1652.3	0.9503	1738.6
	4	2009.7	1.1235	1788.7
2015	1	1952.8	1.0790	1809.7
	2	1524.1	0.8471	1799.2
	3	1774.4	0.9503	1867.1
	4	2198.6	1.1235	1956.9
...	
...	
2018	1	2489.5	1.0790	2307.1
	2	1935.9	0.8471	2285.4
	3	2206.3	0.9503	2321.6
	4	2614.1	1.1235	2326.7
2019	1	2494.6	1.0790	2311.9
	2	2029.0	0.8471	2395.3
	3	2274.6	0.9503	2393.4
	4	2667.6	1.1235	2374.3

FIGURE 17.16 Time series and seasonally adjusted time series for Example 17.7



In Section 17.8, we will present a forecasting technique that uses the seasonal indexes in another way.

EXERCISES

The following exercises require the use of a computer and software.

Learning the techniques

- 17.22 a XR17-22** For the following time series, calculate the five-day moving averages to remove the seasonal and random variation.

Day	Week			
	1	2	3	4
Monday	12	11	14	17
Tuesday	18	17	16	21
Wednesday	16	19	16	20
Thursday	25	24	28	24
Friday	31	27	25	32

- b** Calculate the seasonal (daily) indexes and seasonally adjust the series.

- 17.23 XR17-23** Using the data in Exercise 17.22, the regression trend line is

$$\hat{y}_t = 16.8 + 0.366t \quad (t = 1, 2, \dots, 20)$$

Calculate the seasonal indexes based on the regression trend line and deseasonalise the series.

- 17.24 XR17-24** Given the following time series, calculate the seasonal (quarterly) indexes using a four-quarter centred moving average and seasonally adjust the series.

Quarter	Year				
	1	2	3	4	5
1	50	41	43	36	30
2	44	38	39	32	25
3	46	37	39	30	24
4	39	30	35	25	22

- 17.25 XR17-25** The regression trend line for the data in Exercise 17.24 is

$$\hat{y}_t = 47.73 - 1.19t \quad (t = 1, 2, \dots, 20)$$

Calculate the seasonal indexes based on this trend line and deseasonalise the series.

- 17.26 XR17-26** University enrolments have increased sharply in Australia since the early 1990s. To help forecast future enrolments, an economist recorded the total Australian university enrolments from 1990 to 2017. These data (in thousands) are listed:

Year	Enrolment	Year	Enrolment	Year	Enrolment
1990	485.1	2000	695.5	2010	1192.7
1991	534.1	2001	842.2	2011	1221.0
1992	559.4	2002	896.6	2012	1257.7
1993	575.6	2003	930.0	2013	1313.8
1994	585.4	2004	945.0	2014	1373.2
1995	604.2	2005	957.2	2015	1410.1
1996	634.1	2006	984.1	2016	1457.2
1997	658.8	2007	1029.8	2017	1513.4
1998	671.9	2008	1066.1		
1999	686.3	2009	1134.9		

Source: Australian Bureau of Statistics, *Year Book Australia*, cat. no. 1301.0, CC BY 2.5 AU <http://creativecommons.org/licenses/by/2.5/au/legalcode>

- a** Plot the time series.
b Use regression analysis to determine the trend.

- 17.27 XR17-27** Foreign trade is important to Australia. To measure the extent of the trade imbalance, an economist recorded the difference between total exports and imports (in billions of dollars) for the years 2010 to 2019.

- a** Plot the trade balance.
b Apply regression analysis to measure the trend.

Applying the techniques

- 17.28 XR17-28 Self-correcting exercise.** The quarterly earnings (in millions of dollars) of a large textile manufacturing company have been recorded for the years 2016–19. These data are shown in the following table. Using an appropriate moving average, measure the quarterly variation by calculating the seasonal (quarterly) indexes. Obtain the seasonally adjusted earnings.

Quarter	Year			
	2016	2017	2018	2019
1	52	57	60	66
2	67	75	77	82
3	85	90	94	98
4	54	61	63	67

- 17.29** The linear trend line calculated by regression analysis for the data in Exercise 17.28 is

$$\hat{y}_t = 61.75 + 1.18t \quad (t = 1, 2, \dots, 16)$$

Calculate the seasonal indexes and the deseasonalised series, using this trend line.

17.30 XR17-30 The quarterly overseas arrivals to Australia during 2014(1)–2019(3) are recorded. (Source: Australian Bureau of Statistics, cat. no. 3401.0 – *Overseas Arrivals and Departures, Australia*, Nov 2019.)

- a Plot the time series.
- b Calculate the four-quarter centred moving averages.
- c Plot the moving averages.
- d Calculate the seasonal (quarterly) indexes, using the time series and moving averages in part (b). Calculate the seasonally adjusted values.

17.31 The linear trend line (from regression analysis) for the data in Exercise 17.30 is

$$\hat{y}_t = 4035.3 + 64.59t \quad (t = 1, 2, \dots, 23)$$

Calculate the seasonal indexes, using this trend line.

17.32 XR17-32 The owner of a pizza shop in a Brisbane suburb wants to forecast the number of pizzas she will sell each day. She records the number sold daily for four weeks. These data are also shown in the following table. Calculate the seasonal (daily) indexes, using a seven-day moving average. Seasonally adjust the number of pizzas sold.

Day	Week			
	1	2	3	4
Sunday	240	221	235	219
Monday	85	80	86	91

Day	Week			
	1	2	3	4
Tuesday	93	75	74	102
Wednesday	106	121	100	89
Thursday	125	110	117	105
Friday	188	202	205	192
Saturday	314	386	402	377

17.33 XR17-33 A manufacturer of ski equipment is in the process of reviewing his accounts receivable. He has noticed that there appears to be a seasonal pattern: accounts receivable increase in the winter and decrease during the summer. The quarterly accounts receivable (\$million) for the years 2017–20 are shown in the accompanying table. To measure the seasonal variation, calculate the seasonal (quarterly) indexes based on the regression trend line, which was calculated as

$$\hat{y}_t = 90.4 + 2.02t \quad (t = 1, 2, \dots, 16)$$

Seasonally adjust the accounts receivable.

Quarter	Year			
	2017	2018	2019	2020
1	106	115	114	121
2	92	100	105	111
3	65	73	79	82
4	121	135	140	163

17.6 Introduction to forecasting

As we have noted before, many different forecasting methods are available. One of the factors we consider in making our choice is the type of component that makes up the time series we are attempting to forecast. Even then, however, we have a variety of techniques from which to choose. One way of deciding which method to use is to select the technique that results in the greatest forecast accuracy. The two most commonly used measures of forecast accuracy are the **mean absolute deviation (MAD)** and the **sum of squares for forecast error (SSFE)**. These are defined as follows:

Mean absolute deviation

$$\text{MAD} = \frac{\sum_{t=1}^n |y_t - F_t|}{n}$$

where

y_t = actual value of the time series at time t

F_t = forecast value of the time series at time t

n = number of forecast periods

Sum of squares for forecast error

$$\text{SSFE} = \sum_{t=1}^n (y_t - F_t)^2$$

The MAD averages the absolute differences between the actual values and the forecast values; in contrast, the SSFE squares these differences. Which measure to use in judging forecast accuracy depends on the circumstances. If avoiding large errors is extremely important, SSFE should be used, since it penalises large deviations more heavily than does the MAD. Otherwise, use MAD.

It is probably best to use some of the observations of the time series to develop several competing forecasting models and then forecast for the remaining time periods. We can then calculate either MAD or SSFE for the latter period. For example, if we have five years of monthly observations, we can use the first four years to develop the forecasting techniques and then use them to forecast the fifth year. As we know the actual values in the fifth year, we can choose the technique that results in the most accurate forecasts.

EXAMPLE 17.8

LO6

Comparing forecasting models

XM17-08 Three different forecasting models were used to forecast the time series for 2017–20. The forecast values and the actual values for these years are shown in the following table. Use MAD and SSFE to determine which models performed best.

Year	Actual value of y_t	Forecast value (F) using model		
		1	2	3
2017	129	136	118	130
2018	142	148	141	146
2019	156	150	158	170
2020	183	175	163	180

Solution

The measures of forecast accuracy for Model 1 are

$$\begin{aligned}\text{MAD} &= \frac{|129 - 136| + |142 - 148| + |156 - 150| + |183 - 175|}{4} \\ &= \frac{7 + 6 + 6 + 8}{4} = \frac{27}{4} = 6.75\end{aligned}$$

$$\begin{aligned}\text{SSFE} &= (129 - 136)^2 + (142 - 148)^2 + (156 - 150)^2 + (183 - 175)^2 \\ &= 49 + 36 + 36 + 64 = 185\end{aligned}$$

For Model 2, we have

$$\begin{aligned}\text{MAD} &= \frac{|129 - 118| + |142 - 141| + |156 - 158| + |183 - 163|}{4} \\ &= \frac{11 + 1 + 2 + 20}{4} = \frac{34}{4} = 8.5\end{aligned}$$

$$\begin{aligned}\text{SSFE} &= (129 - 118)^2 + (142 - 141)^2 + (156 - 158)^2 + (183 - 163)^2 \\ &= 121 + 1 + 4 + 400 = 526\end{aligned}$$



For Model 3, we have

$$\begin{aligned} \text{MAD} &= \frac{|129 - 130| + |142 - 146| + |156 - 170| + |183 - 180|}{4} \\ &= \frac{1+4+14+3}{4} = \frac{22}{4} = 5.5 \end{aligned}$$

$$\begin{aligned} \text{SSFE} &= (129 - 130)^2 + (142 - 146)^2 + (156 - 170)^2 + (183 - 180)^2 \\ &= 1+16+196+9 = 222 \end{aligned}$$

Model	MAD	SSFE
Model 1	6.75	185
Model 2	8.50	526
Model 3	5.50	222

Model 2 is inferior to both Models 1 and 3, no matter how forecast accuracy is measured (either by using MAD or SSFE). Using MAD, Model 3 is best; but using SSFE, Model 1 is the most accurate. The choice between Model 1 and Model 3 should be made on the basis of whether we prefer a model that consistently produces moderately accurate forecasts (Model 1) or one whose forecasts come quite close to most actual values but miss badly in a small number of time periods (Model 3).

EXERCISES

Learning the techniques

- 17.34 XR17-34** The actual values and forecast values of a time series are shown in the following table. Calculate MAD and SSFE.

Period t	Actual value y_t	Forecast value F_t
1	166	173
2	179	186
3	195	192
4	214	211
5	220	223

- 17.35 XR17-35** Calculate MAD and SSFE for the forecasts below.

Period t	Actual value y_t	Forecast value F_t
1	5.7	6.3
2	6.0	7.2
3	7.0	8.6
4	7.5	7.1
5	7.0	6.0

Applying the techniques

- 17.36 XR17-36 Self-correcting exercise.** Two forecasting models were used to predict the future values of a

time series. These are shown below, together with the actual values. For each model, calculate MAD and SSFE to determine which was more accurate.

Period t	Actual value y_t	Forecast value F_t	
		Model 1	Model 2
1	6.0	7.5	6.3
2	6.6	6.3	6.7
3	7.3	5.4	7.1
4	9.4	8.2	7.5

- 17.37 XR17-37** Three forecasting techniques were used to predict the values of a time series. These are given in the following table. For each, calculate MAD and SSFE to determine which is most accurate.

Actual value y_t	Forecast value F_t		
	Technique 1	Technique 2	Technique 3
19	21	22	17
24	27	24	20
28	29	26	25
32	31	28	31
38	35	30	39

17.7 Time series forecasting with exponential smoothing

In Section 17.2 we presented smoothing techniques whose function is to reduce random fluctuation, enabling us to identify the time series components. One of these methods, exponential smoothing, can also be used in forecasting. Recall the exponential smoothing formula

$$ES_1 = y_1 \quad \text{and} \quad ES_t = \omega y_t + (1 - \omega)ES_{t-1} \quad \text{for } t \geq 2$$

for which the choice of the smoothing constant ω determines the degree of smoothing. A value of ω close to 1 results in very little smoothing, whereas a value of ω close to 0 results in a great deal of smoothing.

When a time series exhibits a gradual trend and no evidence of cyclical or seasonal effects, exponential smoothing can be a useful way of forecasting. Suppose that t represents the current time period and we have calculated the smoothed value ES_t . This value is then the forecast value at time $t + 1$. That is,

$$F_{t+1} = ES_t$$

If we wish, we can forecast two, three or any number of time periods into the future.

$$F_{t+2} = ES_t$$

$$F_{t+3} = ES_t$$

It must be understood that the accuracy of the time series forecast decreases rapidly for predictions of more than one period into the future. However, as long as we are dealing with a time series that possesses no cyclical or seasonal effects, we can produce reasonably accurate forecasts for the next period.

EXAMPLE 17.9

LO7

Forecasting wine consumption

XM17-09 The annual Australian consumption of wine per person (in litres) for the years 1945 to 2018 was recorded and part of the data is listed in the accompanying table. Use exponential smoothing and data for 1945 to 2017 to forecast the 2018 wine consumption per person.

Year	Wine consumption (litres/capita)
1945	7.07
1946	5.75
1947	6.59
.	...
2015	28.87
2016	28.68
2017	28.35
2018	28.30

Source: Australian Bureau of Statistics, *Apparent Consumption of Alcohol*, Australia, various issues, ABS, Canberra.





Solution

Calculating manually

A plot of the time series (not shown here) reveals a gradual increase in per capita wine consumption in Australia. There is no cyclical pattern. As a consequence, exponential smoothing is an appropriate forecasting method. There does not appear to be a great deal of random variation. As a consequence, we choose $\omega = 0.8$, which results in very little smoothing.

The graphs of the time series and of the smoothed values will therefore be almost coincident, as you can see from the computer output table and plot below. The smoothed values (shown in the computer output) are obtained from the formulas

$$ES_1 = y_1$$

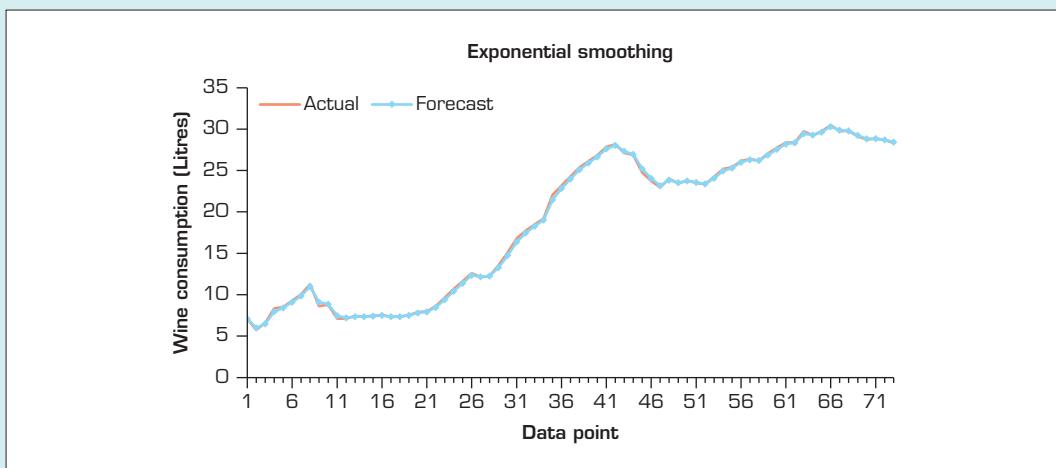
and

$$ES_t = \omega y_t + (1 - \omega)ES_{t-1} \quad (\text{with } \omega = 0.8) \text{ for } t = 2, 3, \dots, 73$$

Excel output for Example 17.9

Wine consumption per person

	A	B	C	D	E
1	Wine consumption per person				
2			Actual	Forecast: $ES_t = \omega Y_t + (1-\omega) ES_{t-1}$	
3	Year	Period t	Y_t	ES_t (Manual)	ES_t (EXCEL)
4	1945	1	7.07	7.07	7.07
5	1946	2	5.75	6.01	6.01
6	1947	3	6.59	6.47	6.47
7	1948	4	8.34	7.97	7.97
8
9	1981	35	22.07	21.46	21.46
10	1982	36	23.18	22.84	22.84
11	1983	37	24.29	24.00	24.00
12
13	2014	70	28.73	28.84	28.84
14	2015	71	28.87	28.86	28.86
15	2016	72	28.68	28.72	28.72
16	2017	73	28.35	28.42	28.42



From the last column of the output, we have $ES_{73} = 28.42$. As $F_{t+1} = ES_t$, the forecast for 2018 ($t = 74$) is given by

$$F_{2018} = F_{74} = ES_{73} = 28.42$$

This forecast of 28.42 litres is very close to the actual wine consumption of 28.3 litres.

In Section 17.2, we presented exponential smoothing as a process to detect the components of a time series, and we pointed out that Excel places the smoothed values in the next period. As you can see, Excel uses exponential smoothing exclusively as a forecasting technique. Thus, each smoothed value represents the forecast for the next period.

EXERCISES

Learning the techniques

- 17.38 XR17-38** Use the exponential smoothing technique with $\omega = 0.3$, to forecast the value of the following time series at time $t = 8$.

Period t	y_t
1	12
2	20
3	16
4	19
5	15
6	11
7	14

- 17.39 XR17-39** Use the following time series for the years 2010–15 to develop forecasts for 2016–19, with:
- a $\omega = 0.3$
 - b $\omega = 0.6$
 - c $\omega = 0.7$

Year	y_t
2010	110
2011	103
2012	111
2013	117
2014	126
2015	115

Applying the techniques

The following exercises require the use of a computer and software.

- 17.40 XR17-40 Self-correcting exercise.** The data file contains the time series values of the rate of inflation for Australia from 1957 to 2019. Use exponential smoothing with $\omega = 0.7$ to forecast the rate of inflation for Australia in 2020.
- 17.41 XR17-41** The meat industry is one of the major exporting sectors in Australia. The annual South Australian production of red meat is recorded for the years 1973–2019. Use exponential smoothing with $\omega = 0.7$ to predict Australian meat production for 2020.

17.8 Time series forecasting with regression

Regression analysis has been applied to various problems. It was used in Chapter 15 to analyse the relationship between two variables, and it was used in Chapter 16 to analyse how a dependent variable is influenced by a group of independent variables. In those chapters, regression analysis was used to predict the value of the dependent variable. In this section we again want to use regression techniques, but now the independent variable or variables will be measures of time. The simplest application would be to a time series in which the only component (in addition to random variation, which is always present) is a *linear trend*. In that case, the model

$$y_t = \beta_0 + \beta_1 t + \varepsilon$$

would be likely to provide excellent forecasts. However, we can take this basic model and augment it so that it can be used in other situations.

17.8a Forecasting time series with trend and seasonality

There are two ways to use regression analysis to forecast time series whose components are *trend* and *seasonal effect*. The first involves using the seasonal indexes developed in Section 17.5. The second involves using indicator variables (or *dummy variables*).

17.8b Forecasting with seasonal indexes

The seasonal indexes measure the season-to-season variation. If we combine these indexes with a forecast of the trend, we produce the following formula.

Forecast of trend and seasonality

$$F_t = (\hat{\beta}_0 + \hat{\beta}_1 t) S_{it}$$

where

F_t = forecast of the time series for period t

S_{it} = seasonal index for period t .

The process that we use to forecast with seasonal indexes is as follows.

- 1 Use simple linear regression to find the trend line.

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 t$$

- 2 Use the trend line to calculate the seasonal indexes.

- 3 For the future time period T , find the trend value.

$$\hat{y}_T = \hat{\beta}_0 + \hat{\beta}_1 T$$

- 4 Multiply the trend value \hat{y}_T by the seasonal index for the season to forecast:

$$F_t = \hat{y}_T \cdot S_{iT}$$

EXAMPLE 17.10

LO8

Forecasting the inbound tourist arrivals to Australia III

Recall that in Example 17.7 we calculated the seasonal (quarterly) indexes for the inbound tourist arrivals to Australia. Use these indexes to forecast the number of inbound tourist arrivals in Australia during the four quarters in 2020.

Solution

The trend line that was calculated from the quarterly data for 2014(Q_1) to 2019(Q_4) ($t = 1, \dots, 24$) is

$$\hat{y}_t = 1656.94 + 34.67t$$

For the four quarters 1, 2, 3 and 4 of 2020, $t = 25, 26, 27$ and 28 , so we find the following trend forecast values:

Year	Quarter	Period t	$\hat{y}_t = 1656.9 + 34.67t$
2020	1	25	2523.8
	2	26	2558.4
	3	27	2593.1
	4	28	2627.8



We now multiply the trend forecast values by the seasonal indexes (calculated from the trend line method in Example 17.7 rather than from the moving averages method) to obtain the seasonalised forecasts. Thus, the seasonalised forecasts are as follows:

Year	Quarter	Trend value \hat{y}_t	Seasonal index S_t	Forecast $F_t = \hat{y}_t \times S_t$
2020	1	2523.8	1.0790	2723.1
	2	2558.4	0.8471	2167.2
	3	2593.1	0.9503	2464.2
	4	2627.8	1.1235	2952.3

We forecast that the number of tourist arrivals ('000) during the four quarters, 1, 2, 3 and 4, of 2020 are 2723.1, 2167.2, 2464.2 and 2952.3 respectively.

Now we will present the solution for the opening example to this chapter.

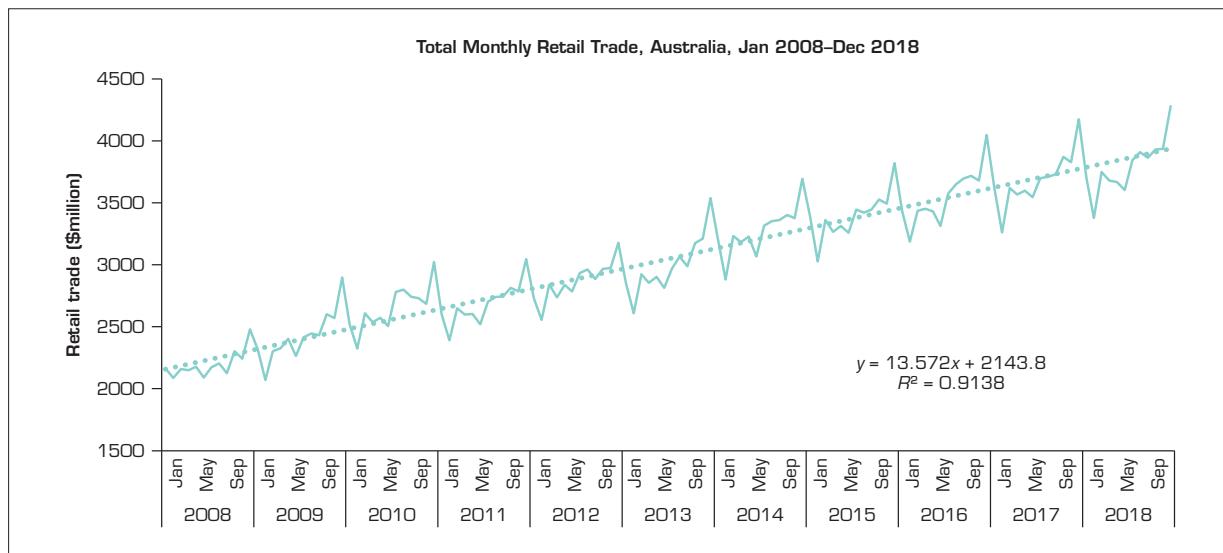
SPOTLIGHT ON STATISTICS

Retail turnover of Australian food services sector: Solution

A preliminary examination of the monthly retail turnover data for the period January 2008 to December 2018 reveals an upward trend over the 11-year period. Moreover, the retail turnover varies by month. We investigated a linear trend versus quadratic and cubic trends. The linear trend fits the data better than a quadratic or cubic trend.



Source: Shutterstock.com/
asr nasib



Using the computer

With retail turnover as the dependent variable and the period as the independent variable, Excel yielded the following regression results:

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.9559					
5	R Square	0.9138					
6	Adjusted R Square	0.9132					
7	Standard Error	160.01					
8	Observations	132					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	35302062.5	35302062	1378.8	0.000	
13	Residual	130	3328497.3	25603.8			
14	Total	131	38630559.8				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	2143.8	28.014	76.526	0.000	2088.3	2199.2
18	t	13.572	0.366	37.132	0.000	12.849	14.295

The estimated linear regression line is given by

$$\hat{y}_t = 2143.8 + 13.57t \quad (t = 1, 2, \dots, 132)$$

The seasonal indexes were calculated as follows. (For instructions, see Example 17.7.)

Excel output

Year	Period (t)	Retail trade (\$million) (y_t)	$\hat{y}_t = 2143.8 + 13.57t$	y_t / \hat{y}_t
2008	1	2163.5	2157.3	1.0029
	2	2086.7	2170.9	0.9612
	3	2159.4	2184.5	0.9885
	4	2148.2	2198.1	0.9773
	5	2176.7	2211.6	0.9842
	6	2089.9	2225.2	0.9392
	7	2172.6	2238.8	0.9704
	8	2204.0	2252.3	0.9785

2018	121	3698.8	3786.0	0.9770
	122	3377.8	3799.5	0.8890
	123	3749.1	3813.1	0.9832
	124	3679.3	3826.7	0.9615
	125	3666.6	3840.3	0.9548
	126	3601.3	3853.8	0.9345
	127	3844.0	3867.4	0.9939
	128	3908.3	3881.0	1.0070
	129	3863.4	3894.5	0.9920
	130	3929.1	3908.1	1.0054
	131	3934.2	3921.7	1.0032
	132	4278.9	3935.3	1.0873

The monthly seasonal indexes are calculated as shown in the following table.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1														
2	Year	1	2	3	4	5	6	7	8	9	10	11	12	
3	2008	1.0029	0.9612	0.9885	0.9773	0.9842	0.9392	0.9704	0.9785	0.9375	1.0101	0.9781	1.0749	
4	2009	0.9992	0.8868	0.9812	0.9856	1.0115	0.9476	1.0056	1.0128	1.0008	1.0654	1.0464	1.1732	
5	2010	1.0154	0.9308	1.0388	1.0048	1.0134	0.9820	1.0845	1.0861	1.0583	1.0480	1.0251	1.1477	
6	2011	0.9820	0.8992	0.9910	0.9673	0.9644	0.9284	0.9907	0.9995	0.9963	1.0167	1.0009	1.0899	
7	2012	0.9718	0.9056	1.0012	0.9607	0.9908	0.9681	1.0145	1.0203	0.9892	1.0122	1.0097	1.0742	
8	2013	0.9589	0.8735	0.9751	0.9473	0.9589	0.9255	0.9723	0.9995	0.9703	1.0264	1.0333	1.1332	
9	2014	1.0238	0.9148	1.0222	1.0022	1.0124	0.9579	1.0310	1.0375	1.0362	1.0445	1.0320	1.1245	
10	2015	1.0285	0.9144	1.0112	0.9786	0.9888	0.9680	1.0197	1.0085	1.0113	1.0311	1.0170	1.1083	
11	2016	0.9918	0.9174	0.9850	0.9859	0.9762	0.9393	1.0088	1.0259	1.0357	1.0375	1.0230	1.1213	
12	2017	0.9995	0.8966	0.9914	0.9736	0.9786	0.9602	0.9983	0.9981	0.9995	1.0336	1.0184	1.1067	
13	2018	0.9770	0.8890	0.9832	0.9615	0.9548	0.9345	0.9939	1.0070	0.9920	1.0054	1.0032	1.0873	
14	Average	0.9955	0.9081	0.9972	0.9768	0.9849	0.9501	1.0082	1.0158	1.0025	1.0301	1.0170	1.1128	11.9989
15	SI	0.9956	0.9082	0.9973	0.9769	0.9850	0.9501	1.0083	1.0159	1.0026	1.0302	1.0171	1.1129	12.0000

The regression equation was employed again to predict the retail turnover for the 12 months of year 2019, based on the linear trend.

$$\hat{y}_t = 2143.8 + 13.57t \quad (t = 133, 134, \dots, 144)$$

These values were multiplied by the seasonal index, which resulted in the following forecasts:

Month-year	Period <i>t</i>	Actual(<i>Y_t</i>)	Trend forecast (<i>ŷ_t</i>)	Seasonal index (<i>S_i</i>)	Forecast <i>F_t</i> = <i>ŷ_t</i> × <i>S_i</i>	Forecast error <i>FE_t</i> = (<i>ŷ_t</i> - <i>F_t</i>)
Jan-19	133	3826.7	3948.8	0.9956	3931.5	-104.8
Feb-19	134	3456.0	3962.4	0.9082	3598.7	-142.7
Mar-19	135	3897.1	3976.0	0.9973	3965.1	-68.0
Apr-19	136	3808.2	3989.6	0.9769	3897.3	-89.1
May-19	137	3829.0	4003.1	0.9850	3943.0	-114.0
Jun-19	138	3706.0	4016.7	0.9501	3816.4	-110.4
Jul-19	139	3903.5	4030.3	1.0083	4063.6	-160.1
Aug-19	140	3948.4	4043.8	1.0159	4108.1	-159.7
Sep-19	141	3926.6	4057.4	1.0026	4067.8	-141.2
Oct-19	142	4043.2	4071.0	1.0302	4193.8	-150.6
Nov-19	143	4067.3	4084.6	1.0171	4154.4	-87.1
Dec-19	144	4389.3	4098.1	1.1129	4561.0	-171.7

We calculated the MAD and SSFE measures to assess the quality of the forecasts. Based on the actual and forecast values presented for the 12 months of 2019 in the above table, $MAD = 124.9$ and $\sqrt{SSFE} = 21.1$. The MAD and \sqrt{SSFE} measures are relatively small compared to the values of y_t . Therefore, the quality of the forecast is acceptable.

17.8c Forecasting seasonal time series with dummy variables

As an alternative to calculating and using seasonal indexes to measure the seasonal variations, we can use dummy variables (or indicator variables). For example, if the seasons are the quarters of a year, the multiple regression model

$$y_t = \beta_0 + \beta_1 t + \beta_2 Q_1 + \beta_3 Q_2 + \beta_4 Q_3 + \varepsilon$$

can be used where

$$\begin{aligned} t &= \text{time period} \\ Q_1 &= 1 (\text{if quarter 1}) \\ &= 0 (\text{if not}) \\ Q_2 &= 1 (\text{if quarter 2}) \\ &= 0 (\text{if not}) \\ Q_3 &= 1 (\text{if quarter 3}) \\ &= 0 (\text{if not}) \end{aligned}$$

Thus, for each time period, the *dummy variables* Q_1 , Q_2 and Q_3 would be used to represent the quarters. The coefficients β_0 , β_1 , β_2 , β_3 and β_4 would be estimated in the usual way, and the regression equation would be used to predict future values.

Notice that for the fourth quarter, the three dummy variables are all equal to zero, and we are simply left with

$$y_t = \beta_0 + \beta_1 t + \varepsilon$$

EXAMPLE 17.11

LO8

Forecasting the number of tourist arrivals in Australia IV

XM17-11 Use dummy (indicator) variables and regression analysis to forecast the number of tourist arrivals in Australia in 2020 (Q_1 , Q_2 , Q_3 , Q_4) using the data in Example 17.6 for the period 2014(1) to 2019(4). (The data for y , t , Q_1 , Q_2 and Q_3 are stored in columns 1 to 5 respectively.)

Solution

Because the seasons are the quarters of the year, we begin by creating dummy variables Q_1 , Q_2 and Q_3 as described earlier. Thus,

- $Q_1 = 1$, $Q_2 = 0$, $Q_3 = 0$ represents an observation in the first quarter
- $Q_1 = 0$, $Q_2 = 1$, $Q_3 = 0$ represents an observation in the second quarter
- $Q_1 = 0$, $Q_2 = 0$, $Q_3 = 1$ represents an observation in the third quarter
- $Q_1 = 0$, $Q_2 = 0$, $Q_3 = 0$ represents an observation in the fourth quarter.

The model to be estimated is

$$y_t = \beta_0 + \beta_1 t + \beta_2 Q_1 + \beta_3 Q_2 + \beta_4 Q_3 + \varepsilon$$

Year	Quarter	Arrivals ('000)	Period, t	Q_1	Q_2	Q_3
2014	1	1800.9	1	1	0	0
	2	1459.4	2	0	1	0
	3	1652.3	3	0	0	1
	4	2009.7	4	0	0	0
2015	1	1952.8	5	1	0	0
	2	1524.1	6	0	1	0





Year	Quarter	Arrivals ('000)	Period, t	Q_1	Q_2	Q_3
	3	1774.4	7	0	0	1
	4	2198.6	8	0	0	0
...
...
2018	1	2489.5	17	1	0	0
	2	1935.9	18	0	1	0
	3	2206.3	19	0	0	1
	4	2614.1	20	0	0	0
2019	1	2494.6	21	1	0	0
	2	2029.0	22	0	1	0
	3	2274.6	23	0	0	1
	4	2667.6	24	0	0	0

Using the data, the following computer regression output was produced.

Excel output for Example 17.11

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.9897					
5	R Square	0.9795					
6	Adjusted R Square	0.9752					
7	Standard Error	53.640					
8	Observations	24					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	4	2613585	653396	227.1	0.000	
13	Residual	19	54668	2877			
14	Total	23	2668252				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	1936.601	31.354	61.766	0.000	1870.977	2002.225
18	Period, t	33.306	1.603	20.780	0.000	29.951	36.661
19	Q1	-103.766	31.340	-3.311	0.004	-169.361	-38.170
20	Q2	-579.538	31.134	-18.614	0.000	-644.703	-514.373
21	Q3	-367.011	31.010	-11.835	0.000	-431.916	-302.105

Interpreting the results

The values $F = 227.1$ ($p\text{-value} = 0$) and adjusted $R^2 = 0.975$ indicate that the model's fit to the data is very good. The t -values for Q_1 , Q_2 , and Q_3 are -3.31 , -18.61 and -11.84 respectively, with the corresponding p -values 0.004 , 0.000 and 0.000 respectively. These values provide enough evidence to allow us to conclude that the number of tourist arrivals in Australia in quarters 1, 2 and 3 differs significantly from that in quarter 4.

The model's fitness appears to be reasonable, so we can use it to forecast the 2020 quarterly inbound tourist arrivals as shown in the accompanying table.

Quarter	Period t	Q_1	Q_2	Q_3	Forecast	
					$\hat{y}_t = 1936.6 + 33.3t - 103.8Q_1 - 579.5Q_2 - 367.0Q_3$	
2020 1	25	1	0	0		2665.5
	26	0	1	0		2223.0
	27	0	0	1		2468.8
	28	0	0	0		2869.2

These forecasts are quite similar to those produced by using seasonal indexes in Example 17.10.

17.8d Autoregressive model

Recall that one of the requirements for the use of regression is that the errors be independent of one another. In Chapter 16 we developed a test for first-order autocorrelation, called the Durbin–Watson test. Although the existence of strong autocorrelation tends to destroy the validity of regression analysis, it also provides an opportunity to produce inaccurate forecasts. If we believe that there is a correlation between consecutive residuals, then the **autoregressive model**

autoregressive model

A model based on the belief that there is correlation between consecutive residuals.

$$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon$$

can be helpful in forecasting future values of y . The model specifies that consecutive values of the time series are correlated (which follows from the conclusion that the residuals are correlated). We estimate β_0 and β_1 by least squares and then use the equation to forecast the values of y .

EXAMPLE 17.12

LO8

Forecasting the Consumer Price Index

XM17-12 The Consumer Price Index (CPI) is used as a general measure of inflation. It is an important measure because it often influences governments to take some corrective action when the rate of inflation is high. The table that follows lists the partial data for percentage annual increase in the Australian CPI from 1957 until 2019.

$$\hat{y}_t = 6.474 - 0.056t$$

Annual percentage increase in the CPI, Australia, 1957–2019

Year	Increase in CPI (%)	Year	Increase in CPI (%)	Year	Increase in CPI (%)
1957	5.6
1958	1.2	1981	9.4	2011	3.4
1959	1.5	1982	10.4	2012	2.4
1960	2.6	1983	11.5	2013	2.3
1961	4.0	1984	6.9	2014	2.6
1962	0.4	1985	4.3	2015	1.5
1963	0.4	1986	8.4	2016	1.3
1964	0.7	1987	9.3	2017	1.9
1965	3.8	1988	7.3	2018	1.9
...	2019	1.6

Source: Australian Bureau of Statistics, January 2020, *Consumer Price Index*, cat. no. 6401.0, ABS, Canberra.

In an attempt to forecast this value on the basis of time period alone, we estimated the model as (see output below):

	B	C	D	E	F
10	Regression of variable % change in CPI:				
11	Goodness of fit statistics (% change in CPI):				
12	Observations	63			
13	DF	61			
14	R ²	0.072			
15	Adjusted R ²	0.057			
16	MSE	13.684			
17	RMSE	3.699			
18	DW	0.317			
19					
20	Model parameters (% change in CPI):				
21	Source	Value	Standard error	t	Pr > t
22	Intercept	6.474	0.943	6.864	< 0.0001
23	Period	-0.056	0.026	-2.183	0.033

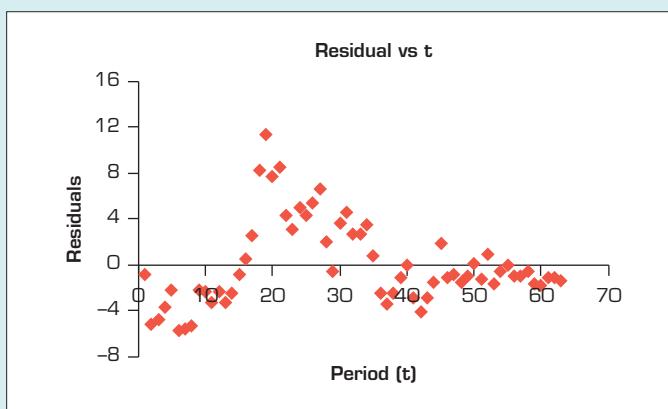
We found that $R^2 = 0.072$ (indicating a poor model), but, more importantly, we also found the Durbin–Watson statistic to be $d = 0.317$ (where using the DW critical values in Table 12(a) of Appendix B for $k = 1$ and $n = 63$,

$d_L = 1.57$ and $d_U = 1.63$), which indicates strong positive autocorrelation. Because of this condition, an autoregressive model appears to be more appropriate.

We now estimate the autoregressive model and forecast the increase in the CPI for 2020. (File **XM17-12** stores the CPI increase for years 1957 to 2019 in column C. Column D contains the lagged value of the CPI increase for years 1958 to 2019.)

Solution

Consider the residual plot from the estimated regression equation, $\hat{y}_t = 6.474 - 0.056t$.



As can be seen from the residual plot, the sign of consecutive residuals remains the same mostly for some time before the sign change, confirming positive autocorrelation (see Section 15.7d, pages 671–2). To remedy this, we consider a first-order autoregressive model.

Therefore, the model to be estimated is

$$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon$$

where y_t = increase in the CPI in year t and y_{t-1} = increase in the CPI in year $t-1$ (which is the lagged value of y_t).

Using the computer

From the Excel output, the estimated result is

$$y_t = 0.616 + 0.856y_{t-1}$$

XLSTAT output for Example 17.12²

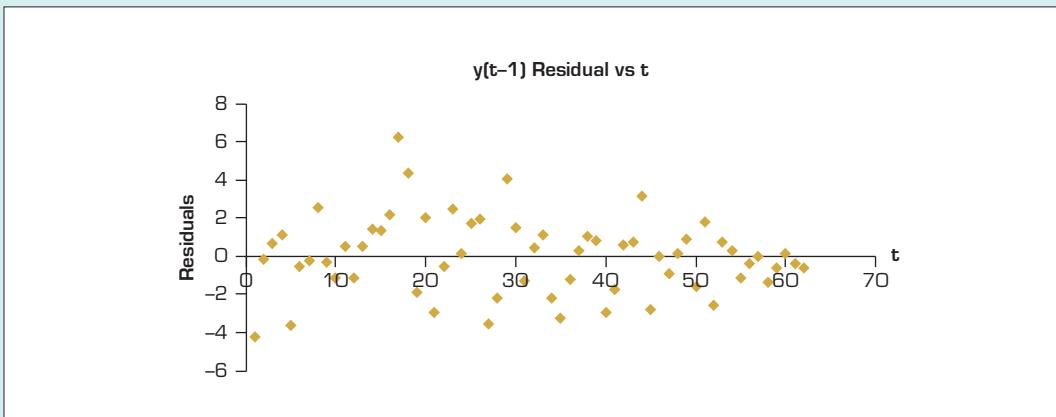
	B	C	D	E	F
1	Regression of variable $y(t)$:				
2					
3	<i>Goodness of fit statistics ($y(t)$):</i>				
4	Observations	62			
5	DF	60			
6	R ²	0.726			
7	Adjusted R ²	0.721			
8	MSE	4.108			
9	RMSE	2.027			
10	DW	1.763			
11					
12	Model parameters (% change in CPI):				
13	<i>Source</i>	<i>Value</i>	<i>Standard error</i>	<i>t</i>	<i>Pr > t </i>
14	Intercept	0.616	0.412	1.496	0.140
15	$y(t-1)$	0.856	0.068	12.603	< 0.0001

Using the first order autoregressive model, we found that $R^2 = 0.73$, indicating better fit, and we also found the Durbin–Watson statistic to be $d = 1.763$. Using the DW critical values in Table 12(a) of Appendix B

² Data Analysis in EXCEL can also be used for the regression estimation. However, it does not provide the value of the Durbin–Watson test statistic to test for autocorrelation. The EXCEL Add-in, **Data Analysis Plus**, can also be used for this purpose. A residual plot in Excel can be used to visually identify positive or negative autocorrelation (see Section 17.7d).



for $k = 1$ and $n = 63$, $d_L = 1.57$, $d_U = 1.63$, $4 - d_U = 2.37$ and $4 - d_L = 2.43$. As $1.63 < d = 1.76 < 2.37$, there is no autocorrelation. Alternatively, we can use Excel to plot the residuals against time.



As can be seen from the residual plot, the plot does not exhibit any pattern and we can safely conclude that there is no autocorrelation. Therefore, the first-order autoregressive model has a better fit and does not have the autocorrelation problem and we can use this estimated model for prediction.

The forecast for 2020 is

$$\hat{y}_{2020} = 0.616 + 0.856y_{2019} = 0.616 + 0.856(1.6) = 1.99\%$$

Interpreting the results

The autoregressive model predicts that in 2020 the CPI should increase by 1.99%.

EXERCISES

Learning the techniques

- 17.42 XR17-42** The following trend line and seasonal indexes were calculated from 10 years of quarterly observations:

$$\hat{y}_t = 150 + 3t$$

Quarter	S_i
1	0.7
2	1.2
3	1.5
4	0.6

Forecast the next four values.

- 17.43 XR17-43** The following trend line and seasonal indexes were calculated from four weeks of daily observations:

$$\hat{y}_t = 120 + 2.3t$$

Day	S_i
Sunday	1.5
Monday	0.4
Tuesday	0.5
Wednesday	0.6
Thursday	0.7
Friday	1.4
Saturday	1.9

Forecast the seven values for the next week.

- 17.44 XR17-44** The following trend line and seasonal indexes were calculated from six years of quarterly observations:

$$\hat{y}_t = 2000 + 80t - 2t^2$$

Quarter	
1	0.6
2	0.9
3	1.1
4	1.4

Forecast the four quarterly values for next year.

- 17.45** Regression analysis with $t = 1$ to $t = 96$ was used to develop the following forecast equation:

$$\hat{y}_t = 220 + 6.5t + 13Q_1 - 1.6Q_2 - 1.3Q_3$$

where

$Q_i = 1$ if quarter i ($i = 1, 2, 3$)

= 0 otherwise

Forecast the next four values.

- 17.46** Daily observations for 52 weeks (five days per week) have produced the following regression model.

$$\hat{y}_t = 1500 + 250t - 20D_1 + 10D_2 + 20D_3 + 50D_4$$

where

$D_1 = 1$ (if Monday)

= 0 (otherwise)

$D_2 = 1$ (if Tuesday)

= 0 (otherwise)

$D_3 = 1$ (if Wednesday)

= 0 (otherwise)

$D_4 = 1$ (if Thursday)

= 0 (otherwise)

Forecast the next week's values.

- 17.47** A daily newspaper wanted to forecast two-day revenues from its classified ads section. The revenues (in \$'000) were recorded for the past 104 weeks. From these data, the regression equation was calculated. Forecast the two-day revenues for the next week. (Note that the newspaper appears six days per week.)

$$\hat{y}_t = 2550 + 0.5t - 205D_1 - 60D_2 \quad (t = 1, 2, \dots, 312)$$

where

$D_1 = 1$ (if Monday or Tuesday)

= 0 (otherwise)

$D_2 = 1$ (if Wednesday or Thursday)

= 0 (otherwise)

- 17.48** Use the following autoregressive model to forecast the value of the time series if the last observed value is 65:

$$\hat{y}_t = 625 - 1.3y_{t-1}$$

- 17.49** The following autoregressive model was developed:

$$\hat{y}_t = 155 + 21y_{t-1}$$

Forecast the time series if the last observation is 11.

- 17.50** For Exercise 17.23, forecast the time series for Monday to Friday of week 5.

- 17.51** For Exercise 17.31 forecast the overseas arrivals over the next eight quarters.

Applying the techniques

- 17.52 Self-correcting exercise.** The following regression model was produced from a set of quarterly observations:

$$\hat{y}_t = 1122.4 + 19.79t - 0.46Q_1 - 91.65Q_2 + 7.16Q_3 \quad (t = 1, 2, \dots, 22)$$

where

$Q_i = 1$ if quarter i ($i = 1, 2, 3$)

= 0 otherwise

Use this model to forecast the number of subscribers over the next four quarters.

- 17.53** The following autoregressive model was produced from the time series in Exercise 17.30:

$$\hat{y}_t = 2685.1 + 0.449y_{t-1}$$

Forecast the overseas arrivals in the fourth quarter of 2019.

- 17.54 XR17-54** Retail turnover in household goods is an important indicator of overall economic activity. It, in turn, is affected by people's perception of the economy and especially by prevailing interest rates and financing availability. Retail turnover is said to exhibit considerable seasonal variability. The monthly retail turnover data for Australia for the period 2005–2019 were recorded. (Source: Australian Bureau of Statistics, 2020, *Retail Trade, Australia*, cat. no. 8501.0, ABS, Canberra.)

a Estimate a linear trend line of the form

$$y_t = \beta_0 + \beta_1 t + \varepsilon \quad (t = 1, 2, \dots, 180)$$

b Using the trend line, calculate the seasonal (monthly) indexes.

c Forecast the monthly total retail turnover for Australia for the 12 months in 2020.

- 17.55** In Exercise 17.32, the linear trend line for the time series is

$$\hat{y}_t = 145 + 1.66t \quad (t = 1, 2, \dots, 28)$$

Use this trend line to obtain the seasonal indexes and forecast the number of pizzas that will be sold daily for the next week.

- 17.56** For Exercise 17.29, forecast the earnings during the four quarters of 2020.

Study Tools

CHAPTER SUMMARY

In this chapter, we discussed the classical *time series* and its decomposition into *long-term trend* and *cyclical, seasonal* and *random variation*. *Moving averages* and *exponential smoothing* were used to remove some of the random fluctuation, enabling us to identify the other components of the time series. The long-term trend was measured more scientifically by one of two regression models – linear or quadratic. The cyclical and seasonal effects are more clearly detected by calculation of *percentage of trend* and *seasonal indexes*.

When the components of a time series are identified, we can select one of many available techniques to forecast the time series. Three forecasting techniques were described in this chapter. When there is no or very little trend, or cyclical and seasonal variation, *exponential smoothing* is recommended. When trend and seasonality are present, we can use *regression analysis with seasonal indexes or indicator (dummy) variables* to make predictions. When the errors are dependent, we can use the *autoregressive model*.

The data files for Examples, Exercises and Cases are provided on the companion website (accessible through <https://login.cengagebrain.com/>).

SYMBOLS

Symbols	Represents
y_t	Time series
S_t	Exponentially smoothed time series
ω	Smoothing constant
F_t	Forecast time series

SUMMARY OF FORMULAS

Autoregressive model	$y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t$
Exponential smoothing	$ES_t = \omega y_t + (1-\omega)ES_{t-1}, t = 2, 3, \dots; ES_1 = y_1$
Forecast of trend and seasonality	$F_t = (\hat{\beta}_0 + \hat{\beta}_1 t)S_t$
Mean absolute deviation	$MAD = \frac{\sum_{i=1}^n y_i - F_t }{n}$
Sum of squares for error	$SSFE = \sum_{i=1}^n (y_i - F_t)^2$

SUPPLEMENTARY EXERCISES

17.57 XR17-57 The revenue of a chain of ice-cream stores is listed for each quarter during the years 2016–20.

Revenue (\$million)

Quarter	2016	2017	2018	2019	2020
1	16	14	17	18	21
2	25	27	31	29	30
3	31	32	40	45	52
4	24	23	27	24	32

- a Plot the time series.
- b Discuss why exponential smoothing is not recommended as a forecasting method in this case.
- c Calculate the four-quarter centred moving averages, and plot these values.
- d Use the moving averages calculated in part (c) to calculate the seasonal (quarterly) indexes.

- e Regression analysis produced the following trend line:

$$\hat{y}_t = 20.2 + 0.732t \quad (t = 1, 2, \dots, 20)$$

Using this trend line, calculate the seasonal indexes.

- f Use the trend line and seasonal indexes calculated in part (e) to forecast revenues for the four quarters of 2021.

- g The following multiple regression model was produced:

$$\hat{y}_t = 18.65 + 0.613t - 6.963Q_1 - 3.625Q_2 + 14.613Q_3 \quad (t = 1, 2, \dots, 20)$$

where

$$Q_i = 1 \text{ if quarter } i \quad (i = 1, 2, 3)$$

$$= 0 \text{ otherwise}$$

Use this model to forecast revenues for the four quarters of 2021.

- h The Durbin–Watson statistic for the regression line in part (e) is $d = 2.08$. What does this tell you about the likelihood that an autoregressive model will produce accurate forecasts?
- i Suppose that the actual 2021 revenues are as shown in the following table.

Quarter	2021 revenues (\$million)
1	25
2	35
3	58
4	37

Calculate MAD for the forecasts in parts (f) and (g).

- j Repeat part (i), using SSFE instead of MAD.

17.58 XR17-58 The monthly Victorian employment participation rates for January 2014 to December 2019 are recorded. (Source: Australian Bureau of Statistics, 2019, various issues of *The Labour Force, Australia*, cat. no. 6202.0, ABS, Canberra.)

- a Plot the time series, to confirm the assertion about seasonal variation.

- b Estimate a linear trend line for the period January 2014 to December 2019 of the form

$$y_t = \beta_0 + \beta_1 t + \varepsilon \quad (t = 1, 2, \dots, 72)$$

Calculate the seasonal (monthly) indexes.

- c Using the trend line and the seasonal indexes calculated in part (b), forecast the monthly Victorian participation rates for January–March 2020.

- d Estimate the following multiple regression model for the Victorian participation rate data:

$$\begin{aligned} y_t = & \beta_0 + \beta_1 t + \beta_1 M_1 + \beta_2 M_2 + \beta_3 M_3 \\ & + \beta_4 M_4 + \beta_5 M_5 + \beta_6 M_6 + \beta_7 M_7 \\ & + \beta_8 M_8 + \beta_9 M_9 + \beta_{10} M_{10} + \beta_{11} M_{11} \end{aligned} \quad (t = 1, 2, \dots, 69)$$

where

$$M_i = 1 \text{ if month } i, \quad (i = 1, 2, \dots, 11)$$

$$= 0 \text{ otherwise}$$

(1 = January, ..., 11 = November)

Use this model to forecast the January, February and March 2020 monthly Victorian participation rates.

- e Obtain the Durbin–Watson statistic in part (b) for the regression line. Does this value indicate strong first-order autocorrelation? Estimate the autoregressive model

$$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon$$

Use this model to forecast the monthly Victorian participation rates for January, February and March 2020.

- f The actual 2020 January–March Victorian employment participation rates were as follows:

Month	Participation rate
Jan 2020	66.1
Feb 2020	67.2
Mar 2020	66.7

For each set of forecasts in parts (c), (d) and (e), calculate MAD to determine which model works best for January–March 2020.

- g Repeat part (f), using SSFE instead of MAD.

Case Studies

CASE 17.1 Part-time employed females

C17-01 For many reasons, the number of people looking for part-time jobs is increasing in Australia, with some seasonal variation. The monthly numbers of part-time employed females (in thousands) for 2009–19 are recorded.

An employment agency would like to model the part-time female jobs market using the data. The manager of the employment agency approaches you for help. Use whatever forecasting technique you judge to be appropriate and the data for 2009–18 to forecast the number of monthly part-time employed females for 2019. Comparing the 2019 monthly forecasts with the actual 2019 monthly data given, comment on the quality of your forecasts.

Source: Australian Bureau of Statistics, *The Labour Force, Australia*, cat. no. 6202.0, December 2019, ABS, Canberra.

CASE 17.2 New Zealand tourism: Tourist arrivals

C17-02 An important performance indicator of a country's tourism industry is the number of tourist arrivals. A tourism operator wants to predict the tourist arrivals to New Zealand in 2020. He has recorded monthly data for tourist arrivals for the period 2010–19. Model the tourist arrivals time series using data for 2010–18 and obtain the monthly predicted values for 2019 and 2020. Use the actual and predicted arrivals for 2019 to assess the quality of the forecasts.

Source: <https://www.stats.govt.nz/topics/tourism>

CASE 17.3 Seasonal and cyclical effects in number of houses constructed in Queensland

C17-03 The building industry in Queensland contributes significantly to the Australian workforce. Information on forecasts for new dwelling starts would be extremely useful in determining a number of variables including demand for housing, availability of labour and price of building materials. Monthly data for the number of dwellings built in Queensland during years 2013–19 are recorded. Analyse the trend, seasonal and cyclical nature of the time series data for the period 2013–18 and forecast the monthly new dwelling starts for 2019. Use the actual 2019 data to assess the quality of your forecast for 2019.

Source: Australian Bureau of Statistics, various issues of *Building Approvals, Australia*, cat. no. 8631.0, December 2019.

CASE 17.4 Measuring the cyclical effect on Woolworths' stock prices

C17-04 The weekly closing Woolworths' stock prices for the period 11 March 2019 to 17 February 2020 was recorded. Assuming a linear trend, measure the cyclical variation by analysing the percentage of trend for each week.

Source: <https://nz.finance.yahoo.com/quote/CMTL/history?p=CMTL&guccounter=1>

Index numbers

Learning objectives

This chapter shows you how to compute the various index numbers such as the simple and weighted aggregate indexes, and discusses their uses and appropriateness.

At the completion of this chapter, you should be able to:

- L01** understand the concept of index numbers
- L02** recognise which index is appropriate for specific uses
- L03** calculate a simple aggregate and weighted aggregate index
- L04** calculate the average of relative price index
- L05** compute Laspeyres, Paasche and Fisher indexes
- L06** understand how the Australian Consumer Price Index (CPI) is calculated
- L07** deflate wages and GDP using the CPI.

CHAPTER OUTLINE

Introduction

- 18.1** Constructing unweighted index numbers
- 18.2** Constructing weighted index numbers
- 18.3** The Australian Consumer Price Index (CPI)
- 18.4** Using the CPI to deflate wages and GDP
- 18.5** Changing the base period of an index number series

SPOTLIGHT ON STATISTICS

How are Aussies and Kiwis performing in their earnings over time?

When people's incomes are reported, they are usually quoted in terms of prices prevailing at the time of reporting and are called 'nominal income (or in current \$)'. Therefore, as price levels usually differ between two different time periods, in order to compare incomes from different time periods, the income figures must be adjusted for price changes over the time periods under consideration and are called 'real income (or in constant \$)'. The Consumer Price Index (CPI) is commonly used to measure the overall movement in prices, hence the overall changes in prices.

A social scientist who is attempting to analyse the earning performance of Australians and New Zealanders over the past 11 years has collected data on the average weekly earnings (nominal) of Australians, average hourly earnings of New Zealanders, and CPI figures for Australia and New Zealand. The data are stored in **CH18:XM18-00**. How could the social scientist use these four series of data to see how the earnings of Australians and New Zealanders have changed over the past 11 years in real terms? For a solution, see pages 819–21.



Source: iStock.com/Hong Li

Introduction

index numbers

Measure the changes over time of particular time-series data.

In Chapter 17 we presented a variety of techniques for analysing and forecasting time series. This chapter is devoted to the simpler task of developing descriptive measurements of the changes in a time series. These measurements are called **index numbers**. Like the descriptive statistics introduced in Chapter 2 and used throughout this book, index numbers allow the statistics practitioner to summarise a large body of data with a single number. An index number is a ratio (generally expressed in percentage form) of one value to another, where one of the values summarises a given group of items and the other value summarises a ‘base’ group of items. The base group is used as a basis for comparison with the given group. Even though many applications of index numbers are for time series, index numbers are also commonly used to make comparisons across different regions/countries where regions/countries play the role of time.

For example, shoppers who observe the frequently changing price of goods and services have difficulty assessing what is happening to overall prices. They might know that the price of T-bone steaks has increased, that babysitters charge more, and that rental and housing prices seem to have increased; but at the same time, desk calculators and personal computers have become cheaper, airline deregulation has lowered the price of many flights, and the cost of certain types of food staples does not seem to have increased significantly. The Australian Bureau of Statistics (ABS) publishes quarterly price changes in various commodity groups. For example, in its media release dated 29 January 2020, the ABS reported that, during the December quarter of 2019, the price of food increased by 1.3%, alcohol and tobacco by 3%, housing by 1%, transport by 1.5%, recreation by 0.9%, education by 1% and insurance and financial services by 0.4%, while household furnishings and equipment decreased by 0.3% and clothing by 0.3% from the previous quarter. The Consumer Price Index gives a general idea of what has happened to overall prices.

Consumer Price Index (CPI)

An economic indicator that measures the changes in the total price of a basket of goods and services.

The **Consumer Price Index (CPI)** measures what has happened to the prices of hundreds of consumer goods and services. The Australian CPI was set to equal 100 in 2012; by 2015 it had risen to 107.5, telling us that prices had generally increased by about 7.5% over the preceding three years; and by 2019 the CPI reading of 114.8 told us that prices had increased by 14.8% over the 7-year period. In Australia, the ABS compiles quarterly CPI figures for the eight capital cities – Sydney, Melbourne, Brisbane, Adelaide, Perth, Hobart, Canberra and Darwin – as well as a weighted average of the eight capital cities for Australia as a whole. A CPI is also published for the all-consumer goods basket, as well as for the all-consumer goods basket excluding housing. In addition, a CPI is published for sub-consumer goods groups such as food, clothing, housing, household equipment, transport, alcohol and tobacco, healthcare, education and recreation. Furthermore, the ABS compiles index series for several other economic indicators such as wholesale prices, retail prices, wage rates, cost levels in industry and volume of agricultural output.

Another popular index is the Dow Jones Industrial Average, which measures the average daily closing share prices of 30 large corporations listed on the New York Stock Exchange. As such, it is perceived by many people (and especially by the media that report it on a daily and sometimes hourly basis) as being a good indicator of the current status of the world stock market. A similar index, which measures the general movement of share prices on the Australian Securities Exchange, is the Australian All Ordinaries Index.

18.1 Constructing unweighted index numbers

In this section, we discuss two methods of constructing unweighted index numbers. We start with the most basic type of index number, known as the *simple price index*.

18.1a Simple price index

Simple price index

A **simple price index** is a ratio of the price of a commodity in current period 1 divided by its value in some base period 0. We can express the ratio of prices as a percentage by multiplying it by 100. That is,

$$I_{1,0} = \frac{p_1}{p_0} \times 100$$

where

$I_{1,0}$ = price index in the current period with base period 0

p_1 = price in the current period

p_0 = price in the base period.

simple price index

The ratio (in percentages) of the price of a commodity in current period 1 divided by its value in some base period 0.

For example, consider the price per litre of beer in Australia for five particular years – 1980, 1990, 2000, 2010 and 2018 – shown in column 2 of the following table. The simple price index for beer is calculated as shown in column 3 of the table.

Constructing the simple price index for beer (1980 = 100)

Year (1)	Price of beer (\$/litre) (2)	Simple price index for beer (base: 1980 = 100) (3)
1980	1.46	$\frac{1.46}{1.46} \times 100 = 100$
1990	3.26	$\frac{3.26}{1.46} \times 100 = 223$
2000	4.60	$\frac{4.60}{1.46} \times 100 = 315$
2010	7.45	$\frac{7.45}{1.46} \times 100 = 510$
2018	9.12	$\frac{9.12}{1.46} \times 100 = 625$

As can be seen from the second row of the table, the value of the price index number for beer in 1990 is 223, with a base price index of 1980 = 100. This means that the price of beer was 123% higher in 1990 than in 1980. Similarly, the price index number for 2018 is 625, meaning that price of beer was 525% higher in 2018 than in 1980.

Even though index numbers are usually mentioned in the context of prices, they can be used to measure the movement of any variable. In the following example, we use index numbers to measure the changes in average weekly earnings of Australian male and female employees.

EXAMPLE 18.1

L01

Average weekly earnings of male and female Australians

XM18-01 Construct an index of average weekly earnings of Australian male and female employees for 1990–2019 using the data in the following table. Use 1990 as the base year, and compare the indexes for males and females.

Average weekly earnings (in dollars) of Australian male and female employees, 1990–2019

Year	Average weekly earnings		Year	Average weekly earnings	
	Male	Female		Male	Female
1990	542.30	352.88	2005	942.70	620.20
1991	574.10	376.88	2006	966.96	632.92
1992	590.35	393.80	2007	1013.90	662.50
1993	605.23	402.35	2008	1059.60	689.70
1994	621.95	414.50	2009	1103.40	719.95
1995	645.10	426.70	2010	1166.45	752.56
1996	664.30	435.00	2011	1213.73	787.42
1997	682.40	451.90	2012	1285.10	821.90
1998	708.30	466.20	2013	1356.70	849.90
1999	726.20	478.30	2014	1364.60	881.30
2000	744.20	490.30	2015	1369.50	907.80
2001	777.30	520.40	2016	1395.10	925.80
2002	815.60	537.40	2017	1417.20	946.80
2003	872.10	567.20	2018	1445.30	976.30
2004	891.20	588.50	2019	1475.60	1010.80

Source: Australian Bureau of Statistics, cat. no. 6302.0, *Average Weekly Earnings*, Table 3, May 2019, Australia.

Solution

For each year (with 1990 as the base year) we calculate the index for the variable 'Earnings (E)'.

$$I_{1,0} = \frac{E_1}{E_0} \times 100$$

Here E refers to the average weekly earnings of Australian male (or female) employees.

For each year (with base 1990 = 100), we calculate the index for average weekly earnings for Australian male and female employees. For example, for a male employee,

$$I_{1991,1990} = \frac{E_{1991}}{E_{1990}} \times 100 = \frac{574.10}{542.30} \times 100 = 105.86$$

and for a female employee,

$$I_{1991,1990} = \frac{E_{1991}}{E_{1990}} \times 100 = \frac{376.88}{552.88} \times 100 = 106.80$$

The results are shown in **Table 18.1**.



TABLE 18.1 Index numbers for average weekly earnings of Australian male and female employees (1990 = 100)

Year (1)	Index number for earnings		Year (1)	Index number for earnings	
	Male (2)	Female (3)		Male (2)	Female (3)
1990	100.0	100.0	2005	173.8	175.8
1991	105.9	106.8	2006	178.3	179.4
1992	108.9	111.6	2007	187.0	187.7
1993	111.6	114.0	2008	195.4	195.4
1994	114.7	117.5	2009	203.5	204.0
1995	119.0	120.9	2010	215.1	213.3
1996	122.5	123.3	2011	223.8	223.1
1997	125.8	128.1	2012	237.0	232.9
1998	130.6	132.1	2013	250.2	240.8
1999	133.9	135.5	2014	251.6	249.7
2000	137.2	138.9	2015	252.5	257.3
2001	143.3	147.5	2016	257.3	262.4
2002	150.4	152.3	2017	261.3	268.3
2003	160.8	160.7	2018	266.5	276.7
2004	164.3	166.8	2019	272.1	286.4

These numbers can be used, for example, to compare female wage increases since 1990 with those of male employees.

Comparing columns 2 and 3, we can see that female earnings grew slightly faster than male earnings during 1990–2019.

Figures 18.1 and **18.2** describe the actual earnings and the index numbers. As you can see, it is difficult to compare relative increases in wages using **Figure 18.1**, but graphing the index numbers in **Figure 18.2** provides additional insights.

FIGURE 18.1 Average weekly earnings, 1990–2019

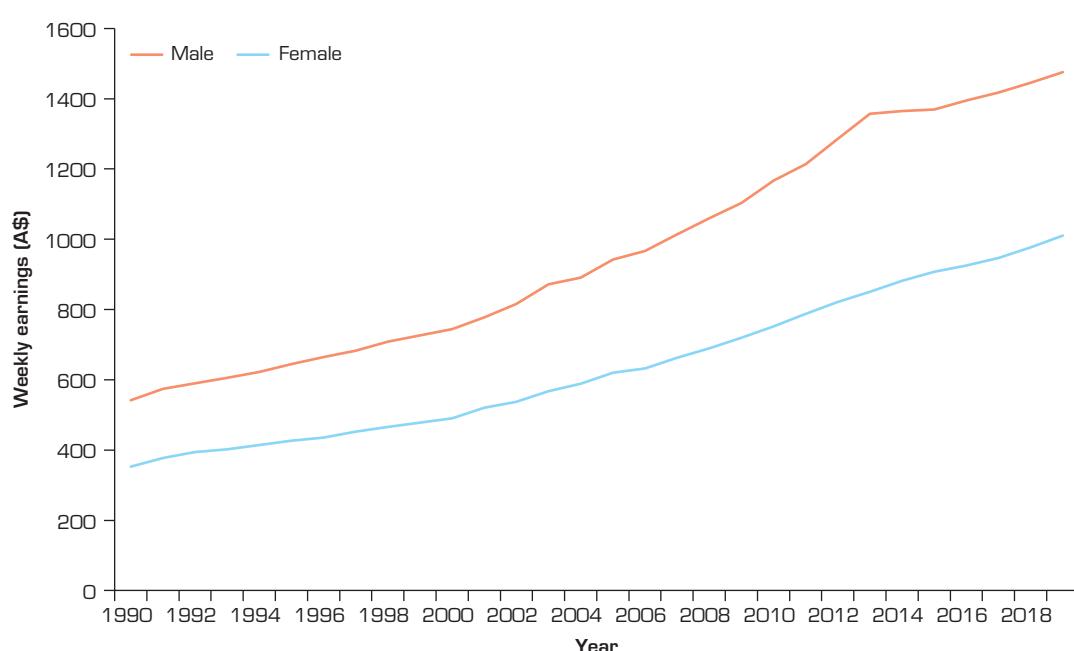
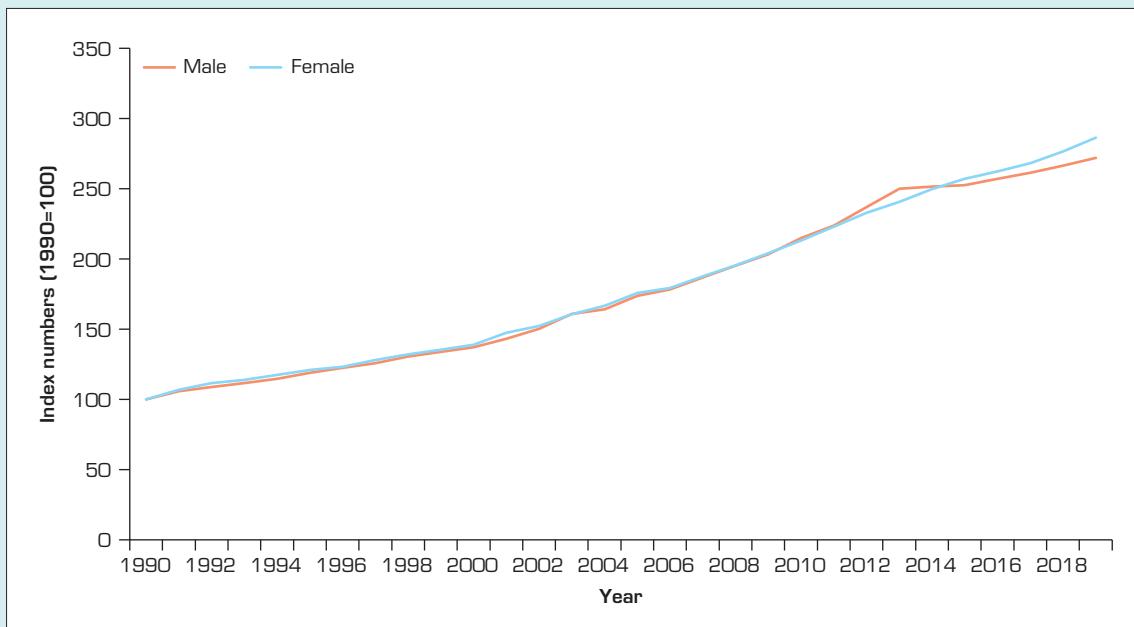


FIGURE 18.2 Index numbers for average weekly earnings (1990 = 100), 1990–2019



The simple price index number measures the price changes of only one item over two time periods. Very frequently, though, we want to measure price changes for a group of commodities. For example, instead of measuring the price changes in the individual beverages beer, wine and spirits, we might want to measure the price changes in the alcoholic beverages group (comprising beer, wine and spirits). Our next index number performs this function.

18.1b Simple aggregate price index

Consider a commodity group consisting of n items.

simple aggregate price index

The ratio (in percentages) of the sum of the prices of n commodities in the current period 1 divided by the sum of the prices of the same n commodities in the base period 0.

Simple aggregate price index

A **simple aggregate price index** is defined as the ratio of the sum of the prices of n commodities in the current period 1 over the sum in the base period 0, multiplied by 100. That is,

$$I_{1,0} = \left[\frac{\sum_{i=1}^n p_{i1}}{\sum_{i=1}^n p_{i0}} \right] \times 100$$

where

$I_{1,0}$ = price index in the current period 1 with base period, 0, price index = 100

p_{i1} = price of commodity i in the current period 1 ($i = 1, 2, \dots, n$)

p_{i0} = price of commodity i in the base period 0 ($i = 1, 2, \dots, n$).

EXAMPLE 18.2

LO3

Unweighted price index for the alcoholic beverages group

XM18-02 The following table gives the prices (\$/litre) of beer, wine and spirits in Australia for the years 1980, 1990, 2010 and 2018. Construct an unweighted index number for the alcoholic beverages group with base 1980 = 100.

Year	Price (\$/litre)		
	Beer	Wine	Spirits
1980	1.46	3.60	17.12
1990	3.26	6.35	37.47
2000	4.60	8.90	50.20
2010	7.45	9.69	80.75
2018	9.12	10.05	95.74

Source: Australian Bureau of Statistics, various issues of *Apparent Consumption of Selected Foodstuffs, Australia*, cat. no. 4315.0.

Solution

For each year the price of the alcoholic beverages group is calculated as the sum of the prices of beer, wine and spirits and presented in column 5 of the following table. The simple aggregate price index for alcohol (base 1980 = 100) is calculated as shown in column 6 of the table.

Constructing the simple aggregate price index for alcohol (1980 = 100)

Year (1)	Price (\$/litre)			Sum of price (5)	Simple aggregate price index for alcohol (base: 1980 = 100) (6)
	Beer (2)	Wine (3)	Spirits (4)		
1980	1.46	3.60	17.12	22.18	$\frac{22.18}{22.18} \times 100 = 100.0$
1990	3.26	6.35	37.47	47.08	$\frac{47.08}{22.18} \times 100 = 212.3$
2000	4.60	8.90	50.20	63.70	$\frac{63.70}{22.18} \times 100 = 287.2$
2010	7.45	9.69	80.75	97.89	$\frac{97.89}{22.18} \times 100 = 441.3$
2018	9.12	10.05	95.74	114.91	$\frac{114.91}{22.18} \times 100 = 518.1$

The index numbers in the last column of the table above indicate that the price of alcohol as a whole increased by 112.3% in 1990 compared to 1980. Further, compared to 1980, it is about 3 times higher in 2000, about 4.5 times higher in 2010 and about 5 times higher in 2018.

EXAMPLE 18.3

LO3

Simple aggregate price index for the fruit and vegetable group

XM18-03 The Australian Bureau of Statistics regularly publishes the average retail prices of selected items for each Australian capital city. The following table presents the prices per kilogram of oranges, bananas, potatoes, tomatoes, carrots and onions in Brisbane during 1995 and 2020. Construct a simple aggregate index of the prices for the fruit and vegetable group using 1995 as the base period.



Fruit and vegetable prices (\$/kg) in Brisbane, 1995 and 2020

Item	Fruit and vegetable prices (\$/kg)	
	1995	2020
Oranges	1.36	3.20
Bananas	2.51	3.00
Potatoes	1.17	2.10
Tomatoes	1.76	3.50
Carrots	1.00	1.70
Onions	1.10	2.05

Source: Australian Bureau of Statistics, CC BY 2.5 AU
<http://creativecommons.org/licenses/by/2.5/au/legalcode>.

Solution

$$\begin{aligned}
 I_{2020,1995} &= \left[\frac{\sum_{i=1}^6 p_{i,2020}}{\sum_{i=1}^n p_{i,1995}} \right] \times 100 \\
 &= \left[\frac{3.20 + 3.00 + 2.10 + 3.50 + 1.70 + 2.05}{1.36 + 2.51 + 1.17 + 1.76 + 1.00 + 1.10} \right] \times 100 \\
 &= \left[\frac{15.55}{8.90} \right] \times 100 \\
 &= 174.7
 \end{aligned}$$

Thus, the increase in total fruit and vegetable prices in Brisbane between 1995 and 2020 was 74.7%.

Simple aggregate price index numbers are very rarely used. Below, we discuss a particular instance in which it is inappropriate to use simple aggregate price index numbers.

18.1c Average of relative price index

In Examples 18.2 and 18.3, when constructing the simple aggregate price index, we assumed that the prices were for the same unit of volume or weight (\$/litre and \$/kg respectively). In many situations this may not be the case. Consider a situation in which the prices are given in dollars per different units of volume, for example, beer in \$/bottle, wine in \$/1000 litres and spirits in \$/litre. In this case, we cannot simply add the prices to obtain the total, as the units are different. To overcome this difficulty, we use price ratios (also called ‘price relatives’), which are unit-free, to construct a price index number. This unweighted index number is known as the *average of relatives*.

Average of relative price index

An **average of relative price index** is a simple average of the n price relatives, p_{i1}/p_{i0} ($i = 1, 2, \dots, n$) multiplied by 100.

$$I_{1,0} = \frac{1}{n} \left[\sum_{i=1}^n \left(\frac{p_{i1}}{p_{i0}} \right) \right] \times 100$$

average of relative price index

A simple average of the n price relatives, multiplied by 100.

EXAMPLE 18.4

LO4

Average of relative price index for the alcoholic beverages group

Refer to the Australian beer, wine and spirits price data in Example 18.2. Construct the average of relative price index for the alcoholic beverages group for the years 1990, 2000, 2010 and 2018 (with base 1980 = 100).

Solution

As the alcoholic beverages group consists of beer, wine and spirits, $n = 3$. To obtain the price index for 2018, we average the three price relatives of the three beverages.

$$\begin{aligned} I_{2018,1980} &= \frac{1}{3} \left[\sum_{i=1}^3 \left(\frac{p_{i,2018}}{p_{i,1980}} \right) \right] \times 100 \\ &= \frac{1}{3} \left[\frac{9.12}{1.46} + \frac{10.05}{3.60} + \frac{95.74}{17.12} \right] \times 100 \\ &= 487.7 \end{aligned}$$

Similarly,

$$\begin{aligned} I_{2010,1980} &= \frac{1}{3} \left[\sum_{i=1}^3 \left(\frac{p_{i,2010}}{p_{i,1980}} \right) \right] \times 100 \\ &= \frac{1}{3} \left[\frac{7.45}{1.46} + \frac{9.69}{3.60} + \frac{80.75}{17.12} \right] \times 100 \\ &= 417.0 \\ I_{2000,1980} &= \frac{1}{3} \left[\sum_{i=1}^3 \left(\frac{p_{i,2000}}{p_{i,1980}} \right) \right] \times 100 \\ &= \frac{1}{3} \left[\frac{4.60}{1.46} + \frac{8.90}{3.60} + \frac{50.20}{17.12} \right] \times 100 \\ &= 285.2 \end{aligned}$$

and

$$\begin{aligned} I_{1990,1980} &= \frac{1}{3} \left[\sum_{i=1}^3 \left(\frac{p_{i,1990}}{p_{i,1980}} \right) \right] \times 100 \\ &= \frac{1}{3} \left[\frac{3.26}{1.46} + \frac{6.35}{3.60} + \frac{37.47}{17.12} \right] \times 100 \\ &= 206.2 \end{aligned}$$

These four index numbers indicate that the price of the alcoholic beverages group increased by 106.2% in 1990, 185.2% in 2000, 317% in 2010 and 387.7% in 2018, relative to the 1980 price level.

EXAMPLE 18.5

LO4

Average of relative price index for the fruit and vegetable group

Refer to the fruit and vegetable price data in Example 18.3. Construct the average of the relative price index for the fruit and vegetable group for 2020 (with base 1995 = 100).

Solution

As the fruit and vegetable group in Example 18.3 is made up of six items, $n = 6$. Therefore,

$$\begin{aligned} I_{2020,1995} &= \frac{1}{6} \left[\sum_{i=1}^6 \left(\frac{p_{i,2020}}{p_{i,1995}} \right) \right] \times 100 \\ &= \frac{1}{6} \left[\frac{3.20}{1.36} + \frac{3.00}{2.51} + \frac{2.10}{1.17} + \frac{3.50}{1.76} + \frac{1.70}{1.00} + \frac{2.05}{1.10} \right] \times 100 \\ &= 181.6 \end{aligned}$$

This index number indicates that fruit and vegetable prices in Brisbane in 2020 had increased by 81.6% relative to the 1995 price level.

Even though the average of a relative price index is made up of unit-free price relatives, the disadvantage with this index is that it gives equal weight to each price relative. Therefore, finding the price changes for the group of goods simply by aggregating the relative prices of all items in the commodity group of interest wouldn't make much sense. Thus, the unweighted aggregate price index numbers considered above must be modified if we do not consume equal quantities of each item.

To remedy this deficiency, in Section 18.2 we introduce the *weighted index number*, which is a weighted sum of all the price relatives, where each price relative is weighted by its relative importance to the average consumer.

18.2 Constructing weighted index numbers

Consider the n price relatives $p_{i1}/p_{i0}, \dots, p_{n1}/p_{n0}$. Let w_1, \dots, w_n be the corresponding weights.

weighted aggregate price index

The weighted sum of the n price relatives, multiplied by 100, where the weights are non-negative and sum to 1.

Weighted aggregate price index

The **weighted aggregate price index** for period 1 (with a base period 0) is the weighted sum of the n price relatives

$$I_{1,0} = \left[\sum_{i=1}^n w_i \frac{p_{i1}}{p_{i0}} \right] \times 100$$

where the weights w_i are non-negative and sum to 1. That is,

$$0 \leq w_i \leq 1 \text{ and } \sum_{i=1}^n w_i = 1$$

18.2a The Laspeyres price index

Now the question is how to choose the weights $w_i, i = 1, \dots, n$. A popular choice for the weight w_i is the base period budget share. The budget shares are calculated as the expenditure on each item as a proportion of the total expenditure on all items.

If a consumer in base period 0 purchases quantities $q_{i0}, q_{i1}, \dots, q_{n0}$ of the n items at prices $p_{i0}, p_{i1}, \dots, p_{n0}$, then the budget share for good i is given by

$$w_i = \frac{p_{i0}q_{i0}}{M_0}, \quad i = 1, 2, \dots, n$$

where $M_0 = \sum_{i=1}^n p_{i0}q_{i0}$ is the total expenditure in the base year. For example, if a consumer spends \$20 per week on alcohol and \$400 per week on all goods including alcohol, then the budget share for alcohol is $(20/400) = 0.05$ or 5%.

Substituting the value of w_i in $I_{1,0}$ gives

$$\begin{aligned} I_{1,0}^{LP} &= \sum_{i=1}^n \left[\frac{p_{i0}q_{i0}}{M_0} \times \frac{p_{i1}}{p_{i0}} \right] \times 100 \\ &= \frac{\sum_{i=1}^n p_{i1}q_{i0}}{M_0} \times 100 \\ &= \frac{\sum_{i=1}^n p_{i1}q_{i0}}{\sum_{i=1}^n p_{i0}q_{i0}} \times 100 \end{aligned}$$

Laspeyres price index

Ratio (in percentages) of current price expenditure to the base price expenditure of purchasing a base period basket of goods.

This weighted aggregate price index is known as the **Laspeyres price index**.

Laspeyres price index

The Laspeyres price index is the ratio (expressed as a percentage) of the total cost in the *given year* of the quantity of each item consumed in the base year to the total cost for the same quantity in the *base year*.

$$I_{1,0}^{\text{LP}} = \left[\frac{\sum_{i=1}^n p_i q_{i,0}}{\sum_{i=1}^n p_{i,0} q_{i,0}} \right] \times 100$$

An advantage of this index is that it requires the consumption for the (base) year 0 only, which is easily available in period 1.

EXAMPLE 18.6

LO5

Laspeyres price index of the cost of food 2010 and 2020

XM18-06 In 2010, a Brisbane family's weekly diet included 6 kg of fish, 2 kg of beef and 2 kg of veal. In 2020, because of the cost increases in these products, the family's diet changed so that each week they consumed 4 kg of chicken, 1 kg each of beef, veal and pork, and 5 kg of fish. The prices for each of these items are listed in the following table. Calculate the Laspeyres price index for 2020, using 2010 as the base.

Meat prices, 2010 and 2020

Item	2010		2020	
	Price (\$) $p_{i,2010}$	Quantity (kg) $q_{i,2010}$	Price (\$) $p_{i,2020}$	Quantity (kg) $q_{i,2020}$
Beef	6.20	2	15.00	1
Veal	8.50	2	12.50	1
Pork	6.50	0	8.00	1
Chicken	3.20	0	5.00	4
Fish	7.00	6	18.00	5

Solution

The Laspeyres price index for 2020 (with base 2010 = 100) is

$$\begin{aligned} I_{2020,2010}^{\text{LP}} &= \left[\frac{\sum_{i=1}^5 p_{i,2020} q_{i,2010}}{\sum_{i=1}^5 p_{i,2010} q_{i,2010}} \right] \times 100 \\ &= \left[\frac{15.00(2) + 12.50(2) + 8.00(0) + 5.00(0) + 18.00(6)}{6.20(2) + 8.50(2) + 6.50(0) + 3.20(0) + 7.00(6)} \right] \times 100 \\ &= \left[\frac{163.00}{71.40} \right] \times 100 \\ &= 228.3 \end{aligned}$$

This index number tells us that, if the 2020 diet had been the same as the 2010 diet, this part of the family budget would have shown a cost increase of 128.3%. It is important to note that this calculation assumes that in 2020 the family still ate 6 kg of fish, 2 kg of beef and 2 kg of veal weekly, as in 2010, even though the family diet changed significantly in 2020.

18.2b The Paasche price index

Another choice for the weight w_i in the weighted aggregate price index number formula above is the current period budget share. If a consumer, in current period 1, purchases quantities $q_{11}, q_{21}, \dots, q_{n1}$ of the n items at the same base period prices, $p_{10}, p_{20}, \dots, p_{n0}$, then the current period budget share for commodity i is given by

$$w_i = \frac{p_{i0}q_{i0}}{M_1}, \quad i=1, 2, \dots, n$$

where $M_1 = \sum_{i=1}^n p_{i0}q_{i0}$ is the total expenditure in current period 1, based on base period prices.

Substituting the value of w_i in $I_{1,0}$ gives

$$\begin{aligned} I_{1,0}^{PP} &= \sum_{i=1}^n \left[\frac{p_{i0}q_{i1}}{M_1} \times \frac{p_{i1}}{p_{i0}} \right] \times 100 \\ &= \sum_{i=1}^n \frac{p_{i1}q_{i1}}{M_1} \times 100 \\ &= \frac{\sum_{i=1}^n p_{i1}q_{i1}}{M_1} \times 100 \\ &= \frac{\sum_{i=1}^n p_{i1}q_{i1}}{\sum_{i=1}^n p_{i0}q_{i1}} \times 100 \end{aligned}$$

This weighted aggregate price index is known as the **Paasche price index**.

Paasche price index
Ratio (in percentages) of current price expenditure to the base price expenditure of purchasing a current period basket of goods.

Paasche price index

The Paasche price index is the ratio (expressed as a percentage) of the total cost in the *current year* of the quantity of each item consumed in that year to what would have been the total cost of these quantities in the *base year*.

$$I_{1,0}^{PP} = \left[\frac{\sum_{i=1}^n p_{i1}q_{i1}}{\sum_{i=1}^n p_{i0}q_{i1}} \right] \times 100$$

One of the disadvantages of this index is that it requires the consumption figures for the (current) year 1, which are not generally available in period 1. Due to this difficulty, this index is not used often.

EXAMPLE 18.7

LO5

Paasche index for the cost of food, 2010 and 2020

Calculate the Paasche price index for 2020, using 2010 as the base, for Example 18.6. The Paasche price index for 2020 (with base year 2010 = 100)

$$\begin{aligned} I_{2020,2010}^{\text{PP}} &= \left[\frac{\sum_{i=1}^5 p_{i,2020} q_{i,2020}}{\sum_{i=1}^5 p_{i,2010} q_{i,2020}} \right] \times 100 \\ &= \left[\frac{15.00(1) + 12.50(1) + 8.00(1) + 5.00(4) + 18.00(5)}{6.20(1) + 8.50(1) + 6.50(1) + 3.20(4) + 7.00(5)} \right] \times 100 \\ &= \left[\frac{145.50}{69} \right] \times 100 \\ &= 210.9 \end{aligned}$$

This index number tells us that, if the 2010 diet had been the same as the 2020 diet, this part of the family budget would have shown an increase of 110.9%.

We now have two different index numbers based on the same data. The Laspeyres price index is

$$I_{2020,2010}^{\text{LP}} = 228.3$$

and the Paasche price index is

$$I_{2020,2010}^{\text{PP}} = 210.9$$

Which index should be used to describe the price change? To answer this question, we must first address another question: What do we want to measure? Do we want to know what happened to the prices, assuming that the 2010 diet remained unchanged? In that case we would use $I_{2020,2010}^{\text{LP}}$ where prices increased by 128.3%. Or do we want to know what happened to the prices, assuming that the 2020 diet was in effect in 2010? Then we would use $I_{2020,2010}^{\text{PP}}$ where prices increased by only 110.9%. Thus, our choice depends on what we wish to measure.

18.2c Comparison of Laspeyres and Paasche index numbers

The attraction of the Laspeyres index is that the weights need not be changed every time, which saves a lot of costly surveys, as it uses weights based on the base period quantities. The use of fixed base-period weights is only realistic during a period when consumption patterns do not change much. Thus, the major criticism of the Laspeyres price index concerns the fact that it does not make allowance for changes in quantities purchased due to changes in tastes, income, product quality, substitution, etc. This problem can be eliminated by using the Paasche index, as it allows the weights, based on the current period quantities, to change over time to reflect the changing consumption pattern. But this causes problems for the construction of the index, as weights have to be altered in each period. This involves additional time and costs in data collection. Due to this difficulty the Laspeyres index is in fact more widely used.

18.2d Fisher price index

In a period of rising prices, and when demand patterns change in response to relative price changes, the Laspeyres index will tend to *overestimate* (or show an upward bias in) the overall price rise, while the Paasche index will tend to *underestimate* (or show a downward bias in) the overall price rise. To remove the over- and underestimating price change characteristic of the Laspeyres and Paasche indexes, the statistician Professor I Fisher introduced the **Fisher price index**, which is the geometric mean of the Laspeyres and Paasche index numbers.

Fisher price index

A geometric mean of the Laspeyres price index and the Paasche price index.

Fisher price index

The Fisher price index is the geometric mean of the Laspeyres price index ($I_{1,0}^{LP}$) and the Paasche price index ($I_{1,0}^{PP}$). That is,

$$I_{1,0}^{FP} = \sqrt{I_{1,0}^{LP} \times I_{1,0}^{PP}}$$

For the diet example above, we have $I_{2020,2010}^{LP} = 228.3$, $I_{2020,2010}^{PP} = 210.9$. Therefore

$$I_{2020,2010}^{FP} = \sqrt{228.3 \times 210.9} = 219.4$$

As the Fisher index requires the calculation of the Paasche index, the problems associated with the Paasche index (such as the additional time and costs in data collection) apply to this index and thus it is not often used.

18.3 The Australian Consumer Price Index (CPI)

The CPI for Australia provides a general measure of the changes in prices of consumer goods and services purchased by Australian households. The CPI was first compiled by the ABS in the June quarter of 1960, but has been calculated retrospectively for all quarters back to September 1948. Initially the CPI was calculated only for the six capital cities – Adelaide, Brisbane, Melbourne, Perth, Sydney and Hobart – but a CPI for Canberra was commenced in 1964 and for Darwin in 1982. Currently, the ABS publishes quarterly (every three months ending March, June, September and December) the CPI for each of the eight capital cities, as well as a combined index that is a weighted average of the eight capital cities, and comprises the CPI for Australia. The CPI is calculated by a variation of the Laspeyres method.

The CPI figures appear in the ABS publication, *Consumer Price Index Australia* (cat. no. 6401.0), and in some other ABS publications such as the *Australian Economic Indicators* (cat. no. 1350.0). In addition, key CPI results appear on the ABS website (<http://www.abs.gov.au>).

In Australia, the CPI has always been an important economic indicator and its movements have had both direct and indirect impacts on all Australians in recent years. The CPI has been used for a variety of purposes, such as in the development and analysis of government economic policy; to determine the size and nature of wage adjustments for the indexation of income ranges for income tax purposes by the taxation department; to adjust or index pension and superannuation payments; and to adjust business contracts, rental agreements, building contracts, insurance coverages, child support payments and some other transactions that are tied in some manner to changes in the CPI. Consequently, at least some basic understanding of the CPI and its construction is essential for every Australian. From time to time the ABS publishes information booklets (for example, cat. nos. 6431.0, 6440.0, 6441.0, 6450.0, 6461.0 and 6470.0) to outline the construction of the CPI and changes in the calculation of the CPI.

18.3a What is the CPI?

The CPI is an important economic indicator that measures the changes in the total price of a typical basket of goods and services (called the *CPI basket*) that are purchased by a large proportion of metropolitan employee households. The composition of the current CPI basket is based on the pattern of the household expenditure in 2015–16 collected from that year's Household Expenditure Survey by the ABS. The ABS termed this household group the *CPI population group*. This group includes a wide variety of subgroups such as wages and

salary earners, self-employed people, self-funded retirees, aged pensioners and social welfare beneficiaries living in the eight capital cities. The current CPI population group represents about 64% of all Australian private households.

The total price of the CPI basket in the reference base period is assigned a base index value of 100, and the total price of the CPI basket in other periods is expressed as a percentage of the total price in the base period. For example, if the price of the CPI basket has increased by 5% since last year, then the index for this year would be 105 with last year's index = 100. Similarly, if the price has declined by 5%, then the index for this year is 95 with last year's index = 100. The current reference base period for the Australian CPI is the financial year 2011–12. The CPI measures the price movement and not the actual price levels. For example, if price indexes for beer and wine are 112 and 103, respectively in 2016 with base 2012 = 100, it does not mean that beer is more expensive than wine. What it means is that the beer price has increased by 12% from last year's beer price, while the wine price has increased by only 3%. Wine may or may not be more expensive than beer.

The introduction of the CPI brought changes in the way retail price movements were measured. Instead of the former emphasis on long-term, fixed-weight indexes, the CPI comprises a series of short-term indexes that are chain-linked together to form a continuous long-term series. This chain-link approach allows changes in consumers' expenditure patterns to be reflected in the CPI.

18.3b Composition of the CPI basket

The CPI basket consists of all the important kinds of consumer purchase items to reflect price changes for a wide range of consumer goods and services. Each quarter, the ABS collects approximately 100 000 price quotations across the eight capital cities. The CPI basket is divided into a number of expenditure classes (groups) – namely, food, clothing, housing, household equipment and operation, transportation, tobacco and alcohol, health and personal care, and recreation and education. These groups are then divided into a number of subgroups and the subgroups into several expenditure classes. For example, the food group consists of subgroups such as dairy produce (fresh milk, flavoured milk, processed cheese and natural cheese, etc.), cereal products (bread, cakes and biscuits, breakfast cereals, etc.), meat and seafood (steak, corned beef, minced steak, beef sausages, etc.) and fresh fruit and vegetables (oranges, apples, washed potatoes, unwashed potatoes, etc.). Indexes are published for each of these groups for each of the eight capital cities and a weighted average of all eight capital cities.

The goods and services included in the CPI basket are selected such that their prices can be associated with an identifiable and specific commodity or service, for example, price changes for a given size and quality of shirt, for a television set of a certain size, for different capacity motor vehicle registration fees, local government rates and charges, etc. These can be measured and are included in the CPI. Sales and excise taxes are also included in the CPI basket. Items such as income tax and personal savings are not included.

18.3c Construction of the CPI

To construct the CPI, the ABS collects prices for most commodities every quarter from retail outlets. The collection of prices of all CPI basket items in each capital city is carried out by trained and experienced staff working in various ABS offices around Australia.

The ABS assigns weights for each expenditure class according to its relative importance. For example, a larger weight will be given to the food expenditure class than to the household equipment and operation expenditure class, as a 10% rise in the price of bread in the food group will have a greater impact on the CPI than a 50% increase in the price of a toaster in the household equipment and operation group, because bread is purchased more frequently than toasters. The weight for each group is determined on the basis of household expenditure

surveys carried out by the ABS. From time to time the ABS reviews these fixed weights and new fixed weights are introduced (usually at five-year intervals) to reflect up-to-date expenditure patterns. In calculating the CPI, price changes for various expenditure classes are combined with their corresponding weights. The four recent CPI series published by the ABS are called the '14th Series', the '15th Series', the '16th Series' and the '17th Series'. The 14th Series was introduced in the September quarter of 2000 using expenditure weights from the 1998–99 Household Expenditure Survey, to reflect the introduction of the new tax system on 1 July 2000. The 15th Series was introduced in the September quarter of 2005 and the 16th series was introduced in the September quarter of 2011, with 87 expenditure classes based on the 2009–10 Household Expenditure Survey data. The 17th series was introduced in 2017.

Detailed information on the CPI is available in the ABS publication *The Australian Consumer Price Index: Concepts, Sources and Methods* (cat. no. 6461.0), *Statistical Concepts Reference Library* (cat. no. 1361.030.001) and in another ABS publication, *Information Paper: Introduction of the 17th Series Australian Consumer Price Index*, Sept 2011 (cat. no. 6470.0.55.001).

The fixed weighting patterns for the 14th, 15th, 16th and 17th Series calculated by the ABS for the eight capital cities combined are given in **Table 18.2**.

TABLE 18.2 Weighting patterns of the CPI 14th, 15th, 16th and 17th Series, Australia

Group	Subgroup	CPI 14th Series		CPI 15th Series		CPI 16th Series		CPI 17th Series	
		Group total	Subgroup total						
Food		17.72		15.44		16.84		16.09	
	Dairy products		1.51		1.19		1.15		0.99
	Cereal products		2.20		1.72		1.71		1.49
	Meat and seafood		2.62		2.42		2.29		2.19
	Fresh fruit and vegetables								
	Processed fruit and vegetables		2.30		2.11		2.95		2.35
	Soft drinks, ice cream and confectionery		2.48		1.96		1.14		0.99
	Meals out, take-away food		4.93		4.56		5.43		5.88
	Other food		1.69		1.49		2.17		2.21
Clothing		5.19		3.91		3.98		3.55	
	Men's and boys' clothing		0.98		0.75		0.74		0.60
	Women's and girls' clothing		1.80		1.41		1.47		1.27
	Piece goods and clothing		0.47		0.4		0.31		0.35
	Footwear		0.83		0.64		0.61		0.54
	Clothing and footwear services		1.1		0.72		0.86		0.80
Housing		16.51		16.43		18.69		18.62	
	Rent		5.60		5.22		6.71		7.22
	Home ownership		10.91		11.21		11.98		11.40

TABLE 18.2 Continued

Group	Subgroup	CPI 14th Series		CPI 15th Series		CPI 16th Series		CPI 17th Series	
		Group total	Subgroup total						
Household equipment and operation		11.32		12.71		12.72		13.44	
	Fuel and light	3.23		3.10		3.61		4.06	
	Furniture and floor coverings	3.58		3.13		1.91		1.70	
	Appliances					1.43		1.56	
	Drapery					0.61		0.49	
	Household utensils and tools	1.98		1.76		2.86		2.65	
	Household supplies and services	1.91		4.72		2.29		2.99	
	Postal and telephone services	0.62							
Transportation		15.25		13.11		11.55		10.32	
	Public motoring	14.40		12.38					
	Urban transport fees	0.85		0.73					
Tobacco and alcohol		7.41		6.79		7.06		7.09	
	Alcoholic beverages	5.14		4.38		4.75		4.49	
	Cigarettes and tobacco	2.27		2.41		2.32		2.60	
Health and personal care		4.69		4.71		5.29		5.43	
	Health services	3.55		3.56		3.97		4.25	
	Personal care products	1.14		1.15		1.32		1.18	
Recreation and education		14.98		14.28		15.74		16.98	
	Books, newspapers and magazines	1.08		0.85					
	Other recreational goods	2.70		2.92					
	Holiday travel and accommodation	4.35		4.06					
	Other recreational services	4.16		3.72					
	Education and childcare	2.69		2.73					
Communication		2.88		3.31		3.05		2.68	
	Postal	0.15		0.11					
	Telecommunication	2.73		3.2					
Financial and insurance service		4.04		9.31		5.08		5.80	
	Financial services			7.81					
	Insurance services			1.50					
TOTAL ALL GROUPS		100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Source: Australian Bureau of Statistics, *Australian Consumer Price Index: Concepts, Sources and Methods*, 2011, cat. no. 6461.0; *Consumer Price Index: Historical Weighting Patterns*, 1948–2011, cat. no. 6431.0; *Consumer Price Index 16th Series Weighting Pattern*, 2011, cat. no. 6471.0, ABS, Canberra, and ABS 6470.0.55.001 *Information Paper: Introduction of the 17th Series Australia, CPI*, 2017, Appendix 2.

18.4 Using the CPI to deflate wages and GDP

In this section we discuss two applications of the consumer price index.

18.4a Using the CPI to deflate wages

The CPI is used in numerous applications. One of these involves adjusting wages by removing the effect of inflation, the proportionate change in CPI [$= (\text{CPI}_t - \text{CPI}_{t-1})/\text{CPI}_{t-1}$]. For example, if a person's nominal income increases at approximately the same rate as prices, then his/her real income (or standard of living) is unchanged as he/she will have the same level of purchasing power this year to purchase all the items he/she purchased last year. If a person's real income increases at a faster rate than price increases, then his/her purchasing power improves compared to last year. Similarly, if a person's real income increases at a slower rate than price increases, then his/her purchasing power declines. Therefore, to measure changes in income over time, we should compare the purchasing power at different periods. This means that we should adjust the nominal income to take into account the price changes. This adjusted income/wages is called *real income/wages*.

To obtain the real income, we deflate the nominal income by the CPI. That is,

$$\text{Real income} = \frac{\text{Nominal income}}{\text{CPI}} \times 100$$

Let us consider the following example, in which we compare the wages of Australian employees over two years, 2010 and 2019, by removing the effect of inflation. To do this, we first identify the Australian CPI from 1990 to 2019 (see **Table 18.3**). Notice that the base year of the index is 1990, so its value for that year is 100.

TABLE 18.3 Consumer price index, Australia, 1990–2019 (1990 = 100)

Year	CPI	Year	CPI
1990	100.0	2005	149.1
1991	105.3	2006	154.4
1992	107.3	2007	158.0
1993	108.4	2008	164.8
1994	110.4	2009	167.8
1995	113.9	2010	172.6
1996	118.7	2011	178.5
1997	120.3	2012	181.4
1998	120.3	2013	185.9
1999	121.8	2014	190.4
2000	124.7	2015	192.9
2001	132.2	2016	195.1
2002	136.0	2017	198.7
2003	140.2	2018	202.6
2004	145.2	2019	205.8

Source: Australian Bureau of Statistics, various issues of ABS cat. no. 6401.0, CPI Australia, Dec 2019.

We can use this table to deflate the annual value of wages. This removes the effect of inflation, making comparisons more realistic. Based on Example 18.1, an Australian male worker earned \$1166.45 per week in 2010 and \$1475.60 per week in 2019. To determine whether his purchasing power has really increased, we deflate both values by dividing by the CPI for each year and then multiplying by 100. These calculations are performed in **Table 18.4**.

TABLE 18.4 Deflated wages, 2010 and 2019 (1990 = 100)

Year	Wage (in current \$)	CPI (1990 = 100)	Deflated wage (in 1990 \$)
2010	1166.45	172.6	$\frac{1166.45}{172.6} \times 100 = 675.81$
2019	1475.60	205.8	$\frac{1475.60}{205.8} \times 100 = 717.01$

The deflated (or real) wages are now being measured in 1990 dollars. In 1990 dollars, the worker earns more in 2019 than in 2010. As you can see, even though the wage increase in current dollars is $(1475.60 - 1166.45) = \$309.15$, the real increase is only $(717.01 - 675.81) = \$41.20$ in real terms (in 1990 dollars).

In the above analysis we expressed the wages in 1990 dollars as the CPI is with base 1990 = 100. We can also make the above comparison by expressing the wages in, for example, 2019 (or any other year) dollars. To do this, first we convert the CPI series with base 1990 = 100 to the CPI series with base 2019 = 100 and then deflate the wages as before using the new CPI series. The calculations are shown in **Table 18.5**.

TABLE 18.5 Deflated wages, 2010 and 2019 (2019 = 100)

Year	Wage (in current \$)	CPI (1990 = 100)	CPI (2019 = 100)	Deflated wage (in 2019 \$)
2010	1166.45	172.6	$\frac{172.6}{205.8} \times 100 = 83.87$	$\frac{1166.45}{83.87} \times 100 = 1390.82$
2019	1475.60	205.8	$\frac{205.8}{205.8} \times 100 = 100.0$	$\frac{1475.60}{100} \times 100 = 1475.60$

In 2019 dollars, the worker earned \$1390.82 in the year 2010. Since he actually earned \$1475.60 in 2019, we can see that our earlier conclusion remains unchanged. That is, in real terms, he earned more in 2019 than in 2010.

18.4b Using the CPI to deflate GDP

The second application of the CPI is its use as a deflator for the Gross Domestic Product (GDP). Such an application is useful in making comparisons of GDP figures over a period of time.

When the GDP of a country is measured in terms of prices prevailing at the time of measurement, it is called 'GDP (in current \$)' or 'Nominal GDP'. Therefore, to compare the GDP at different time periods, these GDP figures must be adjusted for price changes over the two periods under consideration. The adjusted values of GDP are calculated by deflating the nominal GDP by the CPI. This adjusted GDP series is called 'GDP (in constant \$)' or 'Real GDP'. That is

$$\text{Real GDP} = \frac{\text{Nominal GDP}}{\text{CPI}} \times 100$$

or

$$\text{GDP (in constant \$)} = \frac{\text{GDP(current \$)}}{\text{CPI}} \times 100$$

We will discuss the procedure using the following example.

EXAMPLE 18.8

LO7

Deflating GDP

XM18-08 The GDP is often used as a measure of the economic performance of a country. The annual GDP and CPI ($1990 = 100$) of Australia for the years 2004–19 is shown in the following table. Deflate the GDP values to 2004 dollars.

Australia's GDP, 2004–19

Year	GDP (millions of current \$)	CPI (1990 = 100)	Year	GDP (millions of current \$)	CPI (1990 = 100)
2004	890381	145.2	2012	1491046	181.4
2005	960608	149.1	2013	1524383	185.9
2006	1035295	154.4	2014	1584578	190.4
2007	1128881	158.0	2015	1609221	192.9
2008	1232583	164.8	2016	1661000	195.1
2009	1251928	167.8	2017	1764000	198.7
2010	1357034	172.6	2018	1850000	202.6
2011	1445430	178.5	2019	1890000	205.8

Source: Australian Bureau of Statistics, *Australian National Accounts*, cat. no. 5206.1, and World Bank National Accounts data.

Solution

To convert the GDP to 2004 (sometimes referred to as 'constant 2004') dollars, we will first convert the CPI ($1990 = 100$) to CPI ($2004 = 100$) and then divide the GDP by its associated CPI ($2004 = 100$). The results are shown in the following table.

Year	GDP (millions of current \$)	CPI (1990 = 100)	CPI (2004 = 100)	Deflated GDP (2004 constant \$)
2004	890381	145.2	$\frac{145.2}{145.2} \times 100 = 100$	$\frac{890381}{100} \times 100 = 890381$
2005	960608	149.1	$\frac{149.1}{145.2} \times 100 = 102.69$	$\frac{960608}{102.69} \times 100 = 935481$
2006	1035295	154.4	$\frac{154.4}{145.2} \times 100 = 106.34$	$\frac{1035295}{106.34} \times 100 = 973606$
2007	1128881	158.0	$\frac{158.0}{145.2} \times 100 = 108.82$	$\frac{1128881}{108.82} \times 100 = 1037427$
2008	1232583	164.8	$\frac{164.8}{145.2} \times 100 = 113.50$	$\frac{1232583}{113.50} \times 100 = 1085989$
2009	1251928	167.8	$\frac{167.8}{145.2} \times 100 = 115.56$	$\frac{1251928}{115.56} \times 100 = 1083313$
2010	1357034	172.6	$\frac{172.6}{145.2} \times 100 = 118.87$	$\frac{1357034}{118.87} \times 100 = 1141607$
2011	1445430	178.5	$\frac{178.5}{145.2} \times 100 = 122.93$	$\frac{1445430}{122.93} \times 100 = 1175778$
2012	1491246	181.4	$\frac{181.4}{145.2} \times 100 = 124.93$	$\frac{1491246}{124.93} \times 100 = 1193654$



Year	GDP (millions of current \$)	CPI (1990 = 100)	CPI (2004 = 100)	Deflated GDP (2004 constant \$)
2013	1 524 383	185.9	$\frac{185.9}{145.2} \times 100 = 128.03$	$\frac{1524\,383}{128.93} \times 100 = 1190\,642$
2014	1 584 578	190.4	$\frac{190.4}{145.2} \times 100 = 131.13$	$\frac{1584\,578}{131.13} \times 100 = 1208\,407$
2015	1 609 221	192.9	$\frac{192.9}{145.2} \times 100 = 132.85$	$\frac{1609\,221}{132.85} \times 100 = 1211\,295$
2016	1 661 000	195.1	$\frac{195.1}{145.2} \times 100 = 134.37$	$\frac{1661\,000}{134.37} \times 100 = 1236\,172$
2017	1 764 000	198.7	$\frac{198.7}{145.2} \times 100 = 136.85$	$\frac{1764\,000}{136.85} \times 100 = 1289\,043$
2018	1 850 000	202.6	$\frac{202.6}{145.2} \times 100 = 139.53$	$\frac{1850\,000}{139.53} \times 100 = 1325\,864$
2019	1 890 000	205.8	$\frac{205.8}{145.2} \times 100 = 141.74$	$\frac{1890\,000}{141.74} \times 100 = 1333\,469$

As you can see, the GDP in current dollars gives the impression that the economy has grown at a faster rate during the period 2004–2019, whereas when measured in 2004 constant dollars, this is not the case. Our conclusion would not change even if we used a year other than 2004 as our base.

Now we will present the solution for the opening example to this chapter.

SPOTLIGHT ON STATISTICS

How are Aussies and Kiwis performing in their earnings over time?: Solution

In order to compare the average weekly earnings for Australians over the years, we convert the 'nominal' weekly earnings to 'real' weekly earnings. To obtain the real average weekly earnings (in 2009 Australian dollars), we deflate (divide) the nominal earnings by the CPI (2009 = 100).



Source: iStock.com/Hong Li

Year	Nominal average weekly earnings (in current A\$)	CPI (2009 = 100)	Real average weekly earnings (in 2009 constant A\$)
2009	918.60	100.0	$918.60 \times \frac{100.0}{100.0} = 918.60$
2010	977.10	102.9	$977.10 \times \frac{100.0}{102.9} = 949.65$
2011	1015.20	106.3	$1015.2 \times \frac{100.0}{106.3} = 954.88$
2012	1053.20	108.2	$1053.2 \times \frac{100.0}{108.2} = 973.95$
2013	1105.00	110.8	$1105.00 \times \frac{100.0}{110.8} = 997.17$

Year	Nominal average weekly earnings (in current A\$)	CPI (2009 = 100)	Real average weekly earnings (in 2009 constant A\$)
2014	1123.00	113.5	$1123.00 \times \frac{100.0}{113.5} = 989.51$
2015	1136.90	115.0	$1136.90 \times \frac{100.0}{115.0} = 988.70$
2016	1160.90	116.8	$1160.90 \times \frac{100.0}{116.8} = 993.92$
2017	1179.00	119.1	$1179.00 \times \frac{100.0}{119.1} = 989.83$
2018	1207.40	121.3	$1207.40 \times \frac{100.0}{121.3} = 995.38$
2019	1237.90	123.4	$1237.90 \times \frac{100.0}{123.4} = 1003.16$

Source: Australian Bureau of Statistics, *Australian National Accounts*, cat. no. 5206.0; and *National Income, Expenditure and Product*.

As can be seen from the table, the average weekly earnings of Australians (nominal) in current dollars gives the impression that the average weekly earnings of Australians has grown steadily from \$918.60 in 2009 to \$1237.90 in 2019, whereas when measured in real terms (2009 constant dollars) this is not the case. There has been only a moderate growth from \$918.60 in 2009 to \$1003.16 in 2019 with falls in 2014 and 2015, and then another fall in 2017. This conclusion will not change even if we use some year other than 2009 as the base year.

In order to compare the median weekly earnings for New Zealanders over the 11 years, we convert the 'nominal' weekly earnings to 'real' weekly earnings. To obtain the real weekly earnings in 2009 NZ dollars, we deflate (divide) the nominal weekly earnings by the CPI (2009 = 100) for New Zealand.

Year	Nominal weekly earnings (in current NZ\$)	CPI (2009 = 100)	Real weekly earnings (in 2009 constant NZ\$)
2009	760	100.0	$760 \times \frac{100.0}{100.0} = 760.00$
2010	778	102.3	$778 \times \frac{100.0}{102.3} = 760.51$
2011	800	106.4	$800 \times \frac{100.0}{106.4} = 751.88$
2012	806	107.6	$806 \times \frac{100.0}{107.6} = 749.07$
2013	845	108.8	$845 \times \frac{100.0}{108.8} = 776.65$
2014	863	110.2	$863 \times \frac{100.0}{110.2} = 783.12$
2015	882	110.4	$882 \times \frac{100.0}{110.4} = 798.91$

Year	Nominal weekly earnings (in current NZ\$)	CPI (2009 = 100)	Real weekly earnings (in 2009 constant NZ\$)
2016	937	111.1	$937 \times \frac{100.0}{111.1} = 843.38$
2017	959	113.2	$959 \times \frac{100.0}{113.2} = 847.17$
2018	997	115.0	$997 \times \frac{100.0}{115.0} = 866.96$
2019	1016	116.9	$1016 \times \frac{100.0}{116.9} = 869.12$

Source: Statistics New Zealand, various years.

As can be seen, the median weekly earnings of New Zealanders (nominal) in current New Zealand dollars gives the impression that the median weekly earnings of New Zealanders has grown steadily from \$760 to \$1016 between 2009 and 2019, whereas when measured in real terms (2009 constant NZ dollars), this is not the case. There was a fall in real wages in 2011 and 2012. This conclusion will not change even if we use some year other than 2009 as the base year.

18.5 Changing the base period of an index number series

The ABS, like most government statistics departments around the world, frequently updates its index number series by changing the base period to a more recent period. For example, during the 1980s, the CPI figures in the ABS publications used 1980 as the base year. During the 1990s, the base year was changed to 1990 and, recently, it has been changed to 2012. Consequently, the two series are not comparable. Therefore, we need to construct a new comparable CPI series by combining the two incomparable series of CPIs. We will illustrate the procedure by considering two such Australian CPI series. Although, for illustrative purposes, we use the CPI series, this procedure can be used for any index number series.

EXAMPLE 18.9

LO6

Construction of a consistent CPI series

XM18-09 The following table shows the two CPI series (one for 1985–89 with base 1980 = 100, and the other for 1989–2019 with base 1990 = 100) published in various ABS publications. Create a new CPI series for the period 1985–2019 with 1990 as the base year.

► Consumer price index with different base periods, Australia

Year	Base period (1980 = 100)	Base period (1990 = 100)
1985	148.7	
1986	162.6	
1987	174.5	
1988	187.3	
1989	202.3	95.0
1990		100.0
1991		105.3
1992		107.3
...		...
...		...
2018		202.6
2019		205.8

Solution

To construct the new series for the years 1985–2019 with 1990 = 100, first we determine an overlapping year of the two series. Here we can use 1989 for this purpose.

From the two CPI series, for the overlapping year 1989,

$$\text{CPI}_{1989} = 202.3 \text{ (base year } 1980 = 100) = 95.0 \text{ (base year } 1990 = 100)$$

Therefore, for the year 1988:

$$\begin{aligned}\text{CPI}_{1988} &= 187.3 \text{ (base year } 1980 = 100) \\ &= 187.3 \times \frac{95.0}{202.3} \text{ (base year } 1990 = 100) \\ &= 87.96 \text{ (base year } 1990 = 100)\end{aligned}$$

Similarly,

$$\begin{aligned}\text{CPI}_{1987} &= 174.5 \text{ (base year } 1980 = 100) \\ &= 174.5 \times \frac{95.0}{202.3} \text{ (base year } 1990 = 100) \\ &= 81.95 \text{ (base year } 1990 = 100)\end{aligned}$$

That is, to convert the entries (with base year 1980 = 100) for the years 1985–88 to entries with base year 1990 = 100, we multiply each entry by the ratio 95.0/202.3 = 0.4696. The resulting series is shown in the last column of the following table.



Consumer price index, 1985–2019

Year	Base year 1980 = 100	Base year 1990 = 100	New series with base year 1990 = 100
1985	148.7		69.8
1986	162.6		76.4
1987	174.5		81.9
1988	187.3		88.0
1989	202.3	95.0	95.0
1990		100.0	100.0
1991		105.3	105.3
1992		107.3	107.3
...	
...	
2016		195.1	195.1
2017		198.7	198.7
2018		202.6	202.6
2019		205.8	205.8

Study Tools

CHAPTER SUMMARY

In this chapter, we looked at the basic concept of an index number, and through successive examples we have shown what they measure. In particular, we noted that *index numbers* measure the changes over time of particular time-series data. *Simple index numbers* measure changes in only a single data series, while *aggregate index numbers* measure changes in several variables. We examined *simple aggregate index numbers* and *weighted aggregate index numbers*. Two specific examples of the latter are the *Laspeyres price index* and the *Paasche price index*. The most commonly used Laspeyres price index forms the basis for the *Consumer Price Index*. We discussed how the CPI is determined in Australia, and showed how it could be used to deflate wages and GDP to facilitate comparisons. Finally, we discussed how to make consistent series of index numbers by combining two series with different base periods.

The data files for Examples, Exercises and Cases are provided on the companion website (accessible through <https://login.cengagebrain.com/>).

SUMMARY OF FORMULAS

Unweighted price index	
Simple price index	$I_{1,0} = \frac{p_1}{p_0} \times 100$
Simple aggregate price index	$I_{1,0} = \left[\frac{\sum_{i=1}^n p_{i1}}{\sum_{i=1}^n p_{i0}} \right] \times 100$
Average of relative price index	$I_{1,0} = \frac{1}{n} \left[\sum_{i=1}^n \left(\frac{p_{i1}}{p_{i0}} \right) \right] \times 100$
Weighted price index	
Weighted aggregate price index	$I_{1,0} = \left[\sum_{i=1}^n w_i \frac{p_{i1}}{p_{i0}} \right] \times 100$
Laspeyres price index	$I_{1,0}^{LP} = \left[\frac{\sum_{i=1}^n p_{i1} q_{i0}}{\sum_{i=1}^n p_{i0} q_{i0}} \right] \times 100$
Paasche price index	$I_{1,0}^{PP} = \left[\frac{\sum_{i=1}^n p_{i1} q_{i1}}{\sum_{i=1}^n p_{i0} q_{i1}} \right] \times 100$
Fisher price index	$I_{1,0}^{FP} = \sqrt{I_{1,0}^{LP} \times I_{1,0}^{PP}}$
Real GDP	$\text{Real GDP} = \frac{\text{Nominal GDP}}{\text{CPI}} \times 100$

SUPPLEMENTARY EXERCISES

- 18.1 XR18-01** Consider the price of unleaded petrol in Brisbane for the years 1985–2019, presented in the following table.

Year	Petrol price (cents/litre)
1985	50.7
1986	48.7
1987	52.3
1988	51.4
...	...
...	...
2016	119.8
2017	131.1
2018	145.3
2019	150.0

Source: Australian Bureau of Statistics, *Monthly Summary of Statistics, Queensland*, cat. no. 1304.3, and *Average Retail Price of Selected Items, Eight Capital Cities*, various years, cat. no. 6403.0, various years.

- a Taking 1985 as the base year, calculate the simple index.
- b Repeat part (a), taking 1990 as the base year.

- 18.2 XR18-02** Taking 2000 as the base year, calculate the simple index for the per adult consumption of beer in Australia for the years 2000–17, using the data in the following table.

Year	Beer consumption (litres per capita)	Year	Beer consumption (litres per capita)
2000	118.99	2009	108.35
2001	117.6	2010	107.98
2002	117.77	2011	104.94
2003	114.28	2012	100.57
2004	115.58	2013	96.51
2005	111.03	2014	93.02
2006	108.78	2015	92.38
2007	108.79	2016	86.49
2008	108.69	2017	89.43

Source: Australian Bureau of Statistics, various issues of *Apparent Consumption of Selected Foodstuffs, Australia*, cat. no. 4315.0.

- 18.3 XR18-03 Self-correcting exercise.** A cake recipe calls for the following ingredients. The prices of each ingredient in 1995 and in 2020 in Melbourne are also given below.

Ingredient	1995 price (\$)	2020 price (\$)
Butter (per 500g)	1.60	5.00
Sugar (per 2 kg)	1.80	3.00
Flour (per kg)	1.19	2.90
Eggs (per dozen)	2.19	3.20

Calculate a simple aggregate price index and an average of relative price index for the year 2020, taking 1995 as the base year.

- 18.4 XR18-04** Refer to Exercise 18.3. Suppose that a chef at Chez Henri has decided to make his own improvement by adding more butter to the cake recipe he used to make in 1995. The old and new quantities of ingredients and the prices (of the same brand) are listed below. Calculate the Laspeyres and Paasche price index numbers to measure the price increases in 2020 (with 1995 = 100).

Ingredient	1995		2020	
	Price	Quantity	Price	Quantity
Butter	\$3.20/kg	2.0 kg	\$10.00/kg	2.5 kg
Sugar	\$0.70/kg	1.0 kg	\$1.70/kg	1.0 kg
Flour	\$1.19/kg	2.0 kg	\$2.90/kg	2.0 kg
Eggs	\$2.19/doz.	1.0 doz.	\$3.20/doz.	1.5 doz.

- 18.5 XR18-05** The hotel industry is very interested to understand how tourists spend money. In order to measure the price changes in three important components of a tourist's budget, a statistician calculated the average cost of a hotel room (one night), a meal and car rental (one day) in 1985 and in 2020. The results of these calculations are shown in the following table. Calculate a simple aggregate index and the average of relative price index to reflect the costs in 2020, taking 1985 as the base year.

Component	1985 cost (\$)	2020 cost (\$)
Hotel (one night)	55	220
Meal	10	26
Car rental (one day)	25	70

- 18.6** Refer to Exercise 18.5. Suppose that in 1985, the average tourist stayed in the hotel for six days, ate eight meals at the hotel and rented a car for two days. In 2020, the average tourist stayed for four

days, ate six meals at the hotel and rented a car for three days. Calculate the Laspeyres and Paasche indexes with base 1985 = 100.

- 18.7 XR18-07** The mean weekly income of Australians for the years 1995–2019 is recorded. Calculate a simple index showing the wage increases during 1995–2019. Use 1995 as the base year. (Source: Australian Bureau of Statistics, cat. no. 6302.0, *Average Weekly Earnings, Australia*, March 2020.)

- 18.8 XR18-08** Deflate the income in Exercise 18.7 so that the mean weekly income is measured in constant 1990 dollars. Using CPI data, calculate the mean weekly income measured in constant 1995 dollars.

- 18.9 XR18-09** The owner of a service station in Sydney recorded the price (\$) and the number of units sold per day of its four most popular items (after petrol). These data were recorded for the years 1980, 1990 and 2020, as shown below.

Item	p_{1980}	q_{1980}	p_{1990}	q_{1990}	p_{2020}	q_{2020}
Oil (litre)	0.65	23	1.20	10	3.50	5
Tyres	23.00	12	55.00	15	110.00	16
Antifreeze (500mL)	0.80	7	2.00	20	5.50	14
Battery	27.00	13	45.00	22	155.00	20

- a Taking 1980 as the base year, calculate the simple aggregate index for 1990 and 2020.
- b Calculate the average of relative price index for 1990 and 2020, taking 1980 as the base year.
- c Calculate the Laspeyres index for 1990 and 2020, with base 1980 = 100.
- d Calculate the Paasche index for 1990 and 2020, with base 1980 = 100.

- 18.10 XR18-10** The following table gives the prices (in \$/litre) paid and quantities (in litres) of an Australian family's consumption of beer, wine and spirits for the years 1990 and 2020. Use the recorded information to calculate the Laspeyres, Paasche and Fisher price index numbers for alcohol for 2020 with base 1990 = 100.

Beverage	p_{1990}	q_{1990}	p_{2020}	q_{2020}
Beer	3.26	117.60	9.12	87.56
Wine	6.35	22.95	10.05	28.26
Spirits	37.47	3.80	95.74	3.30

Source: Australian Bureau of Statistics, cat. no. 4307.0.55.001, *Apparent Consumption of Alcohol, Australia*, 2020, Canberra.

- 18.11 XR18-11** Is real estate a good investment? In particular, will the value of a house keep up with inflation? To answer these questions, the median sales price of privately owned houses in a particular city council area in Sydney and the CPI (2012 = 100) for Sydney were recorded for the period from 2012(1) to 2019(3). Calculate the prices in 2015 constant dollars. What conclusions do these results indicate?

Source: Australian Bureau of Statistics, cat. no. 6401.0 Consumer Price Index, and cat. no. 6416.0, *Residential Property Price Indexes: Eight Capital Cities, Dec 2019*, Canberra, Australia.

- 18.12 XR18-12** The median sale price of three-bedroom family units in a riverside suburb of Darwin and the CPI (2012 = 100) for Darwin are recorded. Repeat Exercise 18.11 using these data.

Source: Australian Bureau of Statistics, cat. no. 6401.0 Consumer Price Index, and cat. no. 6416.0, *Residential Property Price Indexes: Eight Capital Cities, Dec 2019*, Canberra, Australia.

- 18.13 XR18-13** The data on gross domestic product for the period 2012–19 for Australia are recorded. Convert the data to constant 2012 dollars. What do these values tell you?

- 18.14 XR18-14** The shellfish catch and the wholesale price in an Australian city for 1990 and for 2020 are given in the following table.

Shellfish	Quantity (millions of kg)		Price per kg (\$)	
	1990	2020	1990	2020
Mussels	71	130	1.24	4.20
Crabs	335	300	1.09	4.10
Lobsters	30	35	1.73	5.00
Oysters	55	60	5.00	8.00
Scallops	23	10	1.65	3.00
Prawns	244	400	1.70	10.00

- a Calculate the simple aggregate index to show how prices increased from 1990 to 2020.
- b Repeat part (a) using the Laspeyres index.
- c Repeat part (a) using the Paasche index and the Fisher index.
- d What conclusions can you draw from the indexes calculated in parts (a), (b) and (c)?

- 18.15 XR18-15** Apparent consumption and average meat prices in Sydney for 2001, 2005, 2015 and 2020 are summarised in the following table.

Meat	Price/kg (\$)				Per capita annual consumption (kg)			
	2001	2005	2015	2020	2001	2005	2015	2020
Beef (rump steak)	13.00	18.00	17.89	21.00	34.9	33.0	40.0	38.0
Lamb (leg)	6.60	9.00	12.66	14.00	11.8	10.8	12.0	11.0
Pork (leg)	6.10	7.00	8.21	10.00	19.0	20.0	20.0	21.0
Chicken (frozen)	3.20	4.00	5.51	5.50	30.8	30.8	35.0	40.0

Source: Australian Bureau of Statistics, various issues of *Average Retail Prices of Selected Items and Apparent Per Capita Consumption, Australia, and Monthly Summary Statistics, NSW*.

Calculate the following price indexes for the years 2005, 2015 and 2020 with base 2001 = 100.

- a Calculate the simple aggregate index to measure the meat price increases since 2001.
- b Repeat part (a) using the average relative price index.

- c Repeat part (a) using the Laspeyres index.
- d Repeat part (a) using the Paasche index and the Fisher index.
- e Which of the indexes calculated in parts (a) to (d) most accurately measures the meat price increases since 2001? Explain your answer.

18.16 XR18-16 The average weekly (nominal) earnings of all Australian employees and the CPI (2012 = 100) for the years 1994–2019 are recorded. Find the average weekly 'real' earnings (in constant 2012 dollars) of all Australian employees.

18.17 XR18-17 Data on two annual CPI time series for Australia, one with 1981 = 100 for the period 1975–91, and the other with 1990 = 100 for the period 1990–2019 were recorded.

- a Construct a CPI series with base year 1981 = 100 for the period 1975–2019.
- b Construct a CPI series with base year 1990 = 100 for the period 1975–2019.

Case Studies

CASE 18.1 Soaring petrol prices in Australian capital cities

C18-01 The price of crude oil has been on the decline over the last few years. However, the public is of the opinion that this change is not reflected in the retail petrol prices in most Australian cities. The average annual retail petrol prices in the major Australian cities for the years 2002–18 are recorded. Using index numbers and 2002 as the base year, investigate this opinion.

CASE 18.2 Is the Australian road toll on the increase again?

C18-02 Various police departments in charge of main roads across Australia believe that strict police law enforcement on Australian roads has reduced the number of road deaths in the last few years. However, the number is on the increase again. Road deaths by gender and by state for the period 2010–19 for Australia are recorded. Analyse the data and investigate this belief.

Source: *Roads Deaths Australia monthly bulletins*, Bureau of Infrastructure, Transport and Regional Economics (BITRE), Australian Government, <https://bitre.gov.au>.

Appendix A

Summary Solutions for Selected (Even-Numbered) Exercises

Chapter 1 What is statistics?

- 1.2** Descriptive statistics consists of graphical and numerical methods used to describe sets of data, both populations and samples. Inferential statistics consists of a body of methods used for drawing conclusions about characteristics of a population, based on information available in a sample drawn from the population.
- 1.4**
- a The complete production run of light bulbs
 - b 1000 bulbs selected
 - c The proportion of the light bulbs that are defective in the whole production run.
 - d The proportion of bulbs that are defective in the sample of 1000 bulbs selected.
 - e Parameter
 - f Statistic
 - g Because the sample proportion (1%) is much less than the claimed 5%, we can conclude with confidence that there is evidence to support the claim.
- 1.6**
- a Flip the coin 100 times and count the number of heads and tails.
 - b Outcomes of repeated flips in unlimited number of trials
 - c Outcomes of the 100 flips
 - d The population proportion of heads (p) in unlimited number of trials is expected to be 0.5.
 - e The sample proportion of heads (\hat{p}) in the 100 flips
 - f If the sample proportion \hat{p} is reasonably close to the population proportion $p = 0.5$, we conclude that there is some support for the claim that the coin is a fair coin.

Chapter 2 Types of data, data collection and sampling

- 2.2**
- a Numerical
 - b Numerical
 - c Ordinal
 - d Numerical
 - e Numerical
- 2.4**
- a Numerical
 - b Nominal
 - c Nominal
 - d Ordinal
 - e Numerical
 - f Ordinal.
- 2.6**
- a Numerical
 - b Ordinal

- 2.8**
- c Nominal
 - d Numerical
 - e Ordinal
 - f Nominal.
- 2.10**
- a Numerical
 - b Ordinal
 - c Nominal
 - d Numerical
 - e Nominal
 - f Ordinal.
- 2.12**
- a Australian Bureau of Statistics; Year Book, Australia (annual); rate of unemployment, population
 - b Reserve Bank Bulletin (monthly); interest rate, exchange rate
 - c CIA Fact Book (annual); electricity consumption, flags of the world
- 2.14**
- a This is an observational study, because no attempt is made to control factors that might influence cola sales, such as store location or the way they are displayed in the shop.
 - b Randomly select which stores (both grocery and convenience) receive cola in bottles to reduce the influence of factors like store location or store type. Separately analyse the two types of stores in order to reduce the influence of store type.
- 2.16**
- a A survey can be conducted, for example, by means of a personal interview, a telephone interview, or a self-administered questionnaire.
 - b A personal interview has a high response rate relative to other survey methods, but is expensive because of the need to hire well-trained interviewers and possibly pay travel-related costs if the survey is conducted over a large geographical area. A personal interview will also probably result in fewer incorrect responses arising from respondents misunderstanding some questions. A telephone interview is less expensive, but will probably result in a lower response rate. A self-administered questionnaire is least expensive, but suffers from lower response rates and accuracy than personal interviews.
- 2.18**
- a The sampled population will exclude those who avoid large department

stores in favour of smaller shops, as well as those who consider their time too valuable to spend participating in a survey. The sampled population will therefore differ from the target population of all customers who regularly shop at the mall.

- b** The sampled population will contain a disproportionate number of thick books, because of the manner in which the sample is selected.
- c** The sampled population consists of those eligible voters who are at home in the afternoon, thereby excluding most of those with full-time jobs (or at school).

2.20 We used Excel to generate 30 three-digit random numbers. Because we will ignore any duplicate numbers generated, we generated 30 three-digit random numbers and will use the first 20 unique random numbers to select our sample.

2.22 Stratified random sampling is recommended. The strata are the school of business, the faculty of arts, the graduate school and the all the other schools and faculties would be the fourth stratum. The data can be used to acquire information about the entire campus but also compare the four strata.

- 2.24**
- a Sampling error refers to an inaccuracy in a statement about a population that arises because the statement is based only on sample data. We expect this type of error to occur because we are making a statement based on incomplete information. Nonsampling error refers to mistakes made in the acquisition of data or due to the sample observations being selected improperly.
 - b Nonsampling error is more serious because, unlike sampling error, it cannot be diminished by taking a larger sample.

2.26 Yes. A census will probably contain significantly more nonsampling errors than a carefully conducted sample survey.

Chapter 3 Graphical descriptive techniques – Nominal Data

[Solutions with tables or graphs are not provided here.]

- 3.2**
- a A bar chart would be appropriate.
 - b A pie chart would be appropriate.
- 3.4** A pie chart would be more suitable.

- 3.6** To compare the number of tourist arrivals to Fiji by country a bar chart could be used. On the other hand, to compare the share of tourist arrivals by country, a pie chart could be used.
- 3.8** To compare the exports and imports across regions in a particular year, a bar chart could be used for each year. Also, to compare the exports or imports individually between years, a bar chart could be used.
- 3.10** **a, b** Appropriate graph to compare the marriage rates at different age levels in 1997, 2007 and 2017 for males/females would be a bar chart.
c More and more males and females are waiting longer to get married.
d A bar chart would be more appropriate as the aim is to compare the rates between years and across age groups.
- 3.12** **a, b** A pie chart is more appropriate as our interest is to compare the individual country shares of total imports and exports.
- 3.14** To show the changes within a country over the three years as well as to compare the level of emissions between countries, a combined bar chart would be appropriate. To gauge the reduction in CO₂ emissions, a bar chart of percentage changes in emissions from 2009 to 2016 for the 15 countries would be appropriate.
- 3.16** A bar chart would be appropriate to compare the change in the number of tourist arrivals from 2016 to 2019. The tourist arrivals to NSW and Victoria have increased between 2016 and 2019. Tourist arrivals to other states and territories have remained nearly the same in 2016 and 2019.
- 3.18** (a) If the focus is on the actual numbers who prefer each type of wine, a bar chart would be useful.
(b) A pie chart would be useful to show the share of each category.
- 3.20** A bar chart would be appropriate.
- 3.22** A bar chart would be appropriate.
- 3.24** A bar chart would be appropriate.
- 3.26** The charts show that a majority of applicants are BA graduates, capturing 88 (38%) of the applicants, followed by BEng 51 (22%), then BBus 37 (16%) and BSc 24 (11%).
- 3.28** **a** A bar chart would be appropriate to depict the frequency distribution.
b A pie chart would be appropriate to depict the proportions.
- 3.30** A pivot bar chart by type of worker would be more appropriate.
- 3.32** Constructing a cross-classification table (in percentages) or pivot table (in percentages) and pivot chart would give the information required. There does not appear to be any brand loyalty.
- 3.34** **a** Using Excel, count the frequencies for each software.
b A pie chart would be appropriate to depict the proportions.
c Excel is the choice of about half the sample, one-quarter Minitab, and a small fraction chose SAS and SPSS.
- 3.36** **a** Depict the amount of imports in a bar chart.
b Depict the share of imports in a pie chart.
c China (18%), US (12.3%) and Korea (7.3%) are Australia's top 3 import markets.
- 3.38** A combined bar chart by industry would be appropriate. In the manufacturing and construction, and commercial sectors, energy consumption increased during 2010/11 to 2012/13 and decreased during 2012/13 to 2016/17. In all other sectors, energy consumption continued to increase from 2010/11 to 2016/17.
- 3.40** A bar chart would be appropriate to compare the number of arrivals from the top 10 tourism markets.
- 3.42** A pie chart would be more appropriate. Among the statistics instructors surveyed, 44% use the computer approach, 38% use a combined manual and computer approach, whereas only 18% emphasise manual calculations approach.
- 4.8** Class intervals, $40 < X \leq 50$, $50 < X \leq 60$, ..., $90 < X \leq 100$, frequencies F: 2, 6, 9, 7, 4, 2; Relative F: 0.07, 0.20, 0.30, 0.23, 0.13, 0.07; Cumulative Relative F: 0.07, 0.27, 0.57, 0.80, 0.93, 1.00.
- 4.10** **a, d & e** Class intervals, $16 < X \leq 24$, $24 < X \leq 32$, ..., $48 < X \leq 56$, frequencies F: 1, 8, 9, 5, 2; Relative F: 0.04, 0.32, 0.36, 0.20, 0.08; CF: 1, 9, 18, 23, 25. CRF: 0.04, 0.36, 0.72, 0.92, 1.0.
- 4.12** **a** Class intervals, under 400, $400 < X \leq 800$, $800 < X \leq 1200$, ..., $4000 < X \leq 5000$, Over 5000. Relative F: 0.119, 0.194, 0.148, 0.127, 0.109, 0.104, 0.067, 0.076, 0.027, 0.029; Cumulative Relative F: 0.119, 0.313, 0.461, 0.588, 0.697, 0.801, 0.868, 0.944, 0.971, 1.0.
c Annual income \$62400. Weekly income \$1200. Corresponding CRF = 0.461. Therefore, 46.1% of the annual incomes were less than \$62400.
d 50% of the weekly income were less than approximately \$1250.
- 4.14** **a** Class intervals, under 25, $25 < X \leq 30$, $30 < X \leq 35$, ..., $65 < X \leq 70$,
b. The histogram is unimodal and positively skewed.
- 4.16** **b** Bins: 20, 25, 30, 35, 40; Frequency: 9, 34, 91, 61, 5
c The distribution of the annual incomes of the recently-graduated business graduates is approximately bell-shaped, unimodal, with the modal class consisting of incomes between \$25000 and \$30000.
e **(i)** The proportion of recently-graduated business graduates who earn less than \$20000 is 0.045
(ii) The proportion who earns more than \$35000 is 0.025.
(iii) The proportion who earn between \$25000 and \$40000 is 0.785.
- 4.18** **b** The data are skewed to the right.
d About 27% of the house prices are less than \$400 000.
e About 84% of the house prices are less than \$550 000.
- 4.20** **b** Except for the 12:00–1:00pm hour, other distributions are unimodal and roughly bell-shaped.
c The bar graph shows that

Chapter 4 Graphical descriptive methods – Numerical data

[Solutions with tables or graphs are not provided here.]

4.2 Number of classes $K = 1 + 3.3\log_{10}(1500) = 11.48$. Use 11 to 12 classes

4.4 Number of classes $K = 1 + 3.3\log_{10}(40) = 6.3$. Use 6 or 7 classes. Class width $d = (6.1-5.2)/6 = 0.2$. Choose class intervals $5.0 < X \leq 5.2$, $5.2 < X \leq 5.4$, ..., $6.0 < X \leq 6.2$.

4.6 **a** Class intervals: 5–10, 10–15, 15–20, 20–25, 25–30; Frequency: 5, 3, 9, 7, 1; Relative Frequency: 0.20, 0.12, 0.36, 0.28, 0.04

c The area of each rectangular strip is

- 12:00–1:00pm attracts the most number of customers followed by 2:00–3:00pm, while 11am–12pm and 1:00–2:00pm attract the least.
- d** Use 11:00am–12:00pm and 1:00–2:00pm for tea/lunch breaks for the bank employees.
- 4.22** The histogram is unimodal and positively skewed.
- 4.24** The histogram of the number of books shipped daily is negatively skewed. It appears that there is a maximum number that the company can ship.
- 4.26** The histogram is unimodal, symmetric and bell-shaped.
- 4.28** **c, d** The new scorecards are relatively good predictors.
- 4.30** **c** The proportion of indigenous Australian population is about 3%, while the proportion of deaths in custody of indigenous Australians is more than 10% and sometimes as high as 27%.
- 4.32** **b** The line graph shows the trend in movements of the exchange rate over time.
- 4.34** Except for a slight decline in 1983, 1990 and 2009, the quarterly GDP series has been steadily increasing throughout 1959 to 2019.
- 4.36** For the two time-series data, a line chart would be appropriate. The charts show that more homes are sold than apartments. The patterns of sale of both types seem to be similar.
- 4.38** A positive linear model is well suited. The estimated line is $\hat{y} = 0.933 + 2.114x$, $R^2 = 0.491$
- 4.40** Sales levels seem to have a positive linear relationship with advertising expenditure.
- 4.42** **b** A positive linear relationship exists between hours of machine usage (x , '000) and electricity cost (y , \$). **d** The least squares line of fit is $\hat{y} = 404.8 + 62.368x$.
- 4.44** **b** A positive linear relationship exists between hours of Internet use and years of education. $\hat{y} = -11.3 + 1.68x$ and $R^2 = 0.496$.
- 4.46** **b** Overall, it seems that there is a positive linear relationship between gold price and silver price. $\hat{y} = -0.13 + 0.0168x$, $R^2 = 0.85$.
- 4.48** **a** A scatter diagram with age on the horizontal x -axis and the number of hours of Internet use on the vertical y -axis would be appropriate. **b** There appears a strong negative linear relationship. The older the person, the lesser the Internet use.
- Internet use $\hat{y} = 46.4 - 0.435 \text{Age}$, $R^2 = 0.56$.
- 4.50** A bubble chart is more appropriate.
- 4.52** A heatmap is more appropriate.
- 4.54** **c** The data are slightly skewed to the right and is bimodal.
- 4.56** **a** Line graphs for male and female on the same plot would be appropriate. **b** Line graphs are the appropriate plot for time series data.
- 4.58** **b** It appears that prices of houses have increased and less dispersed this year compared to the house prices 5 years ago.
- 4.60** Most of the IQ is somewhere between 80 and 110. The distribution is slightly symmetrical.
- 4.64** A time series line graph of number of cigarettes per capita against year would be appropriate to show the decline in per capita cigarettes consumption.
- 5.2** **a** Mean = 46.67; Median = 42.5; Mode = 40
- 5.4** **a** Mean = \$1165000; Median = \$970000; Mode = \$970000
b Mean > Median = Mode. Therefore, house price is skewed to the right. Median is the best measure to represent the house prices.
- 5.6** Mean = 8.8; weighted mean = 9.2.
- 5.8** Average compounding rate of return is $[(1.25)(0.90)(1.50)]^{(1/3)} - 1 = 0.1906$ or 19.06%.
- 5.10** Mean = 0.044, Median = -0.08; Average compounding rate of return = 0.015 or 1.5%.
- 5.12** **a** Mean = 72, median = 72, mode = N/A
b Mean = 82, median = 73, mode = N/A
c The outlier made a significant difference to the mean, but not to the median
- 5.14** **a** Mean = 975.5, median = 856.
b Right skewed as mean > median > mode.
c Mean = 862.8, median = 856.
- 5.16** **a** Mean = 69.41;
b Median = 72.35
c Mean = 75.53 and median = 73.7
- 5.18** Unweighted mean = 3.19%.
- 5.20** **a** Years, 1–6: -0.167, 0.40, 0.071, 0.467, 0.364, -0.167
b 0.161; Median = 0.218
c 0.130
- d** Geometric mean is best because $12(1.130)^6 = 25$.
- 5.22** **a** Mean = 5.85 mins, median = 5 mins, mode = 5 mins.
b The average of all the times was 5.85 mins. Half the times were less than 5 mins and half were greater. The time most frequently taken was 5 mins. A few larger observations have pulled the mean upward.
- 5.24** **a** Mean = \$2432.88; Median = \$2446.10
b The distribution is reasonably symmetrical. But a few low incomes have pulled the mean below the median, resulting in a distribution slightly skewed to the left.
c Either measure could be used, but the median is better as it is not affected by the few low incomes.
- 5.26** **a** Mean = \$475910; Median = \$435000
b Mean > Median, the house prices distribution may be skewed to the right.
- 5.28** **a** Mean = 31.66 seconds; Median = 32 seconds; Mode = 33 seconds
b All three measures in part (a) are approximately equal. The distribution of times is therefore approximately symmetrical with a single mode of 33. Half the times are less than 32 seconds.
- 5.30** **a** Mean = \$47194.60; Median = \$47353.20
c Mean < Median, the salaries distribution may be slightly skewed to the left.
- 5.32** **a** No, a standard deviation cannot be negative.
b Yes, a standard deviation is larger than its corresponding variance when the variance is between 0 and 1.
c Yes, when every value of a data set is the same, the variance and standard deviation will be zero.
- 5.34** **a** $\bar{x} = 9$; $s^2 = 12.5$; $s = 3.54$; $cv = 39.3\%$
b $\bar{x} = 0$; $s^2 = 4.67$; $s = 2.16$; $cv = \text{not applicable as } \bar{x} = 0$
c $\bar{x} = 6$; $s^2 = 5.33$; $s = 2.31$; $cv = 38.5\%$
d $\bar{x} = 5$; by inspection, $s^2 = 0$ and $s = 0$; $cv = 0\%$
- 5.36** Range = 6 hours, $\bar{x} = 3$ hours, $s^2 = 4.67$ (hours)², $s = 2.16$ hours, $cv = 72\%$
- 5.38** **a** Mean $\bar{x} = 69.07$; Median = 66.5.
b The modal class of the frequency distribution is '60 up to 70'. The mode is therefore 65.
d $\sum x_i^2 = 147950$; $\sum x_i = 2072$; $s^2 = 167.03$

- 5.40** **a** Mean = 1.06, $s = 0.60$;
b Median = 0.97;
c Mean = 1.00, $s = 0.14$,
Median = 0.97

- 5.42** **a** About 68%;
b About 95%;
c About 99.7%

- 5.44** **a** $\mu = 5.67\%$; $\sigma^2 = 277.16(\%)^2$;
 $\sigma = 16.65\%$;
b Range = 70%; Median = 5%.

- 5.46** **a** $\bar{x} = 2.55$; $s^2 = 0.03945$; $s = 0.20$
b Using the range approximation,
 $s = 0.2$;
We have assumed that the
measurements have a mound-
shaped distribution.

- 5.48** **a** Portfolio: 13.7, -1.0, 17.15, 8.25,
34.20, 37.60, 24.80, 11.45, 4.40,
24.45
b Mean return = 17.5%;
c Standard deviation = 12.63%;
d $cv = 72.2\%$;
e $\bar{x}_A = 20\%$, $s_A = 16.7\%$, $cv_A = 83.7\%$;
 $\bar{x}_B = 15\%$, $s_B = 10.0\%$, $cv_B = 66.5\%$;
 $\bar{x}_P = 17.5\%$, $s_P = 12.6\%$, $cv_P = 72.2\%$;
Fund A: Highest risk; Fund B: Lowest
risk.

5.50 a-b

	Mean	Median	Range	s	cv
1-year (4*)	9.87	9.82	5.11	1.82	18.39
3-year (4*)	5.77	6.17	2.72	1.10	19.10
1-year (3*)	10.16	10.44	2.42	0.93	9.17
3-year (3*)	5.69	5.86	1.52	0.63	11.11

- c** Among the 1 year investments (i)
4-star (ii) 3-star.
d Among the 4-star rated investments,
1 year, high return and low risk (cv)
e Among the 3-star rated investments,
1 year, high return and low risk (cv)

- 5.52** Range = \$411; $s^2 = 9292.41$ (\$) 2 ;
 $s = \$96.40$; $cv = 40.5\%$

- 5.54** **a** $(\bar{x}, s) = 10\text{am}-11\text{am}: (102.22, 16.07)$;
 $11\text{am}-12\text{pm}: (70.26, 10.58)$; $12\text{pm}-$
 $1\text{pm}: (177.93, 18.24)$, $1\text{pm}-2\text{pm}: (65.87, 9.37)$;
 $1\text{pm}-2\text{pm}: (147.92, 14.63)$

- b** The noon hour (12pm–1pm) is the
busiest, followed by the (2pm–3pm)
and (10am–11am) periods. Staff
lunch breaks and coffee breaks
should be scheduled with this in
mind.

- 5.56** $s^2 = 40.73\text{kmph}^2$ and $s = 6.38\text{kmph}$;
at least 75% of the speeds lie within
12.76kmph (2s) of the mean; at
least 88.9% of the speeds lie within
19.14kmph (3s) of the mean.

- 5.58** $s^2 = 0.0858\text{cm}^2$, and $s = 0.2929\text{cm}$;
at least 75% of the lengths lie within
0.5858cm of the mean; at least 88.9%
of the rods will lie within 0.8787cm of
the mean.

- 5.60** **a** $s = 15.01$
b Approximately 68% of the hours, the
number of arrivals falls between 83
(rounded from 83.04) and 113; 95%
of the hours, the number of arrivals
falls between 68 and 128; 99.7% of
the hours, the number of arrivals fall
between 53 and 143.

- 5.62** $n = 10$; Data (in ascending order): 15, 20,
22, 23, 24, 26, 29, 30, 31, 31;
 $L_{30} = 3.3$, $P_{30} = 22.3$; $L_{80} = 8.8$,
 $P_{80} = 30.8$.

- 5.64** $L_{25} = 3.5$, $Q_1 = P_{25} = 13.05$; $L_{50} = 7$,
 $Q_2 = P_{50} = 14.7$; $L_{75} = 10.5$,
 $Q_3 = P_{75} = 15.6$.

- 5.66** Interquartile range = 7 – 3 = 4.

- 5.68** $n = 10$, $L_{25} = 2.75$, $Q_1 = P_{25} = 5.75$;
 $L_{75} = 8.25$, $Q_3 = P_{75} = 15$; Interquartile
range = 9.25

- 5.70** **a** Min (S) = 9, $Q_1 = 13$; $Q_2 = 20$;
 $Q_3 = 24$; Max (L) = 31.

- 5.72** **a** Min (S) = -2.2, $Q_1 = 5.65$; $Q_2 = 18.6$;
 $Q_3 = 35.275$; Max (L) = 46.9.
Min (S) = 0.2, $Q_1 = 7.475$; $Q_2 = 14.75$;
 $Q_3 = 22.975$; Max (L) = 38.

- b** The median return for Fund A
exceeds that for Fund B. The returns
for Fund A are more variable than
for Fund B, with Fund A having an
IQR of 29.6% and a range of 49.1%,
compared with an IQR of 15.5% and
a range of 37.8% for Fund B. Neither
fund has any outliers.

- 5.74** **a-d** $n = 100$, $L_{25} = 25.25$, $Q_1 = 64.25$;
 $L_{50} = 50.5$, $Q_2 = 81.0$; $L_{75} = 75.75$,
 $Q_3 = 90$.

- Smallest = 11; $Q_1 = 64.25$; $Q_2 = 81$;
 $Q_3 = 90$; Largest = 100; IQR = 25.75;
Outliers: 11, 16, 18, 25. Even though
the marks range from 11 to 100, half
of them are over 81. The distribution
is highly skewed to the left. The
mean mark of about 74 (from
Example 5.6) is less than the median
due to the four small value outliers.

- 5.76** Based on the median salary, "Other
degrees" graduates received better
salary offers followed by BSc, 'BBA' and
then BA graduates.

- 5.78** **a** Box plot (Public). Smallest = 238; Q_1 =
279; $Q_2 = 296$; $Q_3 = 307$; Largest =
359; IQR = 28; Outlier: 359.
Box plot (Private). Smallest = 213;
 $Q_1 = 228$; $Q_2 = 237$; $Q_3 = 245.75$;
Largest = 260; IQR = 17.75; No
outliers.

- b** Times for public course is skewed
to the right with one large outlier
value. The amount of time taken
to complete rounds on the public
course is greater and more variable
than those played on private courses.

- 5.80** Box plot. Smallest = 0; $Q_1 = 50$; Median
= 125; $Q_3 = 260$; Largest = 1960; IQR =
210; Outliers: 1960, 1750, 1560, 1240,
1150, 1080, 840, 830, 690; The amounts
are positively skewed.

- 5.82** **a** Points are scattered around an
upward sloping straight line.
b $s_{xy} = 98.52$.
c $s_x^2 = 57.79$, $s_y^2 = 216.3$, $r = 0.8811$.
d $R^2 = r^2 = 0.7763$.
e $\hat{y} = -0.603 + 1.705x$.

- f** A strong positive linear relationship,
and for every additional hour of study
time, marks increased on average by
1.705.

- 5.84** **a** The graph shows a positive linear
relationship between the cost of
electricity (Y) and the machine
time (X).

- b** $s_{xy} = 215.45$; $r = 0.9527$.
c As r is positive and close to 1, there
is a strong positive linear relationship
between the two variables.
d Cost = $404.8 + 62.368$ (Hours); Fixed
cost = \$404.80 and variable cost =
\$62.37/hour.

- 5.86** $R^2 = 0.4009$; 40.09% of the variation in
the employment rate is explained by
the unemployment rate.

- 5.88** **a** $\text{cov}(x, y) = 20.55$.
b $r = 0.4437$.
c There is a weak positive linear
relationship between smoking and
the incidence of cold.

- 5.90** **a** Mean $\bar{x} = 35$, Median = 36;
b $s = 7.68$.
c Average bone density loss of women
aged 50 and over is 35 and half of
the bone density losses lie below 36.
At least 75% of the observations lie
between 19.64 and 50.36, at least
88.9% of the numbers lie between
11.96 and 58.06.

- 5.92** $\bar{x} = 17$, Median = 18, Mode = 19 and 21,
 $s^2 = 38.5$, $s = 6.20$

- 5.94** **a** $\bar{x} = 7.61$, $s = 1.21$.
c Median = 7.53

- 5.96** **a** $\bar{x} = 47.6$ yrs.
b $s^2 = 115.42$ (yrs) 2 .
c $s = 10.74$ yrs.
d Range = 42 yrs.
e Approximate $s = 10.5$ which is close
to its actual value $s = 10.74$ obtained
in part c.

5.98 b

	\bar{x}	Med	Mode	Min	Max	s
5 yrs ago	3.36	3	3	0	9	1.79
Today	1.56	1	1	0	6	1.25

- c The average number of defects today is 1.56, less than half the average number (3.36) five years ago. The most frequent number of defects today is 1, down from 3 five years ago. The maximum number of defects today is 6, down from 9 five years ago.
- 5.100** a $\bar{x} = 5.23$; $s = 2.36$.
c Median = 5.29. d Approximate $\bar{x} = 5.27$, $s = 2.45$. The actual (from part a) and approximate mean and standard deviation are reasonably close.
- 5.102** a $\bar{x} = 3.37\%$; $s = 1.24\%$.
c Median = 3.35%.
- 5.104** Dogs cost more money than cats. Both sets of expenses are positively skewed.
- 5.106** a $\bar{x} = 26.32$ hrs, Median = 26 hrs.
b $s^2 = 88.57$, $s = 9.41$ hrs.
d The times are positively skewed. Mean time is 26.32 hrs and half the times are above 26 hrs.
e $s_{xy} = 11.598$; $r = 0.64$.
f Internet use and level of education are positively related and the strength is moderately strong.

Chapter 6 Probability

- 6.2** a $S = \{a \text{ is correct}, b \text{ is correct}, c \text{ is correct}, d \text{ is correct}, e \text{ is correct}\}$.
b $P(a \text{ is correct}) = P(b \text{ is correct}) = P(c \text{ is correct}) = P(d \text{ is correct}) = P(e \text{ is correct}) = 0.2$.
c Classical approach.
d In the long run all answers are equally likely to be correct.
- 6.4** a The random experiment consists of observing which two applicants receive job offers.
b Let AB denote the simple event that Anne and Bill receive job offers. Similar notations are used for the other simple events: $S = \{AB, AC, AD, BC, BD, CD\}$.
c $L = \{AC, BC, CD\}$; $M = \{AC, AD, CD\}$; $N = \{AB, AC, AD, BC, CD\}$
- 6.6** a A: Exactly one head is observed; B: At least one head is observed. C: Two tails (or no heads) are observed.
b $A \cup B = \{HT, TH, HH\}$
c $A \cap B = \{HT, TH\}$
d $\bar{A} = \{HH, TT\}$

e A and C are mutually exclusive, since $A \cap C$ contains no simple events. B and C are mutually exclusive, since $B \cap C$ contains no simple events.

- 6.8** $P(S) = 1$, $P(A) = 0.17$, $P(B) = 0.5$, $P(C) = 0.5$, $P(D) = 0.33$.

- 6.10** a $S = \{AB, AC, AD, BC, BD, CD\}$;
 $P(AB) = P(AC) = P(AD) = P(BC) = P(BD) = P(CD) = 0.17$.
b $P(AC, BC, CD) = 0.5$.
c $P(AB, AD, BC, CD) = 0.67$.
d $P(AC, AD, CD) = 0.5$.
e $P(AB, AC, AD, BC, CD) = 0.83$.

- 6.12** a $S = \{BF, WF, MF, B\bar{F}, W\bar{F}, M\bar{F}\}$.
b $F = \{BF, WF, MF\}$.
c $P(B) = 0.65$, $P(W) = 0.25$, $P(M) = 0.10$, $P(F) = 0.55$, $P(\bar{F}) = 1 - P(F) = 0.45$.
d $P(\bar{M}) = 1 - P(M) = 0.9$.

- 6.14** a i $P(L) = 0.37$; ii $P(M) = 0.28$; iii $P(N) = 0.35$; iv $P(C) = 0.57$; v $P(\bar{C}) = 0.43$
b $P(M \cup N) = P(M) + P(N) = 0.63$

- 6.16** a The random experiment is to select a shareowner at random
b $S = \{\text{Individual share owners in each state}\}$ and $n(S) = 7777$
c $P(NSW) = 0.3492$; $P(VIC) = 0.2640$; $P(QLD) = 0.1828$; $P(SA) = 0.0586$; $P(WA) = 0.1098$; $P(TAS) = 0.0120$; $P(NT) = 0.0076$.
d Relative frequency approach
e $P(\text{Number of share owners} > 1,000,000) = 0.3492 + 0.2640 + 0.1828 = 0.7960$

- 6.18** a $P(A_1) = 0.3$, $P(A_2) = 0.4$, $P(A_3) = 0.3$;
 $P(B_1) = 0.6$, $P(B_2) = 0.4$.
b $P(A_1 | B_1) = 0.17$.
c $P(A_2 | B_1) = 0.50$.
d $P(A_3 | B_1) = 0.33$.
e $P(A_1 | B_1) + P(A_2 | B_1) + P(A_3 | B_1) = 1.0$

Yes. It is not a coincidence. Given B_1 , the events A_1 , A_2 , A_3 constitute the entire sample space.

- 6.20** $P(A_1 | B_1) = 0.25$; $P(A_1) = 0.25$;
 $P(A_2 | B_1) = 0.75$; $P(A_2) = 0.75$;
 $P(A_2 | B_2) = 0.75$; $P(A_2) = 0.75$;
 $P(A_1 | B_2) = 0.25$; $P(A_1) = 0.25$.

Therefore, the events are independent.

- 6.22** a $P(A_1) = 0.15 + 0.25 = 0.40$,
 $P(A_2) = 0.20 + 0.25 = 0.45$,
 $P(A_3) = 0.10 + 0.05 = 0.15$.
 $P(B_1) = 0.15 + 0.20 + 0.10 = 0.45$,
 $P(B_2) = 0.25 + 0.25 + 0.05 = 0.55$.
b $P(A_2 | B_2) = 0.4545$.
c $P(B_2 | A_2) = 0.5556$.
d $P(B_1 | A_2) = 0.4444$.
e $P(A_1 \text{ or } A_2) = P(A_1) + P(A_2) = 0.40 + 0.45 = 0.85$.
f $P(A_2 \cup B_2) = P(A_2) + P(B_2) - P(A_2 \cap B_2) = 0.45 + 0.55 - 0.25 = 0.75$

g $P(A_3 \cup B_1) = P(A_3) + P(B_1) - P(A_3 \cap B_1) = 0.15 + 0.45 - 0.10 = 0.50$

- 6.24** a $P(\text{promoted} | \text{female}) = 0.20$
b $P(\text{promoted} | \text{male}) = 0.20$
c $P(\text{promoted}) = 0.20 = P(\text{promoted} / \text{female}) = P(\text{promoted} / \text{male})$.
Therefore, promotion and gender are independent events. No, because promotion and gender are independent events.

- 6.26** a $P(\text{He or she works for the construction industry}) = 0.174$
b $P(\text{Works in the transport and communication industry} | \text{Female}) = 0.065$
c $P(\text{Male} | \text{Information and Media}) = 0.568$

6.28 A: balance under \$100; B: balance over \$500. C: balance is \$500 or less; D: account is overdue.

- a $P(A | D) = 0.4$
b $P(D | B) = 0.2$
c $P(D | C) = 0.2$

- 6.30** a F and M are independent events since $P(F | M) = 0.55 = P(F)$
b F and B are dependent events since $P(F | B) = 0.52 \neq P(F) = 0.55$

- 6.32** a $P(\text{customer will return and rate the restaurant's food as Good}) = 0.35$
b $P(\text{customer rates the restaurant's food as good} | \text{say will return}) = 0.538$
c $P(\text{Customer says will return} | \text{customer rates the restaurant's food as good}) = 0.714$
d (a) is the joint probability and (b) and (c) are conditional probabilities.

- 6.34** Number of Australians ('000) = 20561
a $P(\text{Unemployed}) = 0.034$
b $P(\text{Female}) = 0.509$; $P(\text{Unemployed and female}) = 0.016$
 $P(\text{Unemployed} | \text{Female}) = 0.031$;
c $P(\text{Male}) = 0.491$;
 $P(\text{Unemployed and Male}) = 0.018$;
 $P(\text{Unemployed} | \text{Male}) = 0.037$

- 6.36** a $P(\text{Manual} | \text{Stats}) = 0.390$
b $P(\text{Computer}) = 0.66$
c No, because $P(\text{Manual}) = 0.34$, which is not equal to $P(\text{Manual} | \text{Stats})$.

- 6.38** a $P(\text{Ask} | \text{Male}) = 0.24$
b $P(\text{Consult a map}) = 0.39$
c No, because $P(\text{Consult map} | \text{Male}) = 0.50 \neq P(\text{Consult map}) = 0.39$

- 6.40** a $P(\text{Under 20}) = 0.8309$
b $P(\text{Retail}) = 0.5998$
c $P(20 \text{ to } 99 | \text{construction}) = 0.0751$.

- 6.42** $P(A \cap B) = 0.2$; $P(A | B) = 0.4$

- 6.44** a $P(A \cap B) = 0.12$
b $P(A \cup B) = 0.78$

- 6.46** **a** $P(A \cup B) = 0$ since A and B are mutually exclusive.
b $P(A \cup B) = 0.85$
c $P(A|B) = 0$
- 6.48** **a** $P(A \cap B) = 0$
b $P(A \cup B) = 0.55$
c $P(B|A) = 0$
- 6.50** Define the events. A : The company wins contract A . B : The company wins contract B .
 $P(A) = 0.6$; $P(B) = 0.3$; $P(A|B) = 0.8$
a $P(A \cap B) = 0.24$
b $P(A \cup B) = 0.66$
c $P(\bar{A}|B) = 1 - P(A|B) = 0.2$
- 6.52** Define the event. T : A student plays tennis. C : A student plays cricket.
 $P(T) = 0.10$, $P(C) = 0.05$, $P(C|T) = 0.40$.
a $P(T \cap C) = 0.04$
b $P(\bar{T} \cap \bar{C}) = P(\bar{T} \cup \bar{C}) = 1 - P(T \cup C) = 0.89$
- 6.54** Define the events: A_1 : Four spots turn up on the first die. A_2 : Four spots turn up on the second die.
 $P(A_1 \cup A_2) = 1 - P(\bar{A}_1 \cup \bar{A}_2) = 1 - P(\bar{A}_1 \cap \bar{A}_2)$
 $= 1 - P(\bar{A}_1) \cdot P(\bar{A}_2) = 0.31$
- 6.56** $P(A \cap B) = P(B|A) \cdot P(A) = 0.21$
- 6.58** $P(A \cap B) = 0.32$; $P(A \cap \bar{B}) = 0.48$; $P(\bar{A} \cap B) = 0.14$; $P(\bar{A} \cap \bar{B}) = 0.06$;
- 6.60** **b** $P(A \cap B) = 0.06$; $P(\bar{B}) = 0.66$
- 6.62** **b** $P(A \cup B) = 0.84$; $P(\bar{A} \cap B) = 0.24$
- 6.64** $P(\text{Neither defective}) = 0.47$
- 6.66** **a** $P(\text{No sale}) = 0.70$
b $P(L \cup G) = 0.18$
- 6.68** **a** $(0.18)(20000) = 3600$
b $P(B) = 0.41$
c $P(A) = 0.34$
d $P(B \cup C|W) = 0.805$
- 6.70** $P(\text{Increase}) = 0.6125$
- 6.72** $P(A|B) = 0.5$, $P(A|\bar{B}) = 0.63$
- 6.74** $P(A|B) = 0.125$, $P(\bar{A}|\bar{B}) = 0.67$
- 6.76** $P(A|R) = 0.16$, $P(\bar{A}|R) = 0.84$
- 6.78** **a** $P(B) = 0.346$, $P(A|B) = 0.246$
b $P(B) = 0.514$, $P(A|B) = 0.661$
- 6.80** $P(\text{PT}) = 0.6397$, $P(\text{NT}) = 0.3603$, $P(C|\text{PT}) = 0.0256$, $P(C|\text{NT}) = 0.0010$
- 6.82** $n(S) = 36$.
a $P(A) = 2/36$
b $P(B) = 10/36$
c $P(C) = 6/36$
d $P(D) = 5/36$
e $P(E) = 18/36$
f $P(D|F) = 5/18$
- 6.84** **a** $P(\bar{S}) = 0.71$
b $P(\bar{M} \cap \bar{S}) = 0.16$
c $P(\bar{S}|M) = 0.69$
- 6.86** $P(\bar{B}) = 0.9097$; $P(A_i|\bar{B}) = 0.2031$
- 6.88** A: Salary over \$35 000. B: Salary under \$50 000.

- 6.90** **a** $P(WWW) = (0.03)^3 = 0.000027$
b $P(\text{Exactly one winner}) = P(W\bar{W}\bar{W}) + P(\bar{W}W\bar{W}) + P(\bar{W}\bar{W}W) = 0.084681$
c $P(\text{at least one winner}) = 1 - P(\bar{W}\bar{W}\bar{W}) = 0.087327$
- 6.92** F: The firm fails. B: The model predicts bankruptcy.
 $P(B|F) = 0.85$, $P(\bar{B}|\bar{F}) = 0.82$, $P(F) = 0.04$.
a The model's prediction is correct if $(B \cap F) \cup (\bar{B} \cap \bar{F})$ occurs.
 $P(B \cap F) \cup (\bar{B} \cap \bar{F}) = 0.8212$
 Hence, the model's prediction will be correct for 82 firms out of 100 firms.
b $P(F|B) = 0.164$ as $P(F \cap B) = 0.034$ and $P(B) = 0.2068$
- ### Chapter 7 Random variables and discrete probability distributions
- 7.2** **a** $\{0, 1, 2, 3, \dots\}$
b Yes;
c Yes;
d Discrete.
- 7.4** **a** $P(X > 0) = 0.5$; **b** $P(X \leq 0) = 0.5$
c $P(0 \leq X \leq 1) = 0.7$ **d** $P(X = -2) = 0$
e $P(X = -4) = 0.2$; **f** $P(X < 2) = 0.9$
- 7.6**

x	1	2	3	4	5	6
p(x)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
- 7.8** **a** $X = \text{Number of persons in different living arrangements in Australia. } P(X = 1) = 0.4913$, $P(X = 2) = 0.2116$, $P(X = 3) = 0.0080$, $P(X = 4) = 0.0363$ and $P(X = 5) = 0.2527$.
 The distribution is skewed to the right (or positively skewed)
b The most likely category of living arrangement in Australia is couple with one child.
c $P(\text{One parent female}) = 0.0363$
d $P(\text{Couple with children family} | \text{Couple}) = 0.6990$
- 7.10** **b** The distribution is slightly skewed to the right.
c $P(2 \leq X \leq 4) = 0.870$
d $P(X > 5) = 0.020$
e $P(X \leq 6) = 0.999$
f $P(X > 1) = 0.970$
- 7.12** **a** $P(4 \text{ books}) = 0.06$
b $P(8 \text{ books}) = 0$
c $P(\text{No books}) = 0.35$
d $P(\text{At least 1 book}) = 0.65$
- 7.14** **a** $P(X > 4) = 0.32$
b $P(X < 5) = 0.68$
c $P(4 \leq X \leq 6) = 0.47$
- 7.16** **a** $E(X) = 15.75$, $V(X) = 33.1875$
b $E(Y) = 60$
c $V(Y) = 531$
- 7.18** **a** $\mu = 1.7$, $\sigma^2 = 0.81$, $\sigma = 0.9$
b $P(X \geq 2) = 0.6$
c $P(X \geq 1.5) = 0.6$
- 7.20** Let X represent the amount of damage incurred.
a $E(X) = 18$. The owner should be willing to pay up to \$18.
b $V(X) = 3076$, $\sigma = \$55.46$
- 7.22** **a** $E(X) = 1.85$
b $V(X) = 1.0275$
- 7.24** **a** $E(X) = \$950$, $V(X) = 322500$
b $E(Y) = \$475$, $V(y) = 80625$
- 7.26** $E(X) = \$0.70$
- 7.28** $E(X) = 4$, $V(X) = 2.40$, $Y = 0.25X$; $E(Y) = 1$, $V(Y) = 0.15$
- 7.30** **a** $P(X = 2) = 0.9619 - 0.8131 = 0.1488$
b $P(X = 5) = 0.7461 - 0.5000 = 0.2461$
c $P(X = 9) = 0.4013 - 0.0861 = 0.3152$
- 7.32** **a** $P(X \geq 7) = 1 - 0.9819 = 0.0181$
b $E(X) = np = 15(0.2) = 3$
- 7.34** **a** $P(X = 3) = 0.2503$
b $P(X \geq 5) = 0.0781$
c $P(X \geq 10 | n = 20 \text{ and } p = 0.25) = 0.0139$
- 7.36** **a** **i** $P(X \leq 4) = 0.9672$;
ii $P(X \geq 6) = 0.0064$;
iii $P(4 \leq X \leq 6) = 0.1200$
b $60 + 2(20) = 100 \text{ minutes.}$
- 7.38** **a** $P(X = 20) = 0.0646$
b $P(X \geq 20) = 0.9666$
c $P(X \leq 24) = 0.9282$
d $E(X) = 22.5$
- 7.40** **a** $P(X \leq 5) = 0.6160$
b $P(X = 5) = 0.1755$
c $P(X \geq 7) = 0.2378$
- 7.42** $\{0.6065, 0.3033, 0.0758, 0.0126, 0.0016, 0.0002\}$
- 7.44** **a** $X \sim \text{Poisson}(\mu = 12)$; $P(X \geq 5) = 0.9924$
b $X \sim \text{Poisson}(\mu = 6)$; $P(X \geq 5) = 0.7149$
c $X \sim \text{Poisson}(\mu = 3)$; $P(X \geq 5) = 0.1847$
- 7.46** **a** $\mu_x = 5$, $P(X < 4) = 0.2650$
b $\mu_y = 15$, $P(14 \leq Y \leq 18) = 0.4563$
c $\mu_z = 10$, $P(Z = 10) = 0.1251$
- 7.48** **a** $\mu = 5$, $P(X = 1) = 0.0337$
b $\mu = 15$, $P(X > 20) = 0.0830$
- 7.50** **a** $P(X) = \{0.6, 0.4\}$
b $P(Y) = \{0.6, 0.4\}$
c $\mu_x = 1.4$, $\sigma_x^2 = 0.24$
d $\mu_y = 1.4$, $\sigma_y^2 = 0.24$
e $E(X, Y) = 2.1$, $\text{Cov}(X, Y) = 0.14$, $\sigma_x = 0.49$, $\sigma_y = 0.49$, $p = 0.5831$
- 7.52** **a** $E(X + Y) = 2.8$, $V(X + Y) = 0.76$
c Yes.
- 7.54** $P(y, x) = \{0.42, 0.21, 0.07, 0.18, 0.09, 0.03\}$, $y = 1, 2$, $x = 0, 1, 2$.
- 7.56** **a** $P(X = 1 | Y = 0) = 0.412$
b $P(Y = 0 | X = 1) = 0.286$
c $P(X = 2 | Y = 2) = 0.148$

- 7.58** T : Project completion time. Therefore, $T = \sum_i x_i E(T) = 145$, $V(T) = 31$, $\sigma_T = 5.57$
- 7.60** T : Project completion time. Therefore, $T = \sum_i x_i E(T) = 168$, $V(T) = 574$, $\sigma_T = 24$
- 7.62** The expected value does not change.
The standard deviation decreases.
- 7.64** $E(R_p) = 0.1060$, $V(R_p) = 0.0212$
- 7.66** **a** Share 1: $\mu_1 = 0.0232$, $\sigma_1^2 = 0.0973$;
Share 2: $\mu_2 = 0.0601$, $\sigma_2^2 = 0.2384$;
Share 3: $\mu_3 = 0.0136$, $\sigma_3^2 = 0.0524$
c Stock 2 (largest mean)
d Stock 3 (smallest variance)
- 7.68** **a** Share 1: $\mu_1 = 0.0232$, $\sigma_1^2 = 0.0973$,
 $\omega_1 = 0.1$;
Share 2: $\mu_2 = 0.0601$, $\sigma_2^2 = 0.2384$,
 $\omega_2 = 0.1$;
Share 3: $\mu_3 = 0.0136$, $\sigma_3^2 = 0.0524$,
 $\omega_3 = 0.8$
 $E(R_p) = 0.0192$, $V(R_p) = 0.0403$
- 7.70** Share 1: $\mu_1 = 0.0187$, $\sigma_1^2 = 0.0615$, $\omega_1 = 0.25$;
Share 2: $\mu_2 = -0.0176$, $\sigma_2^2 = 0.0232$, $\omega_2 = 0.25$;
Share 3: $\mu_3 = 0.0153$, $\sigma_3^2 = 0.0228$, $\omega_3 = 0.25$;
Share 4: $\mu_4 = 0.0495$, $\sigma_4^2 = 0.0517$, $\omega_4 = 0.25$;
 $E(R_p) = 0.0165$, $V(R_p) = 0.0154$
- 7.72** **a** $P(X = 1) = 0.1212$, $P(X = 2) = 0.1966$,
 $P(X = 3) = 0.1559$, $P(X = 4) = 0.1501$, $P(X = 5) = 0.1119$, $P(X \geq 6) = 0.2644$.
b The distribution is not symmetric.
c The most likely number of persons in a NT family is more than 6.
d X = Number of persons in a NT family, $P(X > 3) = 0.5264$
- 7.74** Let X be the number of breakdowns in a day. $X \sim \text{Poisson}(0.2)$.
a $P(X = 1) = 0.1638$
b $P(X \geq 1) = 0.1813$
c $P(1 \leq X \leq 2) = 0.1802$
- 7.76** Let X be the number of emergency calls during a month. $X \sim \text{Poisson}(12)$.
a $P(X \geq 12) = 0.5384$
b Let Y is the number of calls during a day, then $Y \sim \text{Poisson}(0.4)$. $P(Y \geq 4) = 0.0008$
- 7.78** Use $n = 100$ and $p = 0.45$:
a $P(X > 50) = 0.13458$
b $P(X < 44) = 0.38277$
c $P(X = 45) = 0.07999$
- 7.80** **a** $P(X = 1) = 0.3347$
b $P(X \geq 3) = 0.1912$
c $P(X \leq 4) = 0.9814$
- 7.82** **a** $P(\text{Home team wins}) = 0.38$
b $P(\text{Tie}) = 0.28$
c $P(\text{Visiting team wins}) = 0.34$

Chapter 8 Continuous probability distributions

- 8.2** **a** $P(20 < X < 40) = 0.5$
b $P(X < 25) = 0.125$
c $P(35 < X < 65) = 0.625$
- 8.4** **a** $P(25000 < X < 30000) = 1/6$
b $P(X > 40000) = 1/3$
c $P(X = 25000) = 0$
- 8.6** $f(x) = 7.5$; $Q_1 = 37.5$
- 8.8** **a** $P(X > 150) = 0.38$
b $P(120 \leq X \leq 160) = 0.615$
- 8.10** **a** 0.4893;
b 0.4535;
c 0.0345;
d 0.4970;
e 0.0860;
f 0.6461.
- 8.12** **a** 1.645;
b -0.84;
c 0.675;
d -1.28;
e 1.34;
f 1.555.
- 8.14** **a** $P(X < 40) = 0.1056$
b $P(X = 40) = 0$
c $P(X \geq 52) = 0.4013$
d $P(X > 40) = 0.8944$ (using part (b))
e $P(35 < X \leq 64) = 0.9295$
f $P(32 \leq X \leq 37) = 0.0399$
- 8.16** **a** $P(X < 14) = 0.9082$
b $P(X < 10) = 0.0918$
c $P(X > 14) = 0.0918$
d $P(X > 8) = 0.9962$
e $P(10 \leq X \leq 15) = 0.8854$
- 8.18** **a** $P(X > 3) = 0.0475$
b $P(2.5 - 0.15 \leq X \leq 2.5 + 0.15) = 0.3830$
c $P(X > 1.8) = 0.9901$
d If $P(Z > z_0) = 0.09$ then
 $z_0 = 1.34$ (from Table 3)
 $x_0 = 2.5 + (1.34)(0.3) = 2.902$,
or \$2 902 000.
- 8.20** **a** $P(X \leq x_0) = P(Z \leq z_0) = 0.01$.
 $z_0 = -2.33$ (from Table 3).
 $x_0 = 3000 - 200(2.33) = 2534$ hrs
- b** The probability that neither bulb has burned out is $(0.99)^2 = 0.9801$. Hence, the probability that at least one of them has burned out is $1 - 0.9801 = 0.0199$.
- 8.22** $X \sim \text{Normal}(\mu = 80, \sigma = 3.6)$
Bottom 5%: $P(Z < z_0) = 0.05$; $z_0 = -1.645$;
 $x_0 = 80 - (1.645)(3.6) = 74.08$ cm
Top 5%: $P(Z < z_0) = 0.05$; $z_0 = -1.645$;
 $x_0 = 80 + (1.645)(3.6) = 85.92$ cm
- 8.24** $P(X > 28) = 0.1151$
- 8.26** **a** $P(24 < X < 28) = 0.5762$
b $P(X > 28) = 0.2119$
c $P(X < 24) = 0.2119$
- 8.28** **a** $P(X > 12000) = 0.2643$
b $P(X < 10000) = 0.0301$
- 8.30** **a** $P(X > 10) = 0.1170$
b $P(7 < X < 9) = 0.3559$
c $P(X < 3) = 0.0162$
d $P(Z < z_0) = 0.05$; $z_0 = -1.645$;
 $x_0 = 7.5 - (1.645)(2.1) = 4.05$ hours
- 8.32** **a** $P(X < 10) = 0.0099$
b $P(Z < z_0) = 0.10$; $z_0 = -1.28$;
 $x_0 = 16.40 - (1.28)(2.75) = \12.88
- 8.34** $P(Z < z_0) = 0.02$; $z_0 = 2.055$;
 $x_0 = 100 + (2.055)(16) = 132.8$
(rounded to 133)
- 8.36** $P(Z < z_0) = 0.20$; $z_0 = 0.84$;
 $x_0 = 150 + (0.84)(25) = 171$
- 8.38** $P(Z < z_0) = 0.60$; $z_0 = 0.255$;
 $x_0 = 850 + (0.255)(90) = 872.95$
(rounded to 873)
- 8.40** X = expected return,
 $X \sim N(\mu = 0.211 \text{ or } 21.1\%)$,
 $\sigma = 0.1064 \text{ or } 10.64\%$
a $P(X < 0) = 0.0239$
b $P(X > 20) = 0.5398$
- 8.44** **a** $P(X \geq 2) = 0.0025$
b $P(X \leq 4) = 0.999994$
c $P(1 \leq X \leq 3) = 0.0497$
d $P(X = 0) = 0$ as X is a continuous variable.
- 8.46** $P(X < 10) = 1$
- 8.48** **a** $P(X > 2) = 0.6703$
b $P(X > 5) = 0.3679$
c $P(X < 10) = 0.8647$
- 8.50** **a** $P(X \leq 26) = 0.6915$
b $P(X > 30) = 0.0668$
c $P(25 < X < 27) = 0.1747$
d $P(18 \leq X \leq 23) = 0.3345$
- 8.52** **a** $x_0 = 1.0128$
b $P(Y < 1.8) = 0$
- 8.54** $x_0 = 185.275$
- 8.56** $P(Z < z_0) = 0.99$, $z_0 = 2.33$,
 $x_0 = 490 + 61(2.33) = 632.13 \approx 633$
- 8.58** $P(X > 3) = 0.0019$
- A8.2** **a** $P(X \geq 50) = 0.0465$
b $P(X \leq 99) = 1$
c $P(Z < z_0) = 0.3$, $z_0 = -0.525$,
 $x_0 = 40 - 0.5 - 0.525(5.66) = 36.53$
- A8.4** $\mu = 20$, $\sigma^2 = 16$, $\sigma = 4$
 $P(22 \leq X \leq 25) = 0.2682$

Chapter 9 Statistical inference and sampling distributions

- 9.2** As $n = 100 > 30$, using CLT, \bar{X} is approximately normal.
- 9.4** **a** $P(\bar{X} > 1050) = 0.0062$
b $P(\bar{X} < 960) = 0.0228$
c $P(\bar{X} > 1100) = 0$
- 9.6** Since $n/N (= 50/250 = 20\%)$ is large, we need to use finite population correction

factor $f = \sqrt{(250 - 50) / (250 - 1)} = 0.8962$
for the standard deviation.

$X \sim \text{Normal}(\mu = 100, \sigma = 10)$ and

$X : N(\mu = 100, \sigma = 10)$ and .

$\bar{X} : N(\mu_{\bar{x}} = 100, \sigma_{\bar{x}} = 1.267)$

a $P(\bar{X} > 103) = 0.0089$

b $P(98 < \bar{X} < 101) = 0.7270$

9.8 a $\mu_x = 4.5$

b $\sigma_x^2 = 8.25$

c $\mu_{\bar{x}} = \mu_x = 4.5, \sigma_{\bar{x}}^2 = \sigma_x^2 / n = 0.0825$

d i $P(4.4 < \bar{X} < 4.55) = 0.2043$

ii $P(\bar{X} > 5.0) = 0.0409$

iii $P(\bar{X} < 4.2) = 0.1492$

9.10 b $P(\bar{X} = 1.5) = 2/20, P(\bar{X} = 2.0) = 2/20,$
 $P(\bar{X} = 2.5) = 4/20, P(\bar{X} = 3.0) = 4/20,$
 $P(\bar{X} = 3.5) = 4/20, P(\bar{X} = 4.0) = 2/20,$
 $P(\bar{X} = 4.5) = 2/20.$

c $\mu = 3, \sigma^2 = 2; \mu_{\bar{x}} = 3, \sigma_{\bar{x}}^2 = 0.75$

9.12 $P(995 < \bar{X} < 1020) = 0.6687$

9.14 a $P(\bar{X} < 119.4) = 0.0038$

b $P(\bar{X} < 119) = 0$

9.16 $P(\text{Total} > 1120) = 0.1587$

9.18 $P(\text{Total time} > 300) = 0.1170$

9.20 a $P(X < 32) = 0.2514$

b $P(\bar{X} < 32) = 0.0918$

c $P(\bar{X} > 32) = 0.9082$

9.22 a $P(X < 75) = 0.3085$

b $P(\bar{X} < 75) = 0$

9.24 a 0.0606

b 0.025

c 0.0143

9.26 $P(\hat{p} > 0.35) = 0.7852$

9.28 $P(\hat{p} > 0.04) = 0$

9.30 $P(\hat{p} > 0.10) = 0.8749$

9.32 $P(\hat{p} > 0.32) = 0.838$

9.34 As $P(\bar{X} < 105,000) = 0.0026 > 0$, the dean's claim appears to be incorrect.

9.36 $P(\text{Total} > 3000) = P(\bar{X} > 600) = 0.2266$

9.38 $P(\hat{p} < 0.75) = 0.0096$

9.40 (a) $P(X > 336) = P(\hat{p} > 0.28) = 0.0082$ (b)

The claim appears to be false.

9.42 Answer will not change.

Chapter 10 Estimation: Single population

10.10 [3.84; 15.36]

10.12 [16.68; 28.32]

10.14 a [94.18, 105.82]

b [93.07, 106.93]

c [90.90, 109.10]

d The interval widens.

10.16 a [78.04, 81.96]

b [79.02, 80.98]

c [79.51, 80.49]

d The interval narrows.

10.18 a [492.27, 507.73]

b [484.55, 515.45]

c [469.09, 530.91]
d The interval widens.

10.20 a [92.16, 107.84]

b [192.16, 207.84]

c [492.16, 507.84]

d The width of the interval is unchanged.

10.22 Yes, because the expected value of the sample median is equal to the population mean.

10.24 Because the variance of the sample mean is less than the variance of the sample median, the sample mean is relatively more efficient than the sample median.

10.26 [2611.14, 2932.86]

10.28 [13.07, 15.53]

10.30 [5.79, 7.99]

10.32 [16.04, 29.62]

10.34 [61.01, 76.19]

10.36 [127.09, 167.58]

10.38 [65.97, 85.28]

10.40 [11.86, 12.34]

10.42 [18.66, 19.90]

10.44 [579 545; 590 581]

10.46 [6.98; 10.82]

10.48 a [31.64, 48.36]

b [33.07, 46.93]

c The t distribution is more dispersed than the standard normal; $z_{\alpha/2} \leq t_{\alpha/2}$.

10.50 a σ is unknown, use t. [338.48, 361.52]

b If $\sigma = 100$ (known), use z. [338.48, 361.52]

c As $n = 500$ is large, the t -value is almost identical to the z-value. Also, s in (a) is the same as σ in (b).

10.52 [18.11, 35.23]

10.54 a [458.40, 561.60]

b [474.49, 545.51]

c [485.20, 534.80]

d The interval narrows.

10.56 a [691.77, 708.23]

b [690.20, 709.80]

c [687.12, 712.88]

d The interval widens.

10.58 [10.03, 11.01]

10.60 [4.0, 4.6]

10.62 [4.663, 5.587]

10.64 [13.65, 14.23]

10.66 [0.815, 0.865]

10.68 [0.183, 0.357]

10.70 a [0.4108, 0.5492]

b [0.4362, 0.5238]

c [0.449, 0.511]

d The interval narrows.

10.72 [0.063, 0.097]

10.74 [0.372, 0.428]

10.76 [-0.234, 0.330]

10.78 [0.133, 0.207]

10.80 [0.426, 0.520]

10.82 Range = 200 $\approx 4\sigma$ or $\sigma = 50$;
 $n = 165.8 = 166$ (rounded up)

10.84 a $n = 68$

b $n = 271$

c $n = 97$

d $n = 17$

10.86 a $n = 166$

b $n = 7$

c $n = 97$

d $n = 4145$

10.88 a $n = 271$

b 150 ± 1

10.90 $n = 385$

10.92 $n = 1083$

10.94 $n = 55$

10.96 [0.6906, 0.7704]; [\$1.388106 m,
\$1.548504 m]

10.98 a [0.2711, 0.3127]

b [2 501 309, 2 885 018]

10.100 [13 348, 16 651]

10.102 [30.26, 35.87]

10.104 a Range = 1000 $\approx 4\sigma$ or $\sigma = 250$;
 $n = 1037$

b [280, 320]

10.106 $\bar{X} = 12.85; s = 5.80, [12.32, 13.38]$

10.108 a [5.11, 6.47]

10.110 [0.558, 0.776]

Chapter 11 Estimation: Two populations

11.2 [-40.37, -29.63]

11.4 [-1.37, 3.77]

11.6 [-3.44, 1.04]

11.8 $n_1 = n_2 = 829$

11.10 a $s_p^2 = 0.68, [1.01, 1.87]$

b Assume populations are normal with equal variances.

11.12 a Assume equal population variances;
[0.15, 0.35]

b [0.22, 0.28]

11.14 a d.f. = 64.8; [-1.59, 7.59]

b d.f. = 63.1; [-7.37, 13.37]

c Widens.

d d.f. = 130; [-0.22, 6.22]

e Narrows.

11.16 [0.50, 21.50]

11.18 a Equal variances: Pooled variance = 13.70; [3.13, 8.02]; Unequal variances: d.f. = 24.49; [2.93, 8.22]

b Both variables are approximately normally distributed and variances appear to be equal ($s_1 = 4.2$ and $s_2 = 3.35$)

- 11.20** a Assume unequal variances.
d.f = 449.31 ≈ 450. [-0.58, 12.63]
- b The listening times for the first age group appears to be normal and for the second age group do not appear to be normally distributed.
- 11.22** a Assume unequal variances.
d.f = 31.63 ≈ 30. [11.2, 33.2]
- b The times should be normally distributed with unequal variances.
- c The times are approximately normally distributed with unequal variances ($s_1 = 24.02$ and $s_2 = 9.04$).
- 11.24** [0.24, 4.58]
- 11.26** [-0.07, 7.67]
- 11.28** [-6.90, -1.10]
- 11.30** a [-4.86, 0.26]
- 11.32** a [17.38, 41.87]
- 11.34** [-0.002, 0.102]
- 11.36** [0.05, 0.15]
- 11.38** [-0.102, 0.016]
- 11.40** [-0.02, 0.23]
- 11.42** [-0.10, 0.018]
- 11.44** [0.049, 0.282]
- 11.46** Since, $s_1 (= 12)$; $s_2 (= 15)$ we can assume equal variances. Pooled variance = 184.5. [-33.71, -6.29]
- 11.48** [-0.142, -0.040]
- 12.24** a $p\text{-value} = P(Z > 2.00) = 0.0228$
c $p\text{-value} = P(Z > 4.00) = 0$
d When n increases, the z-test statistic increases, $p\text{-value}$ decreases.
e $p\text{-value} = P(Z > 0.60) = 0.2743$
f $p\text{-value} = P(Z > 0.30) = 0.3821$
g The z-test statistic decreases, $p\text{-value}$ increases.
h $p\text{-value} = P(Z > 2.40) = 0.0082$
i $p\text{-value} = P(Z > 3.60) = 0$
j The z-test statistic increases, $p\text{-value}$ decreases.
- 12.26** a $p\text{-value} = P(Z < -4.00) = 0$
b $p\text{-value} = P(Z < -2.00) = 0.0228$
c $p\text{-value} = P(Z < -1.00) = 0.1587$
d The z-test statistic increases, $p\text{-value}$ increases.
- 12.28** $p\text{-value} = P(Z < -.76) = 0.2243 > \alpha = 0.05$. Do not reject H_0 .
- 12.30** $Z = -1.7$; $p\text{-value} = 0.0446 < \alpha = 0.10$. Reject H_0 .
- 12.32** $Z = 1.60$; $p\text{-value} = 0.0548 < \alpha = 0.10$. Reject H_0 .
- 12.34** $Z = 1.61$; $p\text{-value} = 0.0537 > \alpha = 0.05$. Do not reject H_0 .
- 12.36** a Reject H_0 if $t > 1.833$. $t = 1.58 < 1.833$. Do not reject H_0 .
b Reject H_0 if $t > 2.467$. $t = 10.77 > 2.467$. Reject H_0 .
c Reject H_0 if $t < -1.318$. $t = -2.5 < -1.318$. Reject H_0 .
- 12.38** Reject H_0 if $t > 1.383$. $t = 1.82 > 1.383$. Reject H_0 .
- 12.40** $\alpha = 0.05$.
a Reject H_0 if $t > 1.729$; $t = 1.49$; $p\text{-value} = 0.0762$ (Excel). Do not reject H_0 .
b Reject H_0 if $t > 1.833$; $t = 1.05$; $p\text{-value} = 0.1597$ (Excel). Do not reject H_0 .
c Reject H_0 if $t > 1.676$; $t = 2.36$; $p\text{-value} = 0.0112$ (Excel). Reject H_0 .
d t-statistic increases, $p\text{-value}$ decreases.
e Reject H_0 if $t > 1.729$; $t = 2.68$; $p\text{-value} = 0.0074$ (Excel). Reject H_0 .
f Reject H_0 if $t > 1.729$; $t = 0.67$; $p\text{-value} = 0.2552$ (Excel). Do not reject H_0 .
g t-statistic increases, $p\text{-value}$ decreases.
h $t = 0.50$; $p\text{-value} = 0.3125$ (Excel). Do not reject H_0 .
- 12.42** i $t = 2.98$; $p\text{-value} = 0.0038$ (Excel). Reject H_0 .
j t-statistic increases, $p\text{-value}$ decreases.
- 12.44** Reject H_0 if $|t| > 1.972$ or if $p\text{-value} < \alpha = 0.05$.
a $t = -3.21$, $p\text{-value} = 0.0015$. Reject H_0 .
b $t = -1.57$, $p\text{-value} = 0.1177$. Do not reject H_0 .
c $t = -1.18$, $p\text{-value} = 0.2400$. Do not reject H_0 .
d t-statistic increases, $p\text{-value}$ increases.
- 12.46** $\alpha = 0.10$.
a Reject H_0 if $|t| > 1.711$; $t = 0.67$, $p\text{-value} = 0.5113$. Do not reject H_0 .
b Reject H_0 if $|t| > 1.761$; $t = 0.52$, $p\text{-value} = 0.6136$. Do not reject H_0 .
c Reject H_0 if $|t| > 2.132$; $t = 0.30$, $p\text{-value} = 0.7804$. Do not reject H_0 .
d t-statistic decreases, $p\text{-value}$ increases.
- 12.48** a Reject H_0 if $t > 1.812$; $t = 1.66$. Do not reject H_0 .
b Reject H_0 if $Z > 1.645$; $Z = 1.66$. Reject H_0 .
c The student t distribution is more dispersed than the standard normal.
- 12.50** Reject H_0 if $t < -1.729$; $t = -0.97$. Do not reject H_0 . No.
- 12.52** $\alpha = 0.05$. Reject H_0 if $|t| > 1.99$; $t = -1.45$ ($p\text{-value} = 0.1525$). Do not reject H_0 . No.
- 12.54** a $\alpha = 0.05$. Reject H_0 if $t < -1.677$; $t = -0.89$ ($p\text{-value} = 0.1859$). Do not reject H_0 . No.
b Times should be normally distributed, and is approximately satisfied.
- 12.56** Reject H_0 if $\bar{X} < 960.8$ or $\bar{X} > 1039.2$.
 $\beta = P(960.8 < \bar{X} < 1039.2 | \mu)$;
(μ, β): (900, 0.0012) (940, 0.1492) (980, 0.83) (1020, 0.83) (1060, 0.1492) (1100, 0.0012)
- 12.58** Reject H_0 if $\bar{X} < 921.6$ or $\bar{X} > 1078.4$.
 $\beta = P(921.6 < \bar{X} < 1078.4 | \mu)$
(μ, β): (900, 0.295) (940, 0.677) (980, 0.921) (1020, 0.921) (1060, 0.677) (1100, 0.295)
- 12.60** a Reject H_0 if $\bar{X} < 38.355$; $\beta = P(\bar{X} > 38.355 | \mu = 37) = P(Z > 1.36) = 0.0869$
b Reject H_0 if $\bar{X} < 38.96$; $\beta = P(\bar{X} > 38.96 | \mu = 37) = P(Z > 1.96) = 0.0250$
c β decreases.
- 12.62** Reject H_0 if $\bar{X} > 925.48$; $\beta = P(\bar{X} < 925.48 | \mu = 930) = P(Z < -1.36) = 0.0869$.

- 12.64** Reject H_0 if $Z > 1.645$ or if $p\text{-value} < \alpha = 0.05$. $Z = 1.4$; $p\text{-value} = 0.0808$. Do not reject H_0 . No.
- 12.66** **a** $p\text{-value} = 2P(Z < -1.01) = 0.3124$
b $p\text{-value} = P(Z > 3.45) = 0$
c $p\text{-value} = P(Z < -2.07) = 0.0192$
- 12.68** Reject H_0 if $\hat{p} < 0.406$ or $\hat{p} > 0.494$; $\beta = P(0.406 < \hat{p} < 0.494 | p = 0.50) = 0.3936$.
- 12.70** **a** $Z = 0.61$, $p\text{-value} = P(Z > 0.61) = 0.2709$
b $Z = 0.87$, $p\text{-value} = 2P(Z > 0.87) = 0.1922$
c $Z = 1.22$, $p\text{-value} = 2P(Z > 1.22) = 0.1112$
d The $p\text{-value}$ decreases.
- 12.72** Reject H_0 if $Z > 2.33$; $Z = 2.36$. Reject H_0 . Yes.
- 12.74** **a** $P(\hat{p} < 0.45) = 0.0071$
b The claim appears to be false.
- 12.76** $Z = 2.48$, $p\text{-value} = 0.0066 < \alpha = 0.05$. Reject H_0 .
- 12.78** $\hat{p} = 0.205$; $Z = -1.47$, $p\text{-value} = 0.0708 < \alpha = 0.10$. Reject H_0 .
- 12.80** Reject H_0 if $Z > 1.645$ or if $p\text{-value} < \alpha = 0.05$; $Z = 1.125$, $p\text{-value} = 0.1303$. Do not reject H_0 . No.
- 12.82** **a** Reject H_0 if $Z > 1.645$ or if $p\text{-value} < \alpha = 0.05$; $Z = 1.4$, $p\text{-value} = 0.0808$. Do not reject H_0 . No.
b Reject H_0 if $Z > 1.645$ or if $p\text{-value} < \alpha = 0.05$; $Z = 2.0$, $p\text{-value} = 0.0228$. Reject H_0 . Yes.
- 12.84** $\hat{p} = 0.23$; Reject H_0 if $Z > 1.645$. $Z = 0.82$, Do not reject H_0 .
- 12.86** Reject H_0 if $Z < -1.28$ or if $p\text{-value} < \alpha = 0.10$; $Z = -1.18$. $p\text{-value} = 0.119$. Do not reject H_0 . No.
- 12.88** Reject H_0 if $Z < -2.33$; $Z = -2.64$. Reject H_0 . Yes.
- 12.90** Reject H_0 if $t > 1.676$; $\bar{X} = 110$; $s = 12$; $t = 5.89$. $p\text{-value} = 0.0$. Reject H_0 . Yes.
- 12.92** Reject H_0 if $Z > 1.645$ or if $p\text{-value} < \alpha = 0.05$. $Z = 2.90$, $p\text{-value} = 0.0019$. Reject H_0 . Yes.
- 12.94** Reject H_0 if $Z > 1.645$ or if $p\text{-value} < \alpha = 0.05$; $Z = 1.61$, $p\text{-value} = 0.0537$. Do not reject H_0 . No.
- 13.8** $p\text{-value} = P(t > 1.0) = 0.1596$ [in Excel TDIST(1.0,138,1)]
- 13.10** Assume equal variances. Pooled variance = 2.27. Reject H_0 if $|t| > 2.145$ or if $p\text{-value} < \alpha = 0.05$; $t = 1.00$, $p\text{-value} = 0.3343$ [in Excel TDIST(1.0,14,2)]. Do not reject H_0 .
- 13.12** **a** Unequal-variances t-test; Reject H_0 if $t > t_{\alpha,v} = t_{0.5,200} = 1.653$ or if $p\text{-value} < \alpha = 0.05$. $t = 0.62$, $p\text{-value} = 0.2689$. Do not reject H_0 .
b Unequal-variances t-test; Reject H_0 if $t > t_{\alpha,v} = t_{0.5,223} \approx 1.645$ or if $p\text{-value} < \alpha = 0.05$. $t = 2.46$, $p\text{-value} = 0.0074$. Reject H_0 .
c The value of the test statistic increases and the $p\text{-value}$ decreases.
d Unequal-variances t-test; Reject H_0 if $t > t_{\alpha,v} = t_{0.5,26} = 1.706$ or if $p\text{-value} < \alpha = 0.05$. $t = 0.23$, $p\text{-value} = 0.4118$. Do not reject H_0 .
e The value of the test statistic decreases and the $p\text{-value}$ increases.
f Unequal-variances t-test; Reject H_0 if $t > t_{\alpha,v} = t_{0.5,200} = 1.653$ or if $p\text{-value} < \alpha = 0.05$. $t = .35$, $p\text{-value} = 0.3624$. Do not reject H_0 .
g The value of the test statistic decreases and the $p\text{-value}$ increases.
- 13.14** Reject H_0 if $t > 1.895$ or if $p\text{-value} < \alpha = 0.05$; $t = 2.21$, $p\text{-value} = 0$. Reject H_0 .
- 13.16** **a** Pooled variance = 21.75. Reject H_0 if $t > 2.374$; $t = 3.78$. Reject H_0 . Yes.
b $p\text{-value} = P(t > 3.78) = 0$.
- 13.18** d.f. = 189. Reject H_0 if $|t| > 2.60$ or if $p\text{-value} < \alpha = 0.05$; $t = -3.90$. Reject H_0 . Yes.
- 13.20** Equal-variances t-test; Reject H_0 if $|t| > t_{\alpha/2,v} = t_{0.05,13} = 1.771$ or if $p\text{-value} < \alpha = 0.05$. $t = 1.07$, $p\text{-value} = 0.3028$.
- 13.22** d.f. = $8.59 = 9$. Reject H_0 if $|t| > 1.833$ or if $p\text{-value} < \alpha = 0.10$. $t = 0.24$, $p\text{-value} = 0.8130$. Do not reject H_0 .
- 13.24** Reject H_0 if $|t| > 1.973$ or if $p\text{-value} < \alpha = 0.05$. $t = 1.15$, $p\text{-value} = 0.25$; do not reject H_0 .
- 13.26** Equal-variances t-test; Reject H_0 if $|t| > t_{\alpha/2,v} = t_{0.025,238} = 1.960$ or if $p\text{-value} < \alpha = 0.05$. $t = 1.55$, $p\text{-value} = 0.1204$. Do not reject H_0 .
- 13.28** Pooled variance = 485.25. Reject H_0 if $t > 1.2963$. $t = 1.07$ ($p\text{-value} = 0.1440$). Do not reject H_0 .
- 13.30** d.f. = 450. Reject H_0 if $|t| > 1.96$. $t = 1.79$ ($p\text{-value} = 0.074$). Do not reject H_0 . No.
- 13.32** Assume equal variances. Pooled variance = 565.0
a Reject H_0 if $|t| > 1.96$. $t = 1.43$ ($p\text{-value} = 0.1522$). Do not reject H_0 .
- b** $[-1.28, 8.34]$
c Both populations must be normal.
d Histograms are approximately bell-shaped.
- 13.34** d.f. = 190. Reject H_0 if $|t| > 1.973$. $t = 1.16$ ($p\text{-value} = 0.2467$). Do not reject H_0 .
- 13.36** Assume equal variances. Pooled variance = 0.00217. Reject H_0 if $|t| > 1.973$. $t = -1.21$ ($p\text{-value} = 0.2268$). Do not reject H_0 .
- 13.38** Reject H_0 if $|t| > -4.032$. $t = -2.78$ ($p\text{-value} = 0.04$). Do not reject H_0 .
- 13.40** $t = -1.52$; $p\text{-value} = 0.0734 < \alpha = 0.10$. Reject H_0 .
- 13.42** $t = 2.53$; $p\text{-value} = 0.0140 > \alpha = 0.01$. Do not reject H_0 .
- 13.44** Reject H_0 if $t < -t_{0.05,7} = -1.895$. $t = -3.22$ ($p\text{-value} = 0.0073$). Reject H_0 .
- 13.46** Reject H_0 if $t > t_{0.05,6} = 1.943$. $t = 1.98$ ($p\text{-value} = .0473$). Reject H_0 .
- 13.48** **a** Pooled variance = 408.8. Reject H_0 if $|t| > 2.0739$; $t = 1.16$ ($p\text{-value} = 0.2584$). Do not reject H_0 . No.
b $\bar{X}_D = 25.5$, Reject H_0 if $|t| > 2.2010$; $t = 7.25$ ($p\text{-value} = 0$). Reject H_0 . Yes.
c By reducing the variation, the matched pairs experiment is able to detect a difference between μ_1 and μ_2 .
- 13.50** **a** Reject H_0 if $|t| > 2.0739$; $t = 2.10$ ($p\text{-value} = 0.047$). Reject H_0 . Yes.
b $\bar{X}_D = 9.08$, Reject H_0 if $|t| > 2.2010$; $t = 1.68$ ($p\text{-value} = 0.12$). Do not reject H_0 . No.
- 13.52** $\bar{X}_D = -3.47$, Reject H_0 if $t < -1.6766$; $t = -2.44$ ($p\text{-value} = 0.0091$). Reject H_0 . Yes.
- 13.54** **a** Reject H_0 if $|Z| > 2.575$; $\hat{p} = 0.56$; $Z = -1.56$. Do not reject H_0 . No.
b Reject H_0 if $Z > 1.645$; $Z = 3.25$. Reject H_0 .
- 13.56** $\hat{p} = 0.289$. Reject H_0 if $Z < -1.645$; $Z = -0.73$ ($p\text{-value} = P(Z < -0.73) = 0.2327$). Do not reject H_0 . No.
- 13.58** **a** $\hat{p} = 0.425$, $Z = 0.72$, $p\text{-value} = 0.4716$. Do not reject H_0 .
b $\hat{p} = 0.425$, $Z = 1.43$, $p\text{-value} = 0.1528$. Do not reject H_0 .
c $p\text{-value}$ decreases.
- 13.60** Reject H_0 if $|Z| > 1.96$. $\hat{p} = 0.5714$, $Z = 0.88$ ($p\text{-value} = 0.38$). Do not reject H_0 . No.
- 13.62** Reject H_0 if $Z > 1.645$; $Z = 3.56$ ($p\text{-value} = 0.0$). Reject H_0 .
- 13.64** **a** Reject H_0 if $Z > 1.645$; $Z = 4.31$ ($p\text{-value} = 0.0$). Reject H_0 . Yes.
b Reject H_0 if $Z > 1.645$; $Z = 2.16$ ($p\text{-value} = 0.02$). Reject H_0 . Yes.
c $[0.055, 0.145]$
- 13.66** $\alpha = 0.10$. Reject H_0 if $Z > 1.282$; $Z = 3.62$ ($p\text{-value} = 0$). Reject H_0 .

Chapter 13 Hypothesis testing: Two populations

- 13.2** Reject H_0 if $Z < -2.33$ or $p\text{-value} < \alpha = 0.01$; $Z = -1.85$. $p\text{-value} = 0.032$. Do not reject H_0 .
- 13.4** $p\text{-value} = P(Z > 1.92) = 0.0274$
- 13.6** Assume equal variances. $d.f. = 33$. Pooled variance = 193.18. Reject H_0 if $t < -1.69$ or $p\text{-value} < \alpha = 0.05$; $t = -5.27$. $p\text{-value} = 0.0$. Reject H_0 . Yes.

- 13.68** a $\alpha = 0.01$. Reject H_0 if $Z > 2.33$; $Z = 2.75$, $p\text{-value} = 0.003$. Reject H_0 . Yes.
 b $\alpha = 0.01$. Reject H_0 if $Z > 2.33$; $Z = 1.80$, $p\text{-value} = 0.036$. Do not reject H_0 . No.
- 13.70** Reject H_0 if $|Z| > 1.96$; $Z = 1.13$ ($p\text{-value} = 0.2574$). Do not reject H_0 . No.
- 13.72** Reject H_0 if $Z > 1.28$; $Z = 3.01$, $p\text{-value} = 0.0013$. Reject H_0 .
- 13.74** a Assume $\sigma_1^2 \neq \sigma_2^2$. $d.f. = 13$. Reject H_0 if $t < -1.350$. $t = -3.52$ ($p\text{-value} = 0.002$). Reject H_0 . Yes.
 b If the advertising campaign is successful, net profit will be positive, $0.2(\mu_2 - \mu_1) - 100 > 0$. That is, $\mu_1 - \mu_2 < -500$. Reject H_0 if $t < -1.350$. $t = -2.35$ ($p\text{-value} = 0.018$). Reject H_0 . Yes.
- 13.76** Reject H_0 if $t > 1.833$; $t = 2.73$ ($p\text{-value} = 0.01$). Reject H_0 . Yes.
- 13.78** a $\alpha = 0.10$. Reject H_0 if $t < -1.282$; $t = -2.02$ ($p\text{-value} = 0.022$). Reject H_0 . Yes.
- 13.80** Reject H_0 if $t > 1.679$; $t = 0.54$ ($p\text{-value} = 0.2968$). Do not reject H_0 .
- 13.82** Reject H_0 if $p\text{-value} < \alpha = 0.05$. $t = -1.14$, $p\text{-value} = 0.1288$. Do not reject H_0 .
- 13.84** Reject H_0 if $p\text{-value} < \alpha = 0.05$. $Z = 1.26$, $p\text{-value} = 0.1037$. Do not reject H_0 .
- 13.86** a Equal variances t-test. $t = -1.06$, $p\text{-value} = 0.2980$.
 b Unequal variances t-test. $t = -2.87$, $p\text{-value} = 0.0040$. Reject H_0 .

Chapter 14 Chi-squared tests

- 14.2** Reject H_0 if $\chi^2 > \chi_{\alpha,k-1}^2 = \chi_{0.01,4}^2 = 13.28$ or if $p\text{-value} < \alpha = 0.01$. $\chi^2 = 2.26$, $p\text{-value} = 0.6868$. Do not reject H_0 .
- 14.4** The χ^2 value decreases and the $p\text{-value}$ increases.
- 14.6** b Reject H_0 if $\chi^2 > 7.81$ or if $p\text{-value} < \alpha = 0.05$. $\chi^2 = 9.96$, $p\text{-value} = 0.02$. Reject H_0 .
- 14.8** Reject H_0 if $\chi^2 > 7.81$ or if $p\text{-value} < \alpha = 0.05$. $\chi^2 = 6.85$, $p\text{-value} = 0.0769$. Do not reject H_0 .
- 14.10** Reject H_0 if $\chi^2 > 11.07$ or if $p\text{-value} < \alpha = 0.05$. $\chi^2 = 11.40$, $p\text{-value} = 0.044$. Reject H_0 .
- 14.12** Reject H_0 if $\chi^2 > 5.99$ or if $p\text{-value} < \alpha = 0.01$. $\chi^2 = 14.88$, $p\text{-value} = 0.001$, Reject H_0 .
- 14.14** Reject H_0 if $p\text{-value} < \alpha = 0.05$. $\chi^2 = 33.85$, $p\text{-value} = 0.0$. Reject H_0 . Yes.
- 14.16** Reject H_0 if $\chi^2 > 4.61$. $\chi^2 = 6.35$ ($p\text{-value} = 0.0419$ Excel). Reject H_0 .
- 14.18** Reject H_0 if $\chi^2 > 7.81$. $\chi^2 = 6.00$ ($p\text{-value} = 0.1116$ Excel). Do not reject H_0 .
- 14.20** Reject H_0 if $\chi^2 > 3.84$. $\chi^2 = 9.56$ ($p\text{-value} = 0.002$ Excel). Reject H_0 .

- 14.22** The χ^2 statistic decreases in Exercise 14.21 by $\frac{1}{4}$ of that in Exercise 14.19. Consequently, the $p\text{-value}$ increased. But the conclusion is unchanged.
- 14.24** Reject H_0 if $\chi^2 > 5.99$. $\chi^2 = 2.27$ ($p\text{-value} = 0.32$). Do not reject H_0 .
- 14.26** Reject H_0 if $\chi^2 > 12.59$. $\chi^2 = 33.47$ ($p\text{-value} = 0$). Reject H_0 .
- 14.28** Reject H_0 if $\chi^2 > 9.35$. $\chi^2 = 21.98$ ($p\text{-value} = 0$). Reject H_0 .
- 14.30** Reject H_0 if $\chi^2 > 12.79$. $\chi^2 = 31.48$ ($p\text{-value} = 0$). Reject H_0 . Yes.
- 14.32** Reject H_0 if $\chi^2 > 7.81$. $\chi^2 = 41.7645$ ($p\text{-value} = 0$). Reject H_0 .
- 14.34** Reject H_0 if $\chi^2 > 12.59$. $\chi^2 = 20.89$ ($p\text{-value} = 0.0019$). Reject H_0 .
- 14.36** Reject H_0 if $\chi^2 > 7.81$. $\chi^2 = 11.82$ ($p\text{-value} = 0.0080$). Reject H_0 .
- 14.38** Reject H_0 if $\chi^2 > 5.99$. $\chi^2 = 4.53$ ($p\text{-value} = 0.1038$). Do not reject H_0 .
- 14.40** Reject H_0 if $\chi^2 > 2.71$. $\chi^2 = 9.87$ ($p\text{-value} = 0.0017$). Reject H_0 .
- 14.42** a Reject H_0 if $\chi^2 > 9.49$
 b Reject H_0 if $\chi^2 > 11.07$
 c Reject H_0 if $\chi^2 > 12.59$
- 14.44** Reject H_0 if $\chi^2 > 6.25$. $\chi^2 = 1.77$ ($p\text{-value} = 0.94$). Do not reject H_0 .
- 14.46** Reject H_0 if $\chi^2 > 2.71$. $\chi^2 = 4.87$ ($p\text{-value} = 0.027$). Reject H_0 .
- 14.48** Reject H_0 if $\chi^2 > 5.99$. $\chi^2 = 3.03$ ($p\text{-value} = 0.2199$). Do not reject H_0 .
- 14.50** Reject H_0 if $\chi^2 > 4.61$. $\chi^2 = 3.20$ ($p\text{-value} = 0.20$). Do not reject H_0 .
- 14.52** Reject H_0 if $\chi^2 > 9.49$. $\chi^2 = 4.77$ ($p\text{-value} = 0.31$). Do not reject H_0 .
- 14.54** Reject H_0 if $\chi^2 > 3.84$. $\chi^2 = 11.86$ ($p\text{-value} = 0.001$). Reject H_0 .
- 14.56** Reject H_0 if $\chi^2 > 9.49$. $\chi^2 = 16.07$ ($p\text{-value} = 0.003$). Reject H_0 .
- 14.58** Reject H_0 if $\chi^2 > 4.61$ ($d.f.$ is now 2). $\chi^2 = 8.20$ ($p\text{-value} = 0.002$). Reject H_0 .
- 14.60** Reject H_0 if $\chi^2 > 11.07$. $k = 8$, $X = 0.88$; $s = 2.14$; $\chi^2 = 0.24$ ($p\text{-value} = 0.999$); Do not reject H_0 .
- 14.62** Using time to solve problems: Assume equal variances. Pooled variance = 2469.1. Reject H_0 if $t > 1.686$. $t = 1.94$ ($p\text{-value} = 0.03$). Reject H_0 .
 Using successfully repeating 5 letters:
 Use z-test (Case 1): Reject H_0 if $z < -1.645$. $z = -1.99$, $p\text{-value} = 0.0234$. Reject H_0 .
 Using successfully repeating 5 words:
 Use z-test (Case 1): Reject H_0 if $z < -1.645$. $z = -1.58$, $p\text{-value} = 0.0567$. Do not reject H_0 .
- 14.64** Reject H_0 if $\chi^2 > 9.49$ ($d.f. = 4$). $\chi^2 = 5.49$ ($p\text{-value} = 0.2409$). Do not reject H_0 .

Chapter 15 Simple linear regression and correlation

- 15.2** a y -intercept = 0, slope = 4
 b y -intercept = 2.2, slope = 1.8
 c y -intercept = 10.307, slope = -0.9458
- 15.4** a For each additional centimetre of father's height, the son's height increases, on average, by 0.516 centimetres.
 b Sons are taller.
 c Sons are shorter.
- 15.6** a $\hat{y} = 11.25 + 2.62x$
 b SSE = 469.18
- 15.8** $\hat{y} = 7.96 - 1.12x$
- 15.10** a $\hat{y} = 14.77 + 2.13x$
 d Yes. For each additional employee, the average profit per dollar of sales increases by 2.13 cents.
- 15.12** a $\hat{y} = -0.4 + 1.2x$
 c For each additional square to house size, the construction price would increase by \$1.2 ('0000). That is by \$12000.
- 15.14** a $\hat{y} = -644.95 + 195.07x$.
 b Interpretation of the intercept term is not always valid.
 c Slope = 195.07. For each additional advertising dollar, sales increase, on average, by \$195.07.
 d Higher advertising expenditure would result in lower sales.
- 15.16** a $\hat{y} = 475.17 - 39.17x$
 b There is a negative linear relationship between housing starts and mortgage rates. For every additional 1% increase in interest rate, on average, number of housing starts would fall by 39.17 houses.
- 15.18** $\hat{y} = 89.806 + 0.0514$ Test score. For every additional test score, the percentage of non-defective computer produced would increase by 0.05%.
- 15.20** a $\hat{y} = 21.6 + 1.88$ Study time
 b For every additional hour of study time, final mark would increase, on average, by 1.88.
 c The sign of the slope is logical.
- 15.22** a $\hat{y} = -5.97 + 0.226$ Age
 b For every additional year of age daily expense increases by \$0.226.
- 15.24** $\hat{y} = 153.9 + 1.958$ Income
- 15.26** a $\hat{y} = 7.287 + 0.19x$
 b For every extra cigarette consumed, on average, days absent would increase by 0.19 days.
- 15.28** For each commercial length, the memory test scores are normally distributed with constant variance and a mean which is a linear function of the commercial lengths.

- 15.30** For each number of occupants in a household, energy consumptions are normally distributed with constant variance and mean that is a linear function of the number of occupants.
- 15.32** $SSE = 717.43$; $s_e = 6.31$; $\hat{y} = 12.19 + 1.48x$; $s_{\beta_1} = 1.30$ C.I.: $[-1.25, 4.21]$
- 15.34** $SSE = 390.1$; $s_e = 8.83$; $\hat{\beta}_1 = 10.90$; $s_{\beta_1} = 1.08$. Reject H_0 : $\beta_1 = 0$ if $|t| > t_{\alpha/2, n-2}$ = $t_{0.025, 5} = 2.571$ or $p\text{-value} < \alpha = 0.05$. $t = 10.09$, $p\text{-value} = 0$. Reject H_0 . x significantly influences y .
- 15.36** **a** $s_e = 73.52$; $s_e / \bar{y} \times 100 = 13.5\%$. The fit of the model is reasonable.
- b** 195.07 ± 2.16 (11.538); [170.1, 220.0]
- c** Yes. Reject H_0 : $\beta_1 = 0$ if $t > 2.650$. $t = 16.9$. Reject H_0 .
- d** $r = 0.98$, very strong positive linear relationship. $R^2 = 0.96$, excellent fit.
- 15.38** **a** $\hat{y} = -8.5372 + 0.2843x$
- b** $s_e = 4.44$
- c** $\alpha = 0.01$ Reject H_0 if $|t| > 3.707$. $t = 21.87$. Reject H_0 . Yes.
- d** $R^2 = 0.99$. Excellent fit.
- 15.40** $s_e = 1.813$; $(s_e / \bar{y}) \times 100 = (1.813 / 26.28) \times 100 = 6.90\%$. $R^2 = 0.1884$. About 18.8% of the variation in period of employment is explained by their age. Weak linear relationship.
- 15.42** **a** $s_e = 5.888$
- b** Reject H_0 : $\beta_1 = 0$ if $p\text{-value} < \alpha = 0.05$. $t = 4.86$, $p\text{-value} = 0$. Reject H_0 . Yes.
- c** $R^2 = 0.289$. About 28.9% of the variation in the test scores is explained by the length of the commercial.
- 15.44** **a** $s_e = 8.06$
- b** Reject H_0 : $\beta_1 = 0$ if $p\text{-value} < \alpha = 0.05$. $t = 12.03$, $p\text{-value} = 0$; Reject H_0 .
- c** $R^2 = 0.267$. That is, 26.7% of the variation in the son's height is explained by the father's height.
- 15.46** **a** $\hat{y} = 2.05 + 0.091$ Exercise
- b** For every additional minute of exercise, the cholesterol level decreases by 0.0909 units.
- c** Reject H_0 : $\beta_1 = 0$ if $p\text{-value} < \alpha = 0.05$. $t = 7.06$, $p\text{-value} = 0$; Reject H_0 .
- d** $R^2 = 0.51$. About 51% of the variation in the cholesterol level is explained by the amount of exercise.
- 15.48** $SSE = 3657$; $s_e = 3.996$; $\hat{\beta}_1 = 0.1898$; $s_{\beta_1} = 0.0253$. Reject H_0 : $\beta_1 = 0$ if $|t| > t_{\alpha/2, n-2} = t_{0.025, 229} \approx 1.96$ or $p\text{-value} < \alpha = 0.05$. $t = 7.50$, $p\text{-value} = 0$. Reject H_0 . Significant linear relationship exists.
- 15.50** **a** $\hat{y} = 6.5 - 2.0x$, $x = 22.0$, $\hat{y}_{x=22.0} = -37.5$, 90% CI: [-46.61, -28.39]
- b** $x = 27.0$, $\hat{y}_{x=27.0} = -47.5$, 99% CI: [-51.47, -43.53]
- 15.52** $\hat{y} = -644.95 + 195.07x$; $\hat{y}_{x=5} = 330.4$, CI: [194.08, 466.76]
- 15.54** $\hat{y} = 3.636 + 0.2675x$;
- a** [1.375, 25.155]
- b** [11.727, 14.803]
- 15.56** [156.73, 198.39]
- 15.58** **a** [-10.76, 33.05]
- b** $\hat{y} = 24.79$, Reduction CI: [3.41, 46.18]; Cholesterol Prediction CI: [250–46.18, 250–3.41] = [203.82, 246.61]
- 15.60** $\hat{y} = 103.2 - 1.52x$; $\hat{y}_{x=10} = 88.0$. CI: [58.1, 117.9]
- 15.62** **a** Reject H_0 if $t > 1.734$. $t = 1.33$ ($p\text{-value} = 0.099$). Do not reject H_0 .
- b** Reject H_0 if $t < -2.896$. $t = -0.82$ ($p\text{-value} = 0.217$). Do not reject H_0 .
- c** Reject H_0 if $|t| > 2.12$. $t = 2.19$ ($p\text{-value} = 0.044$). Reject H_0 .
- 15.64** Reject H_0 if $p\text{-value} < \alpha = 0.05$. $r = 0.538$, $t = 4.86$, $p\text{-value} = 0$. Reject H_0 . Yes.
- 15.66** **a** $Y = 8.24 - 1.07X$
- e** No outliers
- 15.68** **a** No
- b** No
- c** Yes
- 15.70** **d** None
- e** Yes
- 15.72** **b** Errors appear normal.
- c** Outliers: observations 28, 29, 40, 69, 74.
- d** No change in error variance.
- 15.74** **b** Errors may not be normal.
- c** Several potential outliers exist.
- d** Heteroscedasticity appears to be a problem as error variance increases as y increases.
- 15.76** **a** $\hat{y} = 0.378 + 5.76x$
- b** Reject H_0 if $p\text{-value} < \alpha = 0.10$. $t = 15.67$, $p\text{-value} = 0$. Reject H_0 . Yes.
- c** $R^2 = 0.946$
- d** [17.83, 40.53]
- 15.78** **a** $\hat{y} = -2.06 + 0.345x$
- b** For every 10 additional calls, sales would increase by 3.45.
- c** Reject H_0 if $p\text{-value} < \alpha = 0.05$. $t = 13.62$, $p\text{-value} = 0$. Reject H_0 .
- d** $R^2 = 0.912$
- e** [14.56, 15.82]
- f** [3.65, 12.93]
- 15.80** **a** $\hat{y} = 363.33 + 5.0x$
- b** Reject H_0 if $p\text{-value} < \alpha = 0.05$. $t = 2.37$, $p\text{-value} = 0.04$. Reject H_0 . Yes.
- c** [399.3, 437.4]
- d** [397.4, 419.2]
- 15.82** **a** $\hat{y} = 115 + 2.47\text{Age}$
- c** $s_e = 43.32$
- d** Reject H_0 if $p\text{-value} < \alpha = 0.05$. $t = 4.84$, $p\text{-value} = 0$. Reject H_0 .
- e** $R^2 = 0.566$
- f** [318.1, 505.2]
- 15.84** **a** $\hat{y} = 13836 + 129.2\text{Height}$
- b** $s_e = 2734$
- c** Reject H_0 if $p\text{-value} < \alpha = 0.05$. $t = 1.90$, $p\text{-value} = 0.034$. Reject H_0 . Yes.
- d** $R^2 = 0.114$
- e** i. [35863, 39103] ii. [30756, 42141]
- 15.86** Reject H_0 if $p\text{-value} < \alpha = 0.05$. $t = 13.62$, $p\text{-value} = 0$. Reject H_0 .
- 15.88** Reject H_0 if $p\text{-value} < \alpha = 0.05$. $t = 13.77$, $p\text{-value} = 0$. Reject H_0 .

Chapter 16 Multiple regression

- 16.2** **a** $n = 50$, $k = 4$, $d.f. = n - k - 1 = 45$. Reject H_0 if $|t| > 2.014$. $t = 2.60$. Reject H_0 . Yes
- b** Reject H_0 if $t < -2.412$. $t = -1.16$. Do not reject H_0 . No.
- c** Reject H_0 if $t > 1.301$. $t = 1.23$. Do not reject H_0 . No.
- 16.4** **a** $SSR = 62.2$
- b** $s_e = 2.35$
- c** $R^2 = 0.194$; $\text{Adj-}R^2 = 0.159$
- d** Reject H_0 if $F > 5.08$. $F = 5.64$. Reject H_0 . The model is useful.
- 16.6** **a** (Permits) $\hat{\beta}_1 = 0.062$: For each additional building permit issued, on average, the plaster demand increases by 0.062 hundred sheets. (Rates) $\hat{\beta}_2 = -1.212$: For each one-point increase in the mortgage rate, on average, plaster demand decreases by 1.212 hundred sheets. (GDP) $\hat{\beta}_3 = 0.022$: For each thousand-dollar increase in per capita GDP, on average, plaster demand increases by 0.022 hundred sheets.
- b** $\alpha = 0.05$; Reject H_0 if $|t| > 2.036$.
- i** Test of β_1 : $t = 2.067 > 2.036$. Reject H_0 . Yes.
- ii** Test of β_2 : $t = -1.839 < -2.306$. Do not reject H_0 . No.
- iii** Test of β_3 : $t = 4.400 > 2.306$. Reject H_0 . Yes.
- 16.8** **a** $s_e = 3.752$.
- b** $R^2 = 76.3\%$.
- c** $\text{Adj-}R^2 = 74.5\%$; the model fits well.
- d** $F = 43.43$ ($p\text{-value} = 0 < \alpha = 0.05$). Reject H_0 . The model is useful.
- f** Assignment mark: $t = 0.97$ ($p\text{-value} = 0.342 > \alpha = 0.05$). Do not reject H_0 . No.
- g** Mid-semester: $t = 9.12$ ($p\text{-value} = 0 < \alpha = 0.05$). Reject H_0 . Yes.
- 16.10** **a** $s_e = 8.045$.
- b** $R^2 = 27.1\%$, which is the proportion of the variation in the heights of men that is explained by the model.

- c** Adj- $R^2 = 26.8\%$; the model does not fit very well.
- d** $F = 73.91$ (p -value = $0 < \alpha = 0.05$). Reject H_0 . The model is useful.
- e** Father: $\hat{\beta}_1 = 0.5068$; for each additional centimetre of father's height, the average son's height increases by 0.5068 centimetre. Mother: $\hat{\beta}_2 = -0.0879$; for each additional centimetre of mother's height, the average son's height decreases by 0.0879 centimetre.
- f** $t = 11.70$ (p -value = $0 < \alpha = 0.05$). Reject H_0 . Yes.
- g** $t = -1.61$ (p -value = $0.107 > \alpha = 0.05$). Do not reject H_0 . No.
- 16.12** **a** $\hat{y} = 12.3 + 0.57$ Direct + 3.32 Newspaper + 0.73 Television
- b** $R^2 = 0.195$; Adj- $R^2 = 0.08$. Model fitness is very poor.
- c** $s_e = 2.587$
- d** $F = 1.70$ (p -value = $0.198 > \alpha = 0.05$). Do not reject H_0 : $\beta_1 = \beta_2 = \beta_3 = 0$. Model is not useful.
- e** Direct: $t = 0.33$ (p -value = $0.74 > \alpha = 0.05$). Do not reject H_0 ;
- Newspapers: $t = 2.16$ (p -value = $0.04 < \alpha = 0.05$). Reject H_0 ;
- Television: $t = 0.37$ (p -value = $0.71 > \alpha = 0.05$). Do not reject H_0 .
- f** \$18.213 ('000)
- g** \$18.213 ('000)
- 16.14** **a** $\hat{y} = 6.06 - 0.0078x_1 + 0.603x_2 - 0.070x_3$
- b** $s_e = 1.92$; Adj- $R^2 = 0.68$; $F = 36.12$ (p -value = $0 < \alpha = 0.05$). Model is useful.
- c** Age: $t = -0.12$, p -value = 0.91 ; insignificant. Years: $t = 6.25$, p -value = 0 ; significant. Pay: $t = -1.34$, p -value = 0.19 ; insignificant.
- d** $\hat{y}_{(\text{Age} = 36, \text{Years} = 10, \text{Salary} = 32)} = \9.57 ('000); [5.64, 13.50]; [8.86, 10.27]. 5 weeks' severance pay = $\$32000 \times (5/52) = \3076.92 falls below the prediction interval. Bill is correct.
- 16.16** **a** The histogram is approximately bell-shaped.
- b** No evidence of heteroscedasticity.
- 16.18** **a** The histogram is approximately bell-shaped.
- b** No evidence of heteroscedasticity.
- 16.20** **a** Errors appear to be independent.
- b** Errors are not independent.
- 16.22** **a** Errors appear to be non-normal.
- b** Error variance is not a constant.
- c** Errors do not appear to be independent.
- 16.24** **a** Independent variables are uncorrelated.
- b** t -tests are valid.
- 16.26** **a** $\hat{y} = -103.1 + 5.82x_1 + 8.56x_2$
- b** Observations 63, 81, 82, 97
- c** Histogram is bell-shaped.
- d** Variance grows as \hat{y} increases.
- e** $y' = \log y$ or $y' = 1/y$. Neither transformation is effective.
- 16.28** All standardised residuals are less than 2.
- 16.30** **a** Histogram of the residuals is bell-shaped. Error variance is not a constant.
- b** No multicollinearity.
- c** Observation 3 has a large standardised residual.
- 16.32** **a** Sales = $3719 - 46.8$ Price-A + 58.5 Price-B
- b** The histogram of the residuals does not appear to be bell-shaped. The error variance appears to be constant.
- d** $s_e = 558.7$, $R^2 = 0.493$, Adj- $R^2 = 0.470$, $F = 23.85$ (p -value = $0 < \alpha = 0.05$). Model is useful.
- f** Multicollinearity is not a problem.
- 16.34** $d = 1.38$, $d_L = 1.46$, $d_U = 1.63$. Since $d < d_L$, there is evidence of positive 1st order autocorrelation.
- 16.36** $d = 2.25$, $d_L = 1.19$, $d_U = 1.73$, $4 - d_L = 2.81$, $4 - d_U = 2.27$. Since $d_U < d = 2.25 < 4 - d_U$, there is no negative 1st order autocorrelation.
- 16.38** $d = 1.75$, $d_L = 1.01$, $d_U = 1.78$. As $d_L < d < d_U$, the test is inconclusive.
- 16.40** **a** $\hat{y} = 2260 + 0.423$ Competitor;
- c** Observation 4 is a possible outlier.
- d** The histogram is roughly bell-shaped.
- e** The error variance may be constant (it is difficult to judge from the plot).
- f** $n = 52$, $k = 1$, $a = 0.05$, $d_L = 1.5$, $d_U = 1.59$, $d = 0.79$; Since $d = 0.79 < d_L = 1.50$, there is evidence of positive 1st order autocorrelation.
- g** Company = $\beta_0 + \beta_1$ Compttor + $\beta_2 t + \varepsilon$; $\hat{y} = 446 + 1.10$ Compttor + $38.9t$
- h** The second model fits better. $d_L = 1.46$, $d_U = 1.63$, $d = 2.26$. Since $1.46 = d_L < d = 2.26 < 4 - d_U = 2.37$. There is no evidence of 1st order autocorrelation.
- 16.42** The errors are normally distributed with a constant variance. $d = 2.37$. $n = 16$, $k = 1$, $\alpha = 0.05$, $d_L = 1.10$, $d_U = 1.37$. Since $d_U = 1.37 < d = 2.29 < 4 - d_U = 2.63$, there is no evidence of 1st order autocorrelation.
- 16.44** **a** The histogram is bell-shaped.
- b** The plot of the residuals versus the predicted values of y indicates a constant variance.
- 16.46** **a** $\hat{y} = -19.47 + 0.15838x_1 + 0.9625x_2$, $\hat{\beta}_1 = 0.15838$: For each additional \$1000 in advertising, annual sales increase by an average of 0.15838 (\$million). $\hat{\beta}_2 = 0.9625$: For each
- additional sales agent, annual sales increase by an average of 0.9625 (\$million)
- b** Reject H_0 if $|t| > 2.179$. For β_1 : $t = 2.82 > 2.179$; Reject H_0 . Yes. For β_2 : $t = 1.24 < 2.179$; Do not reject H_0 . No.
- c** Reject H_0 if $F > 3.89$. Since $F = 6.61$; reject H_0 . The model is useful.
- d** $R^2 = 0.52$, Adj- $R^2 = 0.44$; 44% of the variation in y is explained by the independent variables.
- e** Since the correlation between x_1 and x_2 is small ($r_{1,2} = 0.329$), collinearity is not a problem.
- f** The error variable appears to be normal.
- g, h** It appears that σ_e^2 is fixed and that the errors are independent of one another.
- i** No, the required conditions are satisfied.
- k** In the model in part a, there was sufficient evidence to show that x_1 and y were linearly related. $\hat{y} = -8.2 + 0.0905x_1 - 0.071x_2 + 1.93x_3$. (p -values are β_1 : 0.0198, β_2 : 0.941, β_3 : 0.121)
- In this new model with x_3 added, none of the coefficients are significant – the x variables are not linearly related with y .
- l** The high correlation between x_1 and x_3 ($r_{1,3} = 0.647$) and x_2 and x_3 ($r_{2,3} = 0.679$) points to multicollinearity, which helps explain the difference.
- 16.48** **a** $\hat{y} = 35.7 + 0.247$ Math-Dgr + 0.245 Age + 0.067 Income
- b** $F = 6.66$, p -value = $0.001 < \alpha = 0.05$. Reject H_0 . Yes.
- c** The error appears to be normally distributed with constant variance.
- d** Correlations: Math-Dgr and Age = 0.077; Math-Degree and Income = 0.099, Age and Income = 0.570; Age and income are correlated, which probably affects the t -tests.
- e** Only Math-Degree is linearly related to test score. Multicollinearity may have affected the t -tests of β_2 (Age) and β_3 (income).
- f** 63.81
- 16.50** $\hat{y} = 47.8 + 0.78$ Evaluation + 1.06 Articles; $s_e = 7.009$, $R^2 = 72.1\%$, Adj- $R^2 = 0.71$.
- F -test: $F = 60.70$, p -value = $0 < \alpha = 0.05$. Reject H_0 . The model is useful. The error is normally distributed and σ_e^2 is constant. No autocorrelation as $d = 1.77$. Evaluation and Articles are correlated (0.639), which points to possible multicollinearity, making the t -tests of β_1 and β_2 misleading.

- 16.52** $\hat{y} = 77809 + 745.9$ (ownadv) – 87.5 (comadv); Adj $R^2 = 0.02$.
 F-test: $F = 1.16$, p -value = 0.337 > $\alpha = 0.05$. Do not reject H_0 . Model may not be useful. $d = 0.5931$, $\alpha = 0.05$, $n = 20$, $k = 2$, $d_L = 1.10$ and $d_U = 1.54$. Since $d = 0.5931 < d_L = 1.10$, there is evidence for positive 1st order autocorrelation.
 New model with time trend variable, years, added: $\hat{y} = 54945.8 + 701.8$ (ownadv) – 92.3 (comadv) + 2299.7 years; Adj $R^2 = 0.69$.
 F-test: $F = 15.26$, p -value = 0.00 < $\alpha = 0.05$. Model is useful. $d = 1.89$, ($\alpha = 0.05$, $n = 20$, $k = 3$, $d_L = 1.00$, $d_U = 1.68$, $4 - d_U = 2.32$, $4 - d_L = 3.00$). As $d_U = 1.68 < d = 1.89 < 4 - d_U = 2.32$, errors are not autocorrelated. Notice that the model has improved dramatically. All these diagnostic statistics indicate that the model fit is very good. The t -values (and p -values) indicate that number of own advertisements and years are significantly linearly related to the number of tickets sold.

Chapter 17 Time series analysis and forecasting

- 17.2** **a** 42.00, 36.67, ..., 49.67, 39.67
b 38.8, 35.4, ..., 48.8, 46.2
- 17.4** **a** 12, 12.6, ..., 16.79, 16.51
b 12, 16.80, ..., 22.69, 15.74
c Yes.
- 17.6** **a** 38, 42.00, 42.00, ..., 46.62, 45.32
b Yes.
- 17.8** **a** **ii** 29.750, 28.625, ..., 36.375, 33.625
iii There appears to be a gradual trend of increasing sales.
b 18.0, 24.0, ..., 31.6, 37.0
c 18.0, 30.0, ..., 28.8, 41.8
- 17.10** **b** Yes
c Yes, a linear trend is suitable.
- 17.12** **a** The linear trend model fits well.
b $\hat{y} = 63.87 - 3.94t$ ($R^2 = 0.94$);
 $\hat{y} = 57.2 - 0.61t - 0.3t^2$ ($R^2 = 0.98$). The quadratic trend line fits slightly better.

- 17.14** **b** $\hat{y} = 29905x - 361795$ ($R^2 = 0.86$)
c $\hat{y} = 744.96x^2 - 14792x + 92633$, ($R^2 = 0.99$)
 A quadratic model fits better than a linear model.
- 17.16** **a** 101.3, 91.2, ..., 68.0, 61.2
c There appears a cyclical pattern.
- 17.18** **b** (173.4, 173.8, ..., 16.81, 17.02)
c Yes.
- 17.20** **b** (114.3, 111.4, ..., 97.8, 96.1)
c Yes.
- 17.22** **a** 20.4, 20.2, ..., 21.4, 22.8
b Seasonal index: 0.671, 0.865, 0.855, 1.260, 1.349
- 17.24** 1.085, 1.000, 1.020, 0.895
- 17.26** **b** $\hat{y} = 38.21t + 382.3$, $R^2 = 0.98$
- 17.28** SI: 0.839, 1.057, 1.275, 0.829;
 Seasonally adjusted values: 62.0, 63.4, ..., 76.9, 80.8
- 17.30** **b** (4220.7, 4262.8, ..., 5336.3, 5367.5)
c 1.0642, 0.8824, 1.0263, 1.0271
- 17.32** SI: 1.3375, 0.5113, 0.4984, 0.6154, 0.6979, 1.1748, 2.1646
- 17.34** MAD = 4.60, SSFE = 125
- 17.36** Model 1: MAD = 1.225, Model 2: MAD = 0.625; Model 2 is more accurate.
 Model 1: SSFE = 7.39, Model 2: SSFE = 3.75; Model 2 is more accurate.
- 17.38** $F_8 = S_8 = 14.2$
- 17.40** Inflation $F_{(2020)} = 1.7$
- 17.42** 191.1, 331.2, 418.5, 169.2
- 17.44** 1650.0, 2455.2, 2972.2, 3740.8
- 17.46** 66730, 67010, 67270, 67550, 67750
- 17.48** 540.5
- 17.50** 16.53, 22.19, 21.78, 31.02, 35.24
- 17.52** 1585, 1597, 1617, 1545
- 17.54** **a** $\hat{y} = 15.377.3 + 70.66t$, $R^2 = 0.75$
b SI (Jan-Dec): {0.9868, 0.8858, 0.9763, 0.9480, 0.9710, 0.9591, 0.9764, 0.9768, 0.9702, 1.0150, 1.0457, 1.2889}
c Forecast (Jan-Dec 2020): {27796.7, 25012.1, 27638.8, 26903.0, 27624.4, 27352.9, 27916.7, 27998.3, 27877.1, 29236.0, 30192.5, 37306.0}
- 17.56** 68.60, 87.75, 106.83, 71.05
- 17.58** **b** $\hat{y} = 0.0265t + 64.5$, $R^2 = 0.55$
c Forecast (Jan-Mar 2020): 66.13, 67.17, 66.71
d $\hat{y} = 65.1826 + 0.0267t - 0.9923M_1 + 0.0050M_2 - 0.4651M_3 - 0.8275M_4 - 0.5938M_5 - 0.7815M_6 - 0.7779M_7 - 1.0728M_8 - 1.0262M_9 - 0.9963M_{10} - 0.7233M_{11}$
 Forecast (Jan-Mar 2020): 66.14, 67.16, 67.18
e $\hat{y} = 19.2969 + 0.70576y_{t-1}$
 Forecast (Jan-Mar 2020): 66.79, 66.44, 66.19
f-g
- | | Simple regression | Multiple regression | Autoregressive model |
|------|-------------------|---------------------|----------------------|
| MAD | 0.0248 | 0.1862 | 0.6572 |
| SSFE | 0.0021 | 0.2359 | 1.3284 |

Chapter 18 Index numbers

- 18.2** Beer consumption (2000 = 100): {100.0, 96.1, ..., 76.4, 75.4}
- 18.4** Laspeyres index = 263.07; Paasche index = 259.66
- 18.6** Laspeyres index = 362.61; Paasche index = 350.99
- 18.8** Income in 1990\$ (1995–2019): {482, 476, ..., 596, 602}
 Income in 1995\$ (1995–2019): {549, 543, ..., 679, 685}
- 18.10** Laspeyres index = 248.2;
 Paasche index = 237.6;
 Fisher index = 242.9
- 18.12** 569.1, 584.5, ..., 458.6, 459.3
- 18.14** **a** Simple aggregate index = 276.4;
b-c Laspeyres index = 387.0;
 Paasche index = 418.1; Fisher index = 402.2
d LP index indicates that prices increased by about 287% and PP index indicates that prices increased by 318.1% between 1990 and 2020.
- 18.16** Average weekly earnings in 2005 Constant \$: (2005–2019): {881.95, 870.53, ..., 1088.19, 1101.67}.

Appendix B

Statistical Tables

TABLE 1 Binomial Probabilities

Tabulated values are $P(X \leq k) = \sum_{x=0}^k p(x)$ (Values are rounded to four decimal places.)

$n = 5$

<i>k</i>	<i>p</i>														
	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
0	0.9510	0.7738	0.5905	0.3277	0.2373	0.1681	0.0778	0.0313	0.0102	0.0024	0.0010	0.0003	0.0000	0.0000	0.0000
1	0.9990	0.9774	0.9185	0.7373	0.6328	0.5282	0.3370	0.1875	0.0870	0.0308	0.0156	0.0067	0.0005	0.0000	0.0000
2	1.0000	0.9988	0.9914	0.9421	0.8965	0.8369	0.6826	0.5000	0.3174	0.1631	0.1035	0.0579	0.0086	0.0012	0.0000
3	1.0000	1.0000	0.9995	0.9933	0.9844	0.9692	0.9130	0.8125	0.6630	0.4718	0.3672	0.2627	0.0815	0.0226	0.0010
4	1.0000	1.0000	1.0000	0.9997	0.9990	0.9976	0.9898	0.9688	0.9222	0.8319	0.7627	0.6723	0.4095	0.2262	0.0490

$n = 6$

<i>k</i>	<i>p</i>														
	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
0	0.9415	0.7351	0.5314	0.2621	0.1780	0.1176	0.0467	0.0156	0.0041	0.0007	0.0002	0.0001	0.0000	0.0000	0.0000
1	0.9985	0.9672	0.8857	0.6554	0.5339	0.4202	0.2333	0.1094	0.0410	0.0109	0.0046	0.0016	0.0001	0.0000	0.0000
2	1.0000	0.9978	0.9842	0.9011	0.8306	0.7443	0.5443	0.3438	0.1792	0.0705	0.0376	0.0170	0.0013	0.0001	0.0000
3	1.0000	0.9999	0.9987	0.9830	0.9624	0.9295	0.8208	0.6563	0.4557	0.2557	0.1694	0.0989	0.0159	0.0022	0.0000
4	1.0000	1.0000	0.9999	0.9984	0.9954	0.9891	0.9590	0.8906	0.7667	0.5798	0.4661	0.3446	0.1143	0.0328	0.0015
5	1.0000	1.0000	1.0000	0.9999	0.9998	0.9993	0.9959	0.9844	0.9533	0.8824	0.8220	0.7379	0.4686	0.2649	0.0585

$n = 7$

<i>k</i>	<i>p</i>														
	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
0	0.9321	0.6983	0.4783	0.2097	0.1335	0.0824	0.0280	0.0078	0.0016	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000
1	0.9980	0.9556	0.8503	0.5767	0.4449	0.3294	0.1586	0.0625	0.0188	0.0038	0.0013	0.0004	0.0000	0.0000	0.0000
2	1.0000	0.9962	0.9743	0.8520	0.7564	0.6471	0.4199	0.2266	0.0963	0.0288	0.0129	0.0047	0.0002	0.0000	0.0000
3	1.0000	0.9998	0.9973	0.9667	0.9294	0.8740	0.7102	0.5000	0.2898	0.1260	0.0706	0.0333	0.0027	0.0002	0.0000
4	1.0000	1.0000	0.9998	0.9953	0.9871	0.9712	0.9037	0.7734	0.5801	0.3529	0.2436	0.1480	0.0257	0.0038	0.0000
5	1.0000	1.0000	1.0000	0.9996	0.9987	0.9962	0.9812	0.9375	0.8414	0.6706	0.5551	0.4233	0.1497	0.0444	0.0020
6	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9984	0.9922	0.9720	0.9176	0.8665	0.7903	0.5217	0.3017	0.0679

TABLE 1 Binomial Probabilities (Continued)*n* = 8

<i>k</i>	<i>p</i>														
	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
0	0.9227	0.6634	0.4305	0.1678	0.1001	0.0576	0.0168	0.0039	0.0007	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.9973	0.9428	0.8131	0.5033	0.3671	0.2553	0.1064	0.0352	0.0085	0.0013	0.0004	0.0001	0.0000	0.0000	0.0000
2	0.9999	0.9942	0.9619	0.7969	0.6785	0.5518	0.3154	0.1445	0.0498	0.0113	0.0042	0.0012	0.0000	0.0000	0.0000
3	1.0000	0.9996	0.9950	0.9437	0.8862	0.8059	0.5941	0.3633	0.1737	0.0580	0.0273	0.0104	0.0004	0.0000	0.0000
4	1.0000	1.0000	0.9996	0.9896	0.9727	0.9420	0.8263	0.6367	0.4059	0.1941	0.1138	0.0563	0.0050	0.0004	0.0000
5	1.0000	1.0000	1.0000	0.9988	0.9958	0.9887	0.9502	0.8555	0.6846	0.4482	0.3215	0.2031	0.0381	0.0058	0.0001
6	1.0000	1.0000	1.0000	0.9999	0.9996	0.9987	0.9915	0.9648	0.8936	0.7447	0.6329	0.4967	0.1869	0.0572	0.0027
7	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9961	0.9832	0.9424	0.8999	0.8322	0.5695	0.3366	0.0773

n = 9

<i>k</i>	<i>p</i>														
	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
0	0.9135	0.6302	0.3874	0.1342	0.0751	0.0404	0.0101	0.0020	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.9966	0.9288	0.7748	0.4362	0.3003	0.1960	0.0705	0.0195	0.0038	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000
2	0.9999	0.9916	0.9470	0.7382	0.6007	0.4628	0.2318	0.0898	0.0250	0.0043	0.0013	0.0003	0.0000	0.0000	0.0000
3	1.0000	0.9994	0.9917	0.9144	0.8343	0.7297	0.4826	0.2539	0.0994	0.0253	0.0100	0.0031	0.0001	0.0000	0.0000
4	1.0000	1.0000	0.9991	0.9804	0.9511	0.9012	0.7334	0.5000	0.2666	0.0988	0.0489	0.0196	0.0009	0.0000	0.0000
5	1.0000	1.0000	0.9999	0.9969	0.9900	0.9747	0.9006	0.7461	0.5174	0.2703	0.1657	0.0856	0.0083	0.0006	0.0000
6	1.0000	1.0000	1.0000	0.9997	0.9987	0.9957	0.9750	0.9102	0.7682	0.5372	0.3993	0.2618	0.0530	0.0084	0.0001
7	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9962	0.9805	0.9295	0.8040	0.6997	0.5638	0.2252	0.0712	0.0034
8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9980	0.9899	0.9596	0.9249	0.8658	0.6126	0.3698	0.0865

n = 10

<i>k</i>	<i>p</i>														
	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
0	0.9044	0.5987	0.3487	0.1074	0.0563	0.0282	0.0060	0.0010	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.9957	0.9139	0.7361	0.3758	0.2440	0.1493	0.0464	0.0107	0.0017	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.9999	0.9885	0.9298	0.6778	0.5256	0.3828	0.1673	0.0547	0.0123	0.0016	0.0004	0.0001	0.0000	0.0000	0.0000
3	1.0000	0.9990	0.9872	0.8791	0.7759	0.6496	0.3823	0.1719	0.0548	0.0106	0.0035	0.0009	0.0000	0.0000	0.0000
4	1.0000	0.9999	0.9984	0.9672	0.9219	0.8497	0.6331	0.3770	0.1662	0.0473	0.0197	0.0064	0.0001	0.0000	0.0000
5	1.0000	1.0000	0.9999	0.9936	0.9803	0.9527	0.8338	0.6230	0.3669	0.1503	0.0781	0.0328	0.0016	0.0001	0.0000
6	1.0000	1.0000	1.0000	0.9991	0.9965	0.9894	0.9452	0.8281	0.6177	0.3504	0.2241	0.1209	0.0128	0.0010	0.0000
7	1.0000	1.0000	1.0000	0.9999	0.9996	0.9984	0.9877	0.9453	0.8327	0.6172	0.4744	0.3222	0.0702	0.0115	0.0001
8	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9983	0.9893	0.9536	0.8507	0.7560	0.6242	0.2639	0.0861	0.0043
9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9990	0.9940	0.9718	0.9437	0.8926	0.6513	0.4013	0.0956

TABLE 1 Binomial Probabilities (Continued) $n = 15$

<i>k</i>	<i>p</i>														
	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
0	0.8601	0.4633	0.2059	0.0352	0.0134	0.0047	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.9904	0.8290	0.5490	0.1671	0.0802	0.0353	0.0052	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.9996	0.9638	0.8159	0.3980	0.2361	0.1268	0.0271	0.0037	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	1.0000	0.9945	0.9444	0.6482	0.4613	0.2969	0.0905	0.0176	0.0019	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
4	1.0000	0.9994	0.9873	0.8358	0.6865	0.5155	0.2173	0.0592	0.0093	0.0007	0.0001	0.0000	0.0000	0.0000	0.0000
5	1.0000	0.9999	0.9978	0.9389	0.8516	0.7216	0.4032	0.1509	0.0338	0.0037	0.0008	0.0001	0.0000	0.0000	0.0000
6	1.0000	1.0000	0.9997	0.9819	0.9434	0.8689	0.6098	0.3036	0.0950	0.0152	0.0042	0.0008	0.0000	0.0000	0.0000
7	1.0000	1.0000	1.0000	0.9958	0.9827	0.9500	0.7869	0.5000	0.2131	0.0500	0.0173	0.0042	0.0000	0.0000	0.0000
8	1.0000	1.0000	1.0000	0.9992	0.9958	0.9848	0.9050	0.6964	0.3902	0.1311	0.0566	0.0181	0.0003	0.0000	0.0000
9	1.0000	1.0000	1.0000	0.9999	0.9992	0.9963	0.9662	0.8491	0.5968	0.2784	0.1484	0.0611	0.0022	0.0001	0.0000
10	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9907	0.9408	0.7827	0.4845	0.3135	0.1642	0.0127	0.0006	0.0000
11	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9981	0.9824	0.9095	0.7031	0.5387	0.3518	0.0556	0.0055	0.0000
12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9963	0.9729	0.8732	0.7639	0.6020	0.1841	0.0362	0.0004
13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9948	0.9647	0.9198	0.8329	0.4510	0.1710	0.0096
14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9953	0.9866	0.9648	0.7941	0.5367	0.1399

 $n = 20$

<i>k</i>	<i>p</i>														
	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
0	0.8179	0.3585	0.1216	0.0115	0.0032	0.0008	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.9831	0.7358	0.3917	0.0692	0.0243	0.0076	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.9990	0.9245	0.6769	0.2061	0.0913	0.0355	0.0036	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	1.0000	0.9841	0.8670	0.4114	0.2252	0.1071	0.0160	0.0013	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4	1.0000	0.9974	0.9568	0.6296	0.4148	0.2375	0.0510	0.0059	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	1.0000	0.9997	0.9887	0.8042	0.6172	0.4164	0.1256	0.0207	0.0016	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	1.0000	1.0000	0.9976	0.9133	0.7858	0.6080	0.2500	0.0577	0.0065	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
7	1.0000	1.0000	0.9996	0.9679	0.8982	0.7723	0.4159	0.1316	0.0210	0.0013	0.0002	0.0000	0.0000	0.0000	0.0000
8	1.0000	1.0000	0.9999	0.9900	0.9591	0.8867	0.5956	0.2517	0.0565	0.0051	0.0009	0.0001	0.0000	0.0000	0.0000
9	1.0000	1.0000	1.0000	0.9974	0.9861	0.9520	0.7553	0.4119	0.1275	0.0171	0.0039	0.0006	0.0000	0.0000	0.0000
10	1.0000	1.0000	1.0000	0.9994	0.9961	0.9829	0.8725	0.5881	0.2447	0.0480	0.0139	0.0026	0.0000	0.0000	0.0000
11	1.0000	1.0000	1.0000	0.9999	0.9991	0.9949	0.9435	0.7483	0.4044	0.1133	0.0409	0.0100	0.0001	0.0000	0.0000
12	1.0000	1.0000	1.0000	1.0000	0.9998	0.9987	0.9790	0.8684	0.5841	0.2277	0.1018	0.0321	0.0004	0.0000	0.0000
13	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9935	0.9423	0.7500	0.3920	0.2142	0.0867	0.0024	0.0000	0.0000
14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9984	0.9793	0.8744	0.5836	0.3828	0.1958	0.0113	0.0003	0.0000
15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9941	0.9490	0.7625	0.5852	0.3704	0.0432	0.0026	0.0000
16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9987	0.9840	0.8929	0.7748	0.5886	0.3730	0.0159	0.0000
17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9964	0.9645	0.9087	0.7939	0.5231	0.0755	0.0010
18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9924	0.9757	0.9308	0.6083	0.2642	0.0169
19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9992	0.9968	0.9885	0.8784	0.6415	0.1821

TABLE 1 Binomial Probabilities (Continued) $n = 25$

<i>k</i>	<i>p</i>														
	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
0	0.7778	0.2774	0.0718	0.0038	0.0008	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.9742	0.6424	0.2712	0.0274	0.0070	0.0016	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.9980	0.8729	0.5371	0.0982	0.0321	0.0090	0.0004	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.9999	0.9659	0.7636	0.2340	0.0962	0.0332	0.0024	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4	1.0000	0.9928	0.9020	0.4207	0.2137	0.0905	0.0095	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	1.0000	0.9988	0.9666	0.6167	0.3783	0.1935	0.0294	0.0020	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	1.0000	0.9998	0.9905	0.7800	0.5611	0.3407	0.0736	0.0073	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	1.0000	1.0000	0.9977	0.8909	0.7265	0.5118	0.1536	0.0216	0.0012	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
8	1.0000	1.0000	0.9995	0.9532	0.8506	0.6769	0.2735	0.0539	0.0043	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
9	1.0000	1.0000	0.9999	0.9827	0.9287	0.8106	0.4246	0.1148	0.0132	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000
10	1.0000	1.0000	1.0000	0.9944	0.9703	0.9022	0.5858	0.2122	0.0344	0.0018	0.0002	0.0000	0.0000	0.0000	0.0000
11	1.0000	1.0000	1.0000	0.9985	0.9893	0.9558	0.7323	0.3450	0.0778	0.0060	0.0009	0.0001	0.0000	0.0000	0.0000
12	1.0000	1.0000	1.0000	0.9996	0.9966	0.9825	0.8462	0.5000	0.1538	0.0175	0.0034	0.0004	0.0000	0.0000	0.0000
13	1.0000	1.0000	1.0000	0.9999	0.9991	0.9940	0.9222	0.6550	0.2677	0.0442	0.0107	0.0015	0.0000	0.0000	0.0000
14	1.0000	1.0000	1.0000	1.0000	0.9998	0.9982	0.9656	0.7878	0.4142	0.0978	0.0297	0.0056	0.0000	0.0000	0.0000
15	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9868	0.8852	0.5754	0.1894	0.0713	0.0173	0.0001	0.0000	0.0000
16	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9957	0.9461	0.7265	0.3231	0.1494	0.0468	0.0005	0.0000	0.0000
17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9988	0.9784	0.8464	0.4882	0.2735	0.1091	0.0023	0.0000	0.0000
18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9927	0.9264	0.6593	0.4389	0.2200	0.0095	0.0002	0.0000
19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9980	0.9706	0.8065	0.6217	0.3833	0.0334	0.0012	0.0000
20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9905	0.9095	0.7863	0.5793	0.0980	0.0072	0.0000
21	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9976	0.9668	0.9038	0.7660	0.2364	0.0341	0.0001
22	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9996	0.9910	0.9679	0.9018	0.4629	0.1271	0.0020
23	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9984	0.9930	0.9726	0.7288	0.3576	0.0258
24	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9992	0.9962	0.9282	0.7226	0.2222

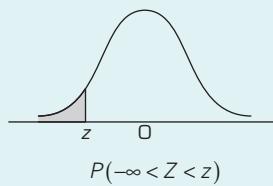
TABLE 2 Poisson Probabilities

Tabulated values are $P(X \leq k) = \sum_{x=0}^k p(x)$ (Values are rounded to four decimal places.)

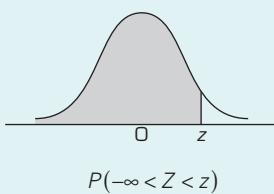
<i>k</i>	<i>μ</i>																
	0.10	0.20	0.30	0.40	0.50	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.3679	0.2231	0.1353	0.0821	0.0498	0.0302	0.0183	0.0111	0.0067	0.0041	0.0025	
1	0.9953	0.9825	0.9631	0.9384	0.9098	0.7358	0.5578	0.4060	0.2873	0.1991	0.1359	0.0916	0.0611	0.0404	0.0266	0.0174	
2	0.9998	0.9989	0.9964	0.9921	0.9856	0.9197	0.8088	0.6767	0.5438	0.4232	0.3208	0.2381	0.1736	0.1247	0.0884	0.0620	
3	1.0000	0.9999	0.9997	0.9992	0.9982	0.9810	0.9344	0.8571	0.7576	0.6472	0.5366	0.4335	0.3423	0.2650	0.2017	0.1512	
4		1.0000	1.0000	0.9999	0.9998	0.9963	0.9814	0.9473	0.8912	0.8153	0.7254	0.6288	0.5321	0.4405	0.3575	0.2851	
5				1.0000	1.0000	0.9994	0.9955	0.9834	0.9580	0.9161	0.8576	0.7851	0.7029	0.6160	0.5289	0.4457	
6					0.9999	0.9991	0.9955	0.9858	0.9665	0.9347	0.8893	0.8311	0.7622	0.6860	0.6063		
7						1.0000	0.9998	0.9989	0.9958	0.9881	0.9733	0.9489	0.9134	0.8666	0.8095	0.7440	
8							1.0000	0.9998	0.9989	0.9962	0.9901	0.9786	0.9597	0.9319	0.8944	0.8472	
9								1.0000	0.9997	0.9989	0.9967	0.9919	0.9829	0.9682	0.9462	0.9161	
10									0.9999	0.9997	0.9990	0.9972	0.9933	0.9863	0.9747	0.9574	
11										1.0000	0.9999	0.9997	0.9991	0.9976	0.9945	0.9890	0.9799
12											1.0000	0.9999	0.9997	0.9992	0.9980	0.9955	0.9912
13												1.0000	0.9999	0.9997	0.9993	0.9983	0.9964
14													1.0000	0.9999	0.9998	0.9994	0.9986
15														1.0000	0.9999	0.9998	0.9995
16															1.0000	0.9999	0.9998
17																1.0000	0.9999
18																	1.0000
19																	
20																	

TABLE 2 Poisson Probabilities (Continued)

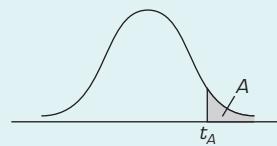
<i>k</i>	μ												
	6.50	7.00	7.50	8.00	8.50	9.00	9.50	10	11	12	13	14	15
0	0.0015	0.0009	0.0006	0.0003	0.0002	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.0113	0.0073	0.0047	0.0030	0.0019	0.0012	0.0008	0.0005	0.0002	0.0001	0.0000	0.0000	0.0000
2	0.0430	0.0296	0.0203	0.0138	0.0093	0.0062	0.0042	0.0028	0.0012	0.0005	0.0002	0.0001	0.0000
3	0.1118	0.0818	0.0591	0.0424	0.0301	0.0212	0.0149	0.0103	0.0049	0.0023	0.0011	0.0005	0.0002
4	0.2237	0.1730	0.1321	0.0996	0.0744	0.0550	0.0403	0.0293	0.0151	0.0076	0.0037	0.0018	0.0009
5	0.3690	0.3007	0.2414	0.1912	0.1496	0.1157	0.0885	0.0671	0.0375	0.0203	0.0107	0.0055	0.0028
6	0.5265	0.4497	0.3782	0.3134	0.2562	0.2068	0.1649	0.1301	0.0786	0.0458	0.0259	0.0142	0.0076
7	0.6728	0.5987	0.5246	0.4530	0.3856	0.3239	0.2687	0.2202	0.1432	0.0895	0.0540	0.0316	0.0180
8	0.7916	0.7291	0.6620	0.5925	0.5231	0.4557	0.3918	0.3328	0.2320	0.1550	0.0998	0.0621	0.0374
9	0.8774	0.8305	0.7764	0.7166	0.6530	0.5874	0.5218	0.4579	0.3405	0.2424	0.1658	0.1094	0.0699
10	0.9332	0.9015	0.8622	0.8159	0.7634	0.7060	0.6453	0.5830	0.4599	0.3472	0.2517	0.1757	0.1185
11	0.9661	0.9467	0.9208	0.8881	0.8487	0.8030	0.7520	0.6968	0.5793	0.4616	0.3532	0.2600	0.1848
12	0.9840	0.9730	0.9573	0.9362	0.9091	0.8758	0.8364	0.7916	0.6887	0.5760	0.4631	0.3585	0.2676
13	0.9929	0.9872	0.9784	0.9658	0.9486	0.9261	0.8981	0.8645	0.7813	0.6815	0.5730	0.4644	0.3632
14	0.9970	0.9943	0.9897	0.9827	0.9726	0.9585	0.9400	0.9165	0.8540	0.7720	0.6751	0.5704	0.4657
15	0.9988	0.9976	0.9954	0.9918	0.9862	0.9780	0.9665	0.9513	0.9074	0.8444	0.7636	0.6694	0.5681
16	0.9996	0.9990	0.9980	0.9963	0.9934	0.9889	0.9823	0.9730	0.9441	0.8987	0.8355	0.7559	0.6641
17	0.9998	0.9996	0.9992	0.9984	0.9970	0.9947	0.9911	0.9857	0.9678	0.9370	0.8905	0.8272	0.7489
18	0.9999	0.9999	0.9997	0.9993	0.9987	0.9976	0.9957	0.9928	0.9823	0.9626	0.9302	0.8826	0.8195
19	1.0000	1.0000	0.9999	0.9997	0.9995	0.9989	0.9980	0.9965	0.9907	0.9787	0.9573	0.9235	0.8752
20			1.0000	0.9999	0.9998	0.9996	0.9991	0.9984	0.9953	0.9884	0.9750	0.9521	0.9170
21				1.0000	0.9999	0.9998	0.9996	0.9993	0.9977	0.9939	0.9859	0.9712	0.9469
22					1.0000	0.9999	0.9999	0.9997	0.9990	0.9970	0.9924	0.9833	0.9673
23						1.0000	0.9999	0.9999	0.9995	0.9985	0.9960	0.9907	0.9805
24							1.0000	1.0000	0.9998	0.9993	0.9980	0.9950	0.9888
25									0.9999	0.9997	0.9990	0.9974	0.9938
26										1.0000	0.9999	0.9995	0.9987
27											0.9999	0.9998	0.9994
28											1.0000	0.9999	0.9997
29												1.0000	0.9999
30													0.9999
31													1.0000
32													1.0000

TABLE 3 Cumulative Standardised Normal Probabilities

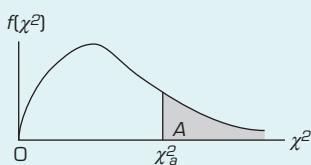
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

TABLE 3 Cumulative Standardised Normal Probabilities (Continued)

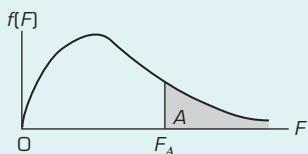
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

TABLE 4 Critical Values of the Student *t* Distribution

Degrees of freedom	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.010}$	$t_{0.005}$	Degrees of freedom	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.010}$	$t_{0.005}$
1	3.078	6.314	12.706	31.821	63.657	29	1.311	1.699	2.045	2.462	2.756
2	1.886	2.920	4.303	6.965	9.925	30	1.310	1.697	2.042	2.457	2.750
3	1.638	2.353	3.182	4.541	5.841	35	1.306	1.690	2.030	2.438	2.724
4	1.533	2.132	2.776	3.747	4.604	40	1.303	1.684	2.021	2.423	2.704
5	1.476	2.015	2.571	3.365	4.032	45	1.301	1.679	2.014	2.412	2.690
6	1.440	1.943	2.447	3.143	3.707	50	1.299	1.676	2.009	2.403	2.678
7	1.415	1.895	2.365	2.998	3.499	55	1.297	1.673	2.004	2.396	2.668
8	1.397	1.860	2.306	2.896	3.355	60	1.296	1.671	2.000	2.390	2.660
9	1.383	1.833	2.262	2.821	3.250	65	1.295	1.669	1.997	2.385	2.654
10	1.372	1.812	2.228	2.764	3.169	70	1.294	1.667	1.994	2.381	2.648
11	1.363	1.796	2.201	2.718	3.106	75	1.293	1.665	1.992	2.377	2.643
12	1.356	1.782	2.179	2.681	3.055	80	1.292	1.664	1.990	2.374	2.639
13	1.350	1.771	2.160	2.650	3.012	85	1.292	1.663	1.988	2.371	2.635
14	1.345	1.761	2.145	2.624	2.977	90	1.291	1.662	1.987	2.368	2.632
15	1.341	1.753	2.131	2.602	2.947	95	1.291	1.661	1.985	2.366	2.629
16	1.337	1.746	2.120	2.583	2.921	100	1.290	1.660	1.984	2.364	2.626
17	1.333	1.740	2.110	2.567	2.898	110	1.289	1.659	1.982	2.361	2.621
18	1.330	1.734	2.101	2.552	2.878	120	1.289	1.658	1.980	2.358	2.617
19	1.328	1.729	2.093	2.539	2.861	130	1.288	1.657	1.978	2.355	2.614
20	1.325	1.725	2.086	2.528	2.845	140	1.288	1.656	1.977	2.353	2.611
21	1.323	1.721	2.080	2.518	2.831	150	1.287	1.655	1.976	2.351	2.609
22	1.321	1.717	2.074	2.508	2.819	160	1.287	1.654	1.975	2.350	2.607
23	1.319	1.714	2.069	2.500	2.807	170	1.287	1.654	1.974	2.348	2.605
24	1.318	1.711	2.064	2.492	2.797	180	1.286	1.653	1.973	2.347	2.603
25	1.316	1.708	2.060	2.485	2.787	190	1.286	1.653	1.973	2.346	2.602
26	1.315	1.706	2.056	2.479	2.779	200	1.286	1.653	1.972	2.345	2.601
27	1.314	1.703	2.052	2.473	2.771	∞	1.282	1.645	1.960	2.326	2.576
28	1.313	1.701	2.048	2.467	2.763						

TABLE 5 Critical Values of the χ^2 Distribution

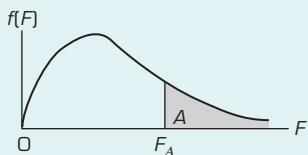
Degrees of Freedom	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000039	0.000157	0.000982	0.00393	0.0158	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	4.61	5.99	7.38	9.21	10.6
3	0.072	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.3	12.8
4	0.207	0.297	0.484	0.711	1.06	7.78	9.49	11.1	13.3	14.9
5	0.412	0.554	0.831	1.15	1.61	9.24	11.1	12.8	15.1	16.7
6	0.676	0.872	1.24	1.64	2.20	10.6	12.6	14.4	16.8	18.5
7	0.989	1.24	1.69	2.17	2.83	12.0	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	3.49	13.4	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	4.17	14.7	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	4.87	16.0	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.57	5.58	17.3	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	6.30	18.5	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	7.04	19.8	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	7.79	21.1	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	8.55	22.3	25.0	27.5	30.6	32.8
16	5.14	5.81	6.91	7.96	9.31	23.5	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	10.1	24.8	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	10.9	26.0	28.9	31.5	34.8	37.2
19	6.84	7.63	8.91	10.1	11.7	27.2	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	12.4	28.4	31.4	34.2	37.6	40.0
21	8.03	8.90	10.3	11.6	13.2	29.6	32.7	35.5	38.9	41.4
22	8.64	9.54	11.0	12.3	14.0	30.8	33.9	36.8	40.3	42.8
23	9.26	10.2	11.7	13.1	14.8	32.0	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.8	15.7	33.2	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	16.5	34.4	37.7	40.6	44.3	46.9
26	11.2	12.2	13.8	15.4	17.3	35.6	38.9	41.9	45.6	48.3
27	11.8	12.9	14.6	16.2	18.1	36.7	40.1	43.2	47.0	49.6
28	12.5	13.6	15.3	16.9	18.9	37.9	41.3	44.5	48.3	51.0
29	13.1	14.3	16.0	17.7	19.8	39.1	42.6	45.7	49.6	52.3
30	13.8	15.0	16.8	18.5	20.6	40.3	43.8	47.0	50.9	53.7
40	20.7	22.2	24.4	26.5	29.1	51.8	55.8	59.3	63.7	66.8
50	28.0	29.7	32.4	34.8	37.7	63.2	67.5	71.4	76.2	79.5
60	35.5	37.5	40.5	43.2	46.5	74.4	79.1	83.3	88.4	92.0
70	43.3	45.4	48.8	51.7	55.3	85.5	90.5	95.0	100	104
80	51.2	53.5	57.2	60.4	64.3	96.6	102	107	112	116
90	59.2	61.8	65.6	69.1	73.3	108	113	118	124	128
100	67.3	70.1	74.2	77.9	82.4	118	124	130	136	140

TABLE 6(a) Values of the F -Distribution: $A = 0.05$ 

		Numerator degrees of freedom																				
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
v_2	v_1	161	199	216	225	230	234	237	239	241	242	243	244	245	245	246	246	246	247	247	248	248
	2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.70	8.69	8.68	8.67	8.67	8.66		
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.86	5.84	5.83	5.82	5.81	5.80		
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64	4.62	4.60	4.59	4.58	4.57	4.56		
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94	3.92	3.91	3.90	3.88	3.87		
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51	3.49	3.48	3.47	3.46	3.44		
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.22	3.20	3.19	3.17	3.16	3.15		
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01	2.99	2.97	2.96	2.95	2.94		
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.85	2.83	2.81	2.80	2.79	2.77		
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72	2.70	2.69	2.67	2.66	2.65		
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62	2.60	2.58	2.57	2.56	2.54		
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53	2.51	2.50	2.48	2.47	2.46		
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.46	2.44	2.43	2.41	2.40	2.39		
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40	2.38	2.37	2.35	2.34	2.33		
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.40	2.37	2.35	2.33	2.31	2.29	2.27	2.26		
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.35	2.33	2.31	2.29	2.27	2.26	2.24	2.23		
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.31	2.29	2.27	2.25	2.23	2.22	2.20	2.19		
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.28	2.26	2.23	2.21	2.20	2.18	2.17	2.16		
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.25	2.22	2.20	2.18	2.17	2.15	2.14	2.12		
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.20	2.17	2.15	2.13	2.11	2.10	2.08	2.07		
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.15	2.13	2.11	2.09	2.07	2.05	2.04	2.03		
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.12	2.09	2.07	2.05	2.03	2.02	2.00	1.99		
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12	2.09	2.06	2.04	2.02	2.00	1.99	1.97	1.96		
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.06	2.04	2.01	1.99	1.98	1.96	1.95	1.93		
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.07	2.04	2.01	1.99	1.96	1.94	1.92	1.91	1.89	1.88		
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.97	1.95	1.92	1.90	1.89	1.87	1.85	1.84		
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05	2.01	1.97	1.94	1.92	1.89	1.87	1.86	1.84	1.82	1.81		
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.99	1.95	1.92	1.89	1.87	1.85	1.83	1.81	1.80	1.78		
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.89	1.86	1.84	1.82	1.80	1.78	1.76	1.75		
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.93	1.89	1.86	1.84	1.81	1.79	1.77	1.75	1.74	1.72		
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.91	1.88	1.84	1.82	1.79	1.77	1.75	1.73	1.72	1.70		
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.90	1.86	1.83	1.80	1.78	1.76	1.74	1.72	1.70	1.69		
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.89	1.85	1.82	1.79	1.77	1.75	1.73	1.71	1.69	1.68		
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.87	1.83	1.80	1.78	1.75	1.73	1.71	1.69	1.67	1.66		
140	3.91	3.06	2.67	2.44	2.28	2.16	2.08	2.01	1.95	1.90	1.86	1.82	1.79	1.76	1.74	1.72	1.70	1.68	1.66	1.65		
160	3.90	3.05	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.81	1.78	1.75	1.73	1.71	1.69	1.67	1.65	1.64		
180	3.89	3.05	2.65	2.42	2.26	2.15	2.06	1.99	1.93	1.88	1.84	1.81	1.77	1.75	1.72	1.70	1.68	1.66	1.64	1.63		
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.84	1.80	1.77	1.74	1.72	1.69	1.67	1.66	1.64	1.62		
∞	3.84	3.00	2.61	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.79	1.75	1.72	1.69	1.67	1.64	1.62	1.60	1.59	1.57		

TABLE 6(a) Values of the *F*-Distribution: $A = 0.05$ [Continued]

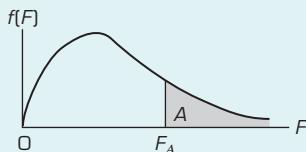
$v_2 \backslash v_1$	Numerator degrees of freedom																		
	22	24	26	28	30	35	40	45	50	60	70	80	90	100	120	140	160	180	200
1	249	249	249	250	250	251	251	251	252	252	252	253	253	253	253	254	254	254	254
2	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5
3	8.65	8.64	8.63	8.62	8.62	8.60	8.59	8.59	8.58	8.57	8.57	8.56	8.56	8.55	8.55	8.55	8.54	8.54	8.53
4	5.79	5.77	5.76	5.75	5.75	5.73	5.72	5.71	5.70	5.69	5.68	5.67	5.67	5.66	5.66	5.65	5.65	5.65	5.63
5	4.54	4.53	4.52	4.50	4.50	4.48	4.46	4.45	4.44	4.43	4.42	4.41	4.41	4.41	4.40	4.39	4.39	4.39	4.37
6	3.86	3.84	3.83	3.82	3.81	3.79	3.77	3.76	3.75	3.74	3.73	3.72	3.72	3.71	3.70	3.70	3.69	3.69	3.67
7	3.43	3.41	3.40	3.39	3.38	3.36	3.34	3.33	3.32	3.30	3.29	3.29	3.28	3.27	3.27	3.26	3.26	3.25	3.23
8	3.13	3.12	3.10	3.09	3.08	3.06	3.04	3.03	3.02	3.01	2.99	2.99	2.98	2.97	2.97	2.96	2.96	2.95	2.93
9	2.92	2.90	2.89	2.87	2.86	2.84	2.83	2.81	2.80	2.79	2.78	2.77	2.76	2.76	2.75	2.74	2.74	2.73	2.71
10	2.75	2.74	2.72	2.71	2.70	2.68	2.66	2.65	2.64	2.62	2.61	2.60	2.59	2.59	2.58	2.57	2.57	2.57	2.54
11	2.63	2.61	2.59	2.58	2.57	2.55	2.53	2.52	2.51	2.49	2.48	2.47	2.46	2.46	2.45	2.44	2.44	2.43	2.41
12	2.52	2.51	2.49	2.48	2.47	2.44	2.43	2.41	2.40	2.38	2.37	2.36	2.36	2.35	2.34	2.33	2.33	2.33	2.30
13	2.44	2.42	2.41	2.39	2.38	2.36	2.34	2.33	2.31	2.30	2.28	2.27	2.27	2.26	2.25	2.25	2.24	2.24	2.21
14	2.37	2.35	2.33	2.32	2.31	2.28	2.27	2.25	2.24	2.22	2.21	2.20	2.19	2.19	2.18	2.17	2.17	2.16	2.13
15	2.31	2.29	2.27	2.26	2.25	2.22	2.20	2.19	2.18	2.16	2.15	2.14	2.13	2.12	2.11	2.11	2.10	2.10	2.07
16	2.25	2.24	2.22	2.21	2.19	2.17	2.15	2.14	2.12	2.11	2.09	2.08	2.07	2.07	2.06	2.05	2.05	2.04	2.01
17	2.21	2.19	2.17	2.16	2.15	2.12	2.10	2.09	2.08	2.06	2.05	2.03	2.03	2.02	2.01	2.00	2.00	1.99	1.96
18	2.17	2.15	2.13	2.12	2.11	2.08	2.06	2.05	2.04	2.02	2.00	1.99	1.98	1.98	1.97	1.96	1.96	1.95	1.92
19	2.13	2.11	2.10	2.08	2.07	2.05	2.03	2.01	2.00	1.98	1.97	1.96	1.95	1.94	1.93	1.92	1.92	1.91	1.88
20	2.10	2.08	2.07	2.05	2.04	2.01	1.99	1.98	1.97	1.95	1.93	1.92	1.91	1.91	1.90	1.89	1.88	1.88	1.84
22	2.05	2.03	2.01	2.00	1.98	1.96	1.94	1.92	1.91	1.89	1.88	1.86	1.86	1.85	1.84	1.83	1.82	1.82	1.78
24	2.00	1.98	1.97	1.95	1.94	1.91	1.89	1.88	1.86	1.84	1.83	1.82	1.81	1.80	1.79	1.78	1.78	1.77	1.73
26	1.97	1.95	1.93	1.91	1.90	1.87	1.85	1.84	1.82	1.80	1.79	1.78	1.77	1.76	1.75	1.74	1.73	1.73	1.69
28	1.93	1.91	1.90	1.88	1.87	1.84	1.82	1.80	1.79	1.77	1.75	1.74	1.73	1.73	1.71	1.71	1.70	1.69	1.65
30	1.91	1.89	1.87	1.85	1.84	1.81	1.79	1.77	1.76	1.74	1.72	1.71	1.70	1.70	1.68	1.68	1.67	1.66	1.62
35	1.85	1.83	1.82	1.80	1.79	1.76	1.74	1.72	1.70	1.68	1.66	1.65	1.64	1.63	1.62	1.61	1.61	1.60	1.56
40	1.81	1.79	1.77	1.76	1.74	1.72	1.69	1.67	1.66	1.64	1.62	1.61	1.60	1.59	1.58	1.57	1.56	1.55	1.51
45	1.78	1.76	1.74	1.73	1.71	1.68	1.66	1.64	1.63	1.60	1.59	1.57	1.56	1.55	1.54	1.53	1.52	1.52	1.47
50	1.76	1.74	1.72	1.70	1.69	1.66	1.63	1.61	1.60	1.58	1.56	1.54	1.53	1.52	1.51	1.50	1.49	1.49	1.44
60	1.72	1.70	1.68	1.66	1.65	1.62	1.59	1.57	1.56	1.53	1.52	1.50	1.49	1.48	1.47	1.46	1.45	1.44	1.39
70	1.70	1.67	1.65	1.64	1.62	1.59	1.57	1.55	1.53	1.50	1.49	1.47	1.46	1.45	1.44	1.42	1.42	1.41	1.35
80	1.68	1.65	1.63	1.62	1.60	1.57	1.54	1.52	1.51	1.48	1.46	1.45	1.44	1.43	1.41	1.40	1.39	1.38	1.33
90	1.66	1.64	1.62	1.60	1.59	1.55	1.53	1.51	1.49	1.46	1.44	1.43	1.42	1.41	1.39	1.38	1.37	1.36	1.30
100	1.65	1.63	1.61	1.59	1.57	1.54	1.52	1.49	1.48	1.45	1.43	1.41	1.40	1.39	1.38	1.36	1.35	1.34	1.28
120	1.63	1.61	1.59	1.57	1.55	1.52	1.50	1.47	1.46	1.43	1.41	1.39	1.38	1.37	1.35	1.34	1.33	1.32	1.26
140	1.62	1.60	1.57	1.56	1.54	1.51	1.48	1.46	1.44	1.41	1.39	1.36	1.35	1.33	1.32	1.30	1.29	1.28	1.23
160	1.61	1.59	1.57	1.55	1.53	1.50	1.47	1.45	1.43	1.40	1.38	1.36	1.35	1.34	1.32	1.31	1.30	1.29	1.22
180	1.60	1.58	1.56	1.54	1.52	1.49	1.46	1.44	1.42	1.39	1.37	1.35	1.34	1.33	1.31	1.30	1.29	1.28	1.20
200	1.60	1.57	1.55	1.53	1.52	1.48	1.46	1.43	1.41	1.39	1.36	1.35	1.33	1.32	1.30	1.29	1.28	1.27	1.19
∞	1.54	1.52	1.50	1.48	1.46	1.42	1.40	1.37	1.35	1.32	1.29	1.28	1.26	1.25	1.22	1.21	1.19	1.18	1.17

TABLE 6(b) Values of the F -Distribution: $A = 0.025$ 

$v_2 \backslash v_1$	Numerator degrees of freedom																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	648	799	864	900	922	937	948	957	963	969	973	977	980	983	985	987	989	990	992	993
2	38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.4
3	17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.4	14.4	14.3	14.3	14.3	14.3	14.2	14.2	14.2	14.2	14.2
4	12.2	10.6	10.0	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.79	8.75	8.71	8.68	8.66	8.63	8.61	8.59	8.58	8.56
5	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.57	6.52	6.49	6.46	6.43	6.40	6.38	6.36	6.34	6.33
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.41	5.37	5.33	5.30	5.27	5.24	5.22	5.20	5.18	5.17
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.71	4.67	4.63	4.60	4.57	4.54	4.52	4.50	4.48	4.47
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.24	4.20	4.16	4.13	4.10	4.08	4.05	4.03	4.02	4.00
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.91	3.87	3.83	3.80	3.77	3.74	3.72	3.70	3.68	3.67
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.66	3.62	3.58	3.55	3.52	3.50	3.47	3.45	3.44	3.42
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.47	3.43	3.39	3.36	3.33	3.30	3.28	3.26	3.24	3.23
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.32	3.28	3.24	3.21	3.18	3.15	3.13	3.11	3.09	3.07
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.20	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96	2.95
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.09	3.05	3.01	2.98	2.95	2.92	2.90	2.88	2.86	2.84
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	3.01	2.96	2.92	2.89	2.86	2.84	2.81	2.79	2.77	2.76
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.93	2.89	2.85	2.82	2.79	2.76	2.74	2.72	2.70	2.68
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.87	2.82	2.79	2.75	2.72	2.70	2.67	2.65	2.63	2.62
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.81	2.77	2.73	2.70	2.67	2.64	2.62	2.60	2.58	2.56
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.76	2.72	2.68	2.65	2.62	2.59	2.57	2.55	2.53	2.51
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.72	2.68	2.64	2.60	2.57	2.55	2.52	2.50	2.48	2.46
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.65	2.60	2.56	2.53	2.50	2.47	2.45	2.43	2.41	2.39
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.59	2.54	2.50	2.47	2.44	2.41	2.39	2.36	2.35	2.33
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.54	2.49	2.45	2.42	2.39	2.36	2.34	2.31	2.29	2.28
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.49	2.45	2.41	2.37	2.34	2.32	2.29	2.27	2.25	2.23
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.46	2.41	2.37	2.34	2.31	2.28	2.26	2.23	2.21	2.20
35	5.48	4.11	3.52	3.18	2.96	2.80	2.68	2.58	2.50	2.44	2.39	2.34	2.30	2.27	2.23	2.21	2.18	2.16	2.14	2.12
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.13	2.11	2.09	2.07
45	5.38	4.01	3.42	3.09	2.86	2.70	2.58	2.49	2.41	2.35	2.29	2.25	2.21	2.17	2.14	2.11	2.09	2.07	2.04	2.03
50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.26	2.22	2.18	2.14	2.11	2.08	2.06	2.03	2.01	1.99
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.22	2.17	2.13	2.09	2.06	2.03	2.01	1.98	1.96	1.94
70	5.25	3.89	3.31	2.97	2.75	2.59	2.47	2.38	2.30	2.24	2.18	2.14	2.10	2.06	2.03	2.00	1.97	1.95	1.93	1.91
80	5.22	3.86	3.28	2.95	2.73	2.57	2.45	2.35	2.28	2.21	2.16	2.11	2.07	2.03	2.00	1.97	1.95	1.92	1.90	1.88
90	5.20	3.84	3.26	2.93	2.71	2.55	2.43	2.34	2.26	2.19	2.14	2.09	2.05	2.02	1.98	1.95	1.93	1.91	1.88	1.86
100	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	2.12	2.08	2.04	2.00	1.97	1.94	1.91	1.89	1.87	1.85
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.10	2.05	2.01	1.98	1.94	1.92	1.89	1.87	1.84	1.82
140	5.13	3.79	3.21	2.88	2.66	2.50	2.38	2.28	2.21	2.14	2.09	2.04	2.00	1.96	1.93	1.90	1.87	1.85	1.83	1.81
160	5.12	3.78	3.20	2.87	2.65	2.49	2.37	2.27	2.19	2.13	2.07	2.03	1.99	1.95	1.92	1.89	1.86	1.84	1.82	1.80
180	5.11	3.77	3.19	2.86	2.64	2.48	2.36	2.26	2.19	2.12	2.07	2.02	1.98	1.94	1.91	1.88	1.85	1.83	1.81	1.79
200	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18	2.11	2.06	2.01	1.97	1.93	1.90	1.87	1.84	1.82	1.80	1.78
∞	5.03	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.99	1.95	1.90	1.87	1.83	1.80	1.78	1.75	1.73	1.71

TABLE 6(b) Values of the *F*-Distribution: $A = 0.025$ (Continued)

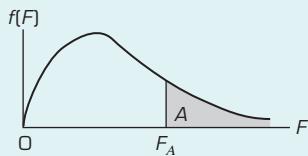
ν_2	ν_1	Numerator degrees of freedom																		
		22	24	26	28	30	35	40	45	50	60	70	80	90	100	120	140	160	180	200
Denominator degrees of freedom	1	995	997	999	1000	1001	1004	1006	1007	1008	1010	1011	1012	1013	1013	1014	1015	1015	1015	1016
	2	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5	39.5
	3	14.1	14.1	14.1	14.1	14.1	14.1	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	13.9	13.9	13.9	13.9	13.9
	4	8.53	8.51	8.49	8.48	8.46	8.43	8.41	8.39	8.38	8.36	8.35	8.33	8.33	8.32	8.31	8.30	8.30	8.29	8.29
	5	6.30	6.28	6.26	6.24	6.23	6.20	6.18	6.16	6.14	6.12	6.11	6.10	6.09	6.08	6.07	6.06	6.06	6.05	6.05
	6	5.14	5.12	5.10	5.08	5.07	5.04	5.01	4.99	4.98	4.96	4.94	4.93	4.92	4.92	4.90	4.90	4.89	4.89	4.88
	7	4.44	4.41	4.39	4.38	4.36	4.33	4.31	4.29	4.28	4.25	4.24	4.23	4.22	4.21	4.20	4.19	4.18	4.18	4.14
	8	3.97	3.95	3.93	3.91	3.89	3.86	3.84	3.82	3.81	3.78	3.77	3.76	3.75	3.74	3.73	3.72	3.71	3.71	3.67
	9	3.64	3.61	3.59	3.58	3.56	3.53	3.51	3.49	3.47	3.45	3.43	3.42	3.41	3.40	3.39	3.38	3.38	3.37	3.33
	10	3.39	3.37	3.34	3.33	3.31	3.28	3.26	3.24	3.22	3.20	3.18	3.17	3.16	3.15	3.14	3.13	3.13	3.12	3.08
	11	3.20	3.17	3.15	3.13	3.12	3.09	3.06	3.04	3.03	3.00	2.99	2.97	2.96	2.96	2.94	2.94	2.93	2.92	2.92
	12	3.04	3.02	3.00	2.98	2.96	2.93	2.91	2.89	2.87	2.85	2.83	2.82	2.81	2.80	2.79	2.78	2.77	2.77	2.73
	13	2.92	2.89	2.87	2.85	2.84	2.80	2.78	2.76	2.74	2.72	2.70	2.69	2.68	2.67	2.66	2.65	2.64	2.64	2.63
	14	2.81	2.79	2.77	2.75	2.73	2.70	2.67	2.65	2.64	2.61	2.60	2.58	2.57	2.56	2.55	2.54	2.54	2.53	2.49
	15	2.73	2.70	2.68	2.66	2.64	2.61	2.59	2.56	2.55	2.52	2.51	2.49	2.48	2.47	2.46	2.45	2.44	2.44	2.40
	16	2.65	2.63	2.60	2.58	2.57	2.53	2.51	2.49	2.47	2.45	2.43	2.42	2.40	2.40	2.38	2.37	2.37	2.36	2.36
	17	2.59	2.56	2.54	2.52	2.50	2.47	2.44	2.42	2.41	2.38	2.36	2.35	2.34	2.33	2.32	2.31	2.30	2.29	2.29
	18	2.53	2.50	2.48	2.46	2.44	2.41	2.38	2.36	2.35	2.32	2.30	2.29	2.28	2.27	2.26	2.25	2.24	2.23	2.23
	19	2.48	2.45	2.43	2.41	2.39	2.36	2.33	2.31	2.30	2.27	2.25	2.24	2.23	2.22	2.20	2.19	2.19	2.18	2.18
	20	2.43	2.41	2.39	2.37	2.35	2.31	2.29	2.27	2.25	2.22	2.20	2.19	2.18	2.17	2.16	2.15	2.14	2.13	2.09
	22	2.36	2.33	2.31	2.29	2.27	2.24	2.21	2.19	2.17	2.14	2.13	2.11	2.10	2.09	2.08	2.07	2.06	2.05	2.00
	24	2.30	2.27	2.25	2.23	2.21	2.17	2.15	2.12	2.11	2.08	2.06	2.05	2.03	2.02	2.01	2.00	1.99	1.99	1.98
	26	2.24	2.22	2.19	2.17	2.16	2.12	2.09	2.07	2.05	2.03	2.01	1.99	1.98	1.97	1.95	1.94	1.94	1.93	1.92
	28	2.20	2.17	2.15	2.13	2.11	2.08	2.05	2.03	2.01	1.98	1.96	1.94	1.93	1.92	1.91	1.90	1.89	1.88	1.88
	30	2.16	2.14	2.11	2.09	2.07	2.04	2.01	1.99	1.97	1.94	1.92	1.90	1.89	1.88	1.87	1.86	1.85	1.84	1.79
	35	2.09	2.06	2.04	2.02	2.00	1.96	1.93	1.91	1.89	1.86	1.84	1.82	1.81	1.80	1.79	1.77	1.76	1.75	1.70
	40	2.03	2.01	1.98	1.96	1.94	1.90	1.88	1.85	1.83	1.80	1.78	1.76	1.75	1.74	1.72	1.71	1.70	1.70	1.69
	45	1.99	1.96	1.94	1.92	1.90	1.86	1.83	1.81	1.79	1.76	1.74	1.72	1.70	1.69	1.68	1.66	1.66	1.65	1.64
	50	1.96	1.93	1.91	1.89	1.87	1.83	1.80	1.77	1.75	1.72	1.70	1.68	1.67	1.66	1.64	1.63	1.62	1.61	1.60
	60	1.91	1.88	1.86	1.83	1.82	1.78	1.74	1.72	1.70	1.67	1.64	1.63	1.61	1.60	1.58	1.57	1.56	1.55	1.54
	70	1.88	1.85	1.82	1.80	1.78	1.74	1.71	1.68	1.66	1.63	1.60	1.59	1.57	1.56	1.54	1.53	1.52	1.51	1.50
	80	1.85	1.82	1.79	1.77	1.75	1.71	1.68	1.65	1.63	1.60	1.57	1.55	1.54	1.53	1.51	1.49	1.48	1.47	1.40
	90	1.83	1.80	1.77	1.75	1.73	1.69	1.66	1.63	1.61	1.58	1.55	1.53	1.52	1.50	1.48	1.47	1.46	1.45	1.44
	100	1.81	1.78	1.76	1.74	1.71	1.67	1.64	1.61	1.59	1.56	1.53	1.51	1.50	1.48	1.46	1.45	1.44	1.43	1.42
	120	1.79	1.76	1.73	1.71	1.69	1.65	1.61	1.59	1.56	1.53	1.50	1.48	1.47	1.45	1.43	1.41	1.40	1.39	1.31
	140	1.77	1.74	1.72	1.69	1.67	1.63	1.60	1.57	1.55	1.51	1.48	1.46	1.45	1.43	1.42	1.39	1.38	1.36	1.28
	160	1.76	1.73	1.70	1.68	1.66	1.62	1.58	1.55	1.53	1.50	1.47	1.45	1.43	1.42	1.39	1.38	1.36	1.35	1.35
	180	1.75	1.72	1.69	1.67	1.65	1.61	1.57	1.54	1.52	1.48	1.46	1.43	1.42	1.40	1.38	1.36	1.35	1.34	1.33
	200	1.74	1.71	1.68	1.66	1.64	1.60	1.56	1.53	1.51	1.47	1.45	1.42	1.41	1.39	1.37	1.35	1.34	1.33	1.23
	∞	1.67	1.64	1.61	1.59	1.57	1.52	1.49	1.46	1.43	1.39	1.36	1.33	1.31	1.30	1.27	1.25	1.23	1.22	1.21

TABLE 6(c) Values of the *F*-Distribution: $A = 0.01$ 

$v_2 \backslash v_1$	Numerator degrees of freedom																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056	6083	6106	6126	6143	6157	6170	6181	6192	6201	6209
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	27.1	27.0	26.9	26.9	26.8	26.8	26.8	26.7	26.7
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.5	14.4	14.3	14.2	14.2	14.2	14.1	14.1	14.0	14.0
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.96	9.89	9.82	9.77	9.72	9.68	9.64	9.61	9.58	9.55
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.66	7.60	7.56	7.52	7.48	7.45	7.42	7.40
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47	6.41	6.36	6.31	6.28	6.24	6.21	6.18	6.16
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67	5.61	5.56	5.52	5.48	5.44	5.41	5.38	5.36
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11	5.05	5.01	4.96	4.92	4.89	4.86	4.83	4.81
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71	4.65	4.60	4.56	4.52	4.49	4.46	4.43	4.41
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.34	4.29	4.25	4.21	4.18	4.15	4.12	4.10
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.10	4.05	4.01	3.97	3.94	3.91	3.88	3.86
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.91	3.86	3.82	3.78	3.75	3.72	3.69	3.66
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.75	3.70	3.66	3.62	3.59	3.56	3.53	3.51
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.61	3.56	3.52	3.49	3.45	3.42	3.40	3.37
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.55	3.50	3.45	3.41	3.37	3.34	3.31	3.28	3.26
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.46	3.40	3.35	3.31	3.27	3.24	3.21	3.19	3.16
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37	3.32	3.27	3.23	3.19	3.16	3.13	3.10	3.08
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.24	3.19	3.15	3.12	3.08	3.05	3.03	3.00
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23	3.18	3.13	3.09	3.05	3.02	2.99	2.96	2.94
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	3.07	3.02	2.98	2.94	2.91	2.88	2.85	2.83
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09	3.03	2.98	2.93	2.89	2.85	2.82	2.79	2.76	2.74
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	3.02	2.96	2.90	2.86	2.81	2.78	2.75	2.72	2.69	2.66
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.96	2.90	2.84	2.79	2.75	2.72	2.68	2.65	2.63	2.60
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91	2.84	2.79	2.74	2.70	2.66	2.63	2.60	2.57	2.55
35	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88	2.80	2.74	2.69	2.64	2.60	2.56	2.53	2.50	2.47	2.44
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.73	2.66	2.61	2.56	2.52	2.48	2.45	2.42	2.39	2.37
45	7.23	5.11	4.25	3.77	3.45	3.23	3.07	2.94	2.83	2.74	2.67	2.61	2.55	2.51	2.46	2.43	2.39	2.36	2.34	2.31
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.63	2.56	2.51	2.46	2.42	2.38	2.35	2.32	2.29	2.27
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50	2.44	2.39	2.35	2.31	2.28	2.25	2.22	2.20
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.59	2.51	2.45	2.40	2.35	2.31	2.27	2.23	2.20	2.18	2.15
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.48	2.42	2.36	2.31	2.27	2.23	2.20	2.17	2.14	2.12
90	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52	2.45	2.39	2.33	2.29	2.24	2.21	2.17	2.14	2.11	2.09
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.43	2.37	2.31	2.27	2.22	2.19	2.15	2.12	2.09	2.07
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.40	2.34	2.28	2.23	2.19	2.15	2.12	2.09	2.06	2.03
140	6.82	4.76	3.92	3.46	3.15	2.93	2.77	2.64	2.54	2.45	2.38	2.31	2.26	2.21	2.17	2.13	2.10	2.07	2.04	2.01
160	6.80	4.74	3.91	3.44	3.13	2.92	2.75	2.62	2.52	2.43	2.36	2.30	2.24	2.20	2.15	2.11	2.08	2.05	2.02	1.99
180	6.78	4.73	3.89	3.43	3.12	2.90	2.74	2.61	2.51	2.42	2.35	2.28	2.23	2.18	2.14	2.10	2.07	2.04	2.01	1.98
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.34	2.27	2.22	2.17	2.13	2.09	2.06	2.03	2.00	1.97
∞	6.64	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.25	2.19	2.13	2.08	2.04	2.00	1.97	1.94	1.91	1.88

TABLE 6(c) Values of the *F*-Distribution: $A = 0.01$ (Continued)

$v_2 \backslash v_1$	Numerator degrees of freedom																			
	22	24	26	28	30	35	40	45	50	60	70	80	90	100	120	140	160	180	200	∞
1	6223	6235	6245	6253	6261	6276	6287	6296	6303	6313	6321	6326	6331	6334	6339	6343	6346	6348	6350	6366
2	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5
3	26.6	26.6	26.6	26.5	26.5	26.5	26.4	26.4	26.4	26.3	26.3	26.3	26.3	26.2	26.2	26.2	26.2	26.2	26.2	26.1
4	14.0	13.9	13.9	13.9	13.8	13.8	13.7	13.7	13.7	13.7	13.6	13.6	13.6	13.6	13.6	13.5	13.5	13.5	13.5	13.5
5	9.51	9.47	9.43	9.40	9.38	9.33	9.29	9.26	9.24	9.20	9.18	9.16	9.14	9.13	9.11	9.10	9.09	9.08	9.08	9.02
6	7.35	7.31	7.28	7.25	7.23	7.18	7.14	7.11	7.09	7.06	7.03	7.01	7.00	6.99	6.97	6.96	6.95	6.94	6.93	6.88
7	6.11	6.07	6.04	6.02	5.99	5.94	5.91	5.88	5.86	5.82	5.80	5.78	5.77	5.75	5.74	5.72	5.72	5.71	5.70	5.65
8	5.32	5.28	5.25	5.22	5.20	5.15	5.12	5.09	5.07	5.03	5.01	4.99	4.97	4.96	4.95	4.93	4.92	4.92	4.91	4.86
9	4.77	4.73	4.70	4.67	4.65	4.60	4.57	4.54	4.52	4.48	4.46	4.44	4.43	4.41	4.40	4.39	4.38	4.37	4.36	4.31
10	4.36	4.33	4.30	4.27	4.25	4.20	4.17	4.14	4.12	4.08	4.06	4.04	4.03	4.01	4.00	3.98	3.97	3.97	3.96	3.91
11	4.06	4.02	3.99	3.96	3.94	3.89	3.86	3.83	3.81	3.78	3.75	3.73	3.72	3.71	3.69	3.68	3.67	3.66	3.66	3.60
12	3.82	3.78	3.75	3.72	3.70	3.65	3.62	3.59	3.57	3.54	3.51	3.49	3.48	3.47	3.45	3.44	3.43	3.42	3.41	3.36
13	3.62	3.59	3.56	3.53	3.51	3.46	3.43	3.40	3.38	3.34	3.32	3.30	3.28	3.27	3.25	3.24	3.23	3.23	3.22	3.17
14	3.46	3.43	3.40	3.37	3.35	3.30	3.27	3.24	3.22	3.18	3.16	3.14	3.12	3.11	3.09	3.08	3.07	3.06	3.06	3.01
15	3.33	3.29	3.26	3.24	3.21	3.17	3.13	3.10	3.08	3.05	3.02	3.00	2.99	2.98	2.96	2.95	2.94	2.93	2.92	2.87
16	3.22	3.18	3.15	3.12	3.10	3.05	3.02	2.99	2.97	2.93	2.91	2.89	2.87	2.86	2.84	2.83	2.82	2.81	2.81	2.75
17	3.12	3.08	3.05	3.03	3.00	2.96	2.92	2.89	2.87	2.83	2.81	2.79	2.78	2.76	2.75	2.73	2.72	2.72	2.71	2.65
18	3.03	3.00	2.97	2.94	2.92	2.87	2.84	2.81	2.78	2.75	2.72	2.70	2.69	2.68	2.66	2.65	2.64	2.63	2.62	2.57
19	2.96	2.92	2.89	2.87	2.84	2.80	2.76	2.73	2.71	2.67	2.65	2.63	2.61	2.60	2.58	2.57	2.56	2.55	2.55	2.49
20	2.90	2.86	2.83	2.80	2.78	2.73	2.69	2.67	2.64	2.61	2.58	2.56	2.55	2.54	2.52	2.50	2.49	2.49	2.48	2.42
22	2.78	2.75	2.72	2.69	2.67	2.62	2.58	2.55	2.53	2.50	2.47	2.45	2.43	2.42	2.40	2.39	2.38	2.37	2.36	2.31
24	2.70	2.66	2.63	2.60	2.58	2.53	2.49	2.46	2.44	2.40	2.38	2.36	2.34	2.33	2.31	2.30	2.29	2.28	2.27	2.21
26	2.62	2.58	2.55	2.53	2.50	2.45	2.42	2.39	2.36	2.33	2.30	2.28	2.26	2.25	2.23	2.22	2.21	2.20	2.19	2.13
28	2.56	2.52	2.49	2.46	2.44	2.39	2.35	2.32	2.30	2.26	2.24	2.22	2.20	2.19	2.17	2.15	2.14	2.13	2.13	2.07
30	2.51	2.47	2.44	2.41	2.39	2.34	2.30	2.27	2.25	2.21	2.18	2.16	2.14	2.13	2.11	2.10	2.09	2.08	2.07	2.01
35	2.40	2.36	2.33	2.30	2.28	2.23	2.19	2.16	2.14	2.10	2.07	2.05	2.03	2.02	2.00	1.98	1.97	1.96	1.96	1.89
40	2.33	2.29	2.26	2.23	2.20	2.15	2.11	2.08	2.06	2.02	1.99	1.97	1.95	1.94	1.92	1.90	1.89	1.88	1.87	1.81
45	2.27	2.23	2.20	2.17	2.14	2.09	2.05	2.02	2.00	1.96	1.93	1.91	1.89	1.88	1.85	1.84	1.83	1.82	1.81	1.74
50	2.22	2.18	2.15	2.12	2.10	2.05	2.01	1.97	1.95	1.91	1.88	1.86	1.84	1.82	1.80	1.79	1.77	1.76	1.76	1.68
60	2.15	2.12	2.08	2.05	2.03	1.98	1.94	1.90	1.88	1.84	1.81	1.78	1.76	1.75	1.73	1.71	1.70	1.69	1.68	1.60
70	2.11	2.07	2.03	2.01	1.98	1.93	1.89	1.85	1.83	1.78	1.75	1.73	1.71	1.70	1.67	1.65	1.64	1.63	1.62	1.54
80	2.07	2.03	2.00	1.97	1.94	1.89	1.85	1.82	1.79	1.75	1.71	1.69	1.67	1.65	1.63	1.61	1.60	1.59	1.58	1.50
90	2.04	2.00	1.97	1.94	1.92	1.86	1.82	1.79	1.76	1.72	1.68	1.66	1.64	1.62	1.60	1.58	1.57	1.55	1.55	1.46
100	2.02	1.98	1.95	1.92	1.89	1.84	1.80	1.76	1.74	1.69	1.66	1.63	1.61	1.60	1.57	1.55	1.54	1.53	1.52	1.43
120	1.99	1.95	1.92	1.89	1.86	1.81	1.76	1.73	1.70	1.66	1.62	1.60	1.58	1.56	1.53	1.51	1.50	1.49	1.48	1.38
140	1.97	1.93	1.89	1.86	1.84	1.78	1.74	1.70	1.67	1.63	1.60	1.57	1.55	1.53	1.50	1.48	1.47	1.46	1.45	1.35
160	1.95	1.91	1.88	1.85	1.82	1.76	1.72	1.68	1.66	1.61	1.58	1.55	1.53	1.51	1.48	1.46	1.45	1.43	1.42	1.32
180	1.94	1.90	1.86	1.83	1.81	1.75	1.71	1.67	1.64	1.60	1.56	1.53	1.51	1.49	1.47	1.45	1.43	1.42	1.41	1.30
200	1.93	1.89	1.85	1.82	1.79	1.74	1.69	1.66	1.63	1.58	1.55	1.52	1.50	1.48	1.45	1.43	1.42	1.40	1.39	1.28
∞	1.83	1.79	1.76	1.73	1.70	1.64	1.59	1.56	1.53	1.48	1.44	1.41	1.38	1.36	1.33	1.30	1.28	1.26	1.25	1.00

TABLE 6(d) Values of the *F*-Distribution: $A = 0.005$ 

		Numerator degrees of freedom																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Denominator degrees of freedom	1	16211	19999	21615	22500	23056	23437	23715	23925	24091	24224	24334	24426	24505	24572	24630	24681	24727	24767	24803	24836
	2	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199
	3	55.6	49.8	47.5	46.2	45.4	44.8	44.4	44.1	43.9	43.7	43.5	43.4	43.3	43.2	43.1	43.0	42.9	42.8	42.8	42.8
	4	31.3	26.3	24.3	23.2	22.5	22.0	21.6	21.4	21.1	21.0	20.8	20.7	20.6	20.5	20.4	20.4	20.3	20.3	20.2	20.2
	5	22.8	18.3	16.5	15.6	14.9	14.5	14.2	14.0	13.8	13.6	13.5	13.4	13.3	13.2	13.1	13.1	13.0	13.0	12.9	12.9
	6	18.6	14.5	12.9	12.0	11.5	11.1	10.8	10.6	10.4	10.3	10.1	10.0	9.95	9.88	9.81	9.76	9.71	9.66	9.62	9.59
	7	16.2	12.4	10.9	10.1	9.52	9.16	8.89	8.68	8.51	8.38	8.27	8.18	8.10	8.03	7.97	7.91	7.87	7.83	7.79	7.75
	8	14.7	11.0	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.10	7.01	6.94	6.87	6.81	6.76	6.72	6.68	6.64	6.61
	9	13.6	10.1	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.31	6.23	6.15	6.09	6.03	5.98	5.94	5.90	5.86	5.83
	10	12.8	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.75	5.66	5.59	5.53	5.47	5.42	5.38	5.34	5.31	5.27
	11	12.2	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42	5.32	5.24	5.16	5.10	5.05	5.00	4.96	4.92	4.89	4.86
	12	11.8	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.99	4.91	4.84	4.77	4.72	4.67	4.63	4.59	4.56	4.53
	13	11.4	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82	4.72	4.64	4.57	4.51	4.46	4.41	4.37	4.33	4.30	4.27
	14	11.1	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60	4.51	4.43	4.36	4.30	4.25	4.20	4.16	4.12	4.09	4.06
	15	10.8	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.33	4.25	4.18	4.12	4.07	4.02	3.98	3.95	3.91	3.88
	16	10.6	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27	4.18	4.10	4.03	3.97	3.92	3.87	3.83	3.80	3.76	3.73
	17	10.4	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14	4.05	3.97	3.90	3.84	3.79	3.75	3.71	3.67	3.64	3.61
	18	10.2	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03	3.94	3.86	3.79	3.73	3.68	3.64	3.60	3.56	3.53	3.50
	19	10.1	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93	3.84	3.76	3.70	3.64	3.59	3.54	3.50	3.46	3.43	3.40
	20	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.76	3.68	3.61	3.55	3.50	3.46	3.42	3.38	3.35	3.32
	22	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81	3.70	3.61	3.54	3.47	3.41	3.36	3.31	3.27	3.24	3.21	3.18
	24	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59	3.50	3.42	3.35	3.30	3.25	3.20	3.16	3.12	3.09	3.06
	26	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60	3.49	3.40	3.33	3.26	3.20	3.15	3.11	3.07	3.03	3.00	2.97
	28	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52	3.41	3.32	3.25	3.18	3.12	3.07	3.03	2.99	2.95	2.92	2.89
	30	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.25	3.18	3.11	3.06	3.01	2.96	2.92	2.89	2.85	2.82
	35	8.98	6.19	5.09	4.48	4.09	3.81	3.61	3.45	3.32	3.21	3.12	3.05	2.98	2.93	2.88	2.83	2.79	2.76	2.72	2.69
	40	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12	3.03	2.95	2.89	2.83	2.78	2.74	2.70	2.66	2.63	2.60
	45	8.71	5.97	4.89	4.29	3.91	3.64	3.43	3.28	3.15	3.04	2.96	2.88	2.82	2.76	2.71	2.66	2.62	2.59	2.56	2.53
	50	8.63	5.90	4.83	4.23	3.85	3.58	3.38	3.22	3.09	2.99	2.90	2.82	2.76	2.70	2.65	2.61	2.57	2.53	2.50	2.47
	60	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.82	2.74	2.68	2.62	2.57	2.53	2.49	2.45	2.42	2.39
	70	8.40	5.72	4.66	4.08	3.70	3.43	3.23	3.08	2.95	2.85	2.76	2.68	2.62	2.56	2.51	2.47	2.43	2.39	2.36	2.33
	80	8.33	5.67	4.61	4.03	3.65	3.39	3.19	3.03	2.91	2.80	2.72	2.64	2.58	2.52	2.47	2.43	2.39	2.35	2.32	2.29
	90	8.28	5.62	4.57	3.99	3.62	3.35	3.15	3.00	2.87	2.77	2.68	2.61	2.54	2.49	2.44	2.39	2.35	2.32	2.28	2.25
	100	8.24	5.59	4.54	3.96	3.59	3.33	3.13	2.97	2.85	2.74	2.66	2.58	2.52	2.46	2.41	2.37	2.33	2.29	2.26	2.23
	120	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	2.62	2.54	2.48	2.42	2.37	2.33	2.29	2.25	2.22	2.19
	140	8.14	5.50	4.47	3.89	3.52	3.26	3.06	2.91	2.78	2.68	2.59	2.52	2.45	2.40	2.35	2.30	2.26	2.22	2.19	2.16
	160	8.10	5.48	4.44	3.87	3.50	3.24	3.04	2.88	2.76	2.66	2.57	2.50	2.43	2.38	2.33	2.28	2.24	2.20	2.17	2.14
	180	8.08	5.46	4.42	3.85	3.48	3.22	3.02	2.87	2.74	2.64	2.56	2.48	2.42	2.36	2.31	2.26	2.22	2.19	2.15	2.12
	200	8.06	5.44	4.41	3.84	3.47	3.21	3.01	2.86	2.73	2.63	2.54	2.47	2.40	2.35	2.30	2.25	2.21	2.18	2.14	2.11
	∞	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.75	2.62	2.52	2.43	2.36	2.30	2.24	2.19	2.14	2.10	2.07	2.03	2.00

TABLE 6(d) Values of the *F*-Distribution: $A = 0.005$ (Continued)

v_2	v_1	Numerator degrees of freedom																		
		22	24	26	28	30	35	40	45	50	60	70	80	90	100	120	140	160	180	200
1	24892	24940	24980	25014	25044	25103	25148	25183	25211	25253	25283	25306	25323	25337	25359	25374	25385	25394	25401	25464
2	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199
3	42.7	42.6	42.6	42.5	42.5	42.4	42.3	42.3	42.2	42.1	42.1	42.1	42.0	42.0	42.0	42.0	41.9	41.9	41.9	41.8
4	20.1	20.0	20.0	19.9	19.9	19.8	19.8	19.7	19.7	19.6	19.6	19.5	19.5	19.5	19.5	19.4	19.4	19.4	19.4	19.3
5	12.8	12.8	12.7	12.7	12.7	12.6	12.5	12.5	12.5	12.4	12.4	12.3	12.3	12.3	12.3	12.3	12.2	12.2	12.2	12.1
6	9.53	9.47	9.43	9.39	9.36	9.29	9.24	9.20	9.17	9.12	9.09	9.06	9.04	9.03	9.00	8.98	8.97	8.96	8.95	8.88
7	7.69	7.64	7.60	7.57	7.53	7.47	7.42	7.38	7.35	7.31	7.28	7.25	7.23	7.22	7.19	7.18	7.16	7.15	7.15	7.08
8	6.55	6.50	6.46	6.43	6.40	6.33	6.29	6.25	6.22	6.18	6.15	6.12	6.10	6.09	6.06	6.05	6.04	6.03	6.02	5.95
9	5.78	5.73	5.69	5.65	5.62	5.56	5.52	5.48	5.45	5.41	5.38	5.36	5.34	5.32	5.30	5.28	5.27	5.26	5.26	5.19
10	5.22	5.17	5.13	5.10	5.07	5.01	4.97	4.93	4.90	4.86	4.83	4.80	4.79	4.77	4.75	4.73	4.72	4.71	4.71	4.64
11	4.80	4.76	4.72	4.68	4.65	4.60	4.55	4.52	4.49	4.45	4.41	4.39	4.37	4.36	4.34	4.32	4.31	4.30	4.29	4.23
12	4.48	4.43	4.39	4.36	4.33	4.27	4.23	4.19	4.17	4.12	4.09	4.07	4.05	4.04	4.01	4.00	3.99	3.98	3.97	3.91
13	4.22	4.17	4.13	4.10	4.07	4.01	3.97	3.94	3.91	3.87	3.84	3.81	3.79	3.78	3.76	3.74	3.73	3.72	3.71	3.65
14	4.01	3.96	3.92	3.89	3.86	3.80	3.76	3.73	3.70	3.66	3.62	3.60	3.58	3.57	3.55	3.53	3.52	3.51	3.50	3.44
15	3.83	3.79	3.75	3.72	3.69	3.63	3.58	3.55	3.52	3.48	3.45	3.43	3.41	3.39	3.37	3.36	3.34	3.34	3.33	3.26
16	3.68	3.64	3.60	3.57	3.54	3.48	3.44	3.40	3.37	3.33	3.30	3.28	3.26	3.25	3.22	3.21	3.20	3.19	3.18	3.11
17	3.56	3.51	3.47	3.44	3.41	3.35	3.31	3.28	3.25	3.21	3.18	3.15	3.13	3.12	3.10	3.08	3.07	3.06	3.05	2.99
18	3.45	3.40	3.36	3.33	3.30	3.25	3.20	3.17	3.14	3.10	3.07	3.04	3.02	3.01	2.99	2.97	2.96	2.95	2.94	2.87
19	3.35	3.31	3.27	3.24	3.21	3.15	3.11	3.07	3.04	3.00	2.97	2.95	2.93	2.91	2.89	2.87	2.86	2.85	2.85	2.78
20	3.27	3.22	3.18	3.15	3.12	3.07	3.02	2.99	2.96	2.92	2.88	2.86	2.84	2.83	2.81	2.79	2.78	2.77	2.76	2.69
22	3.12	3.08	3.04	3.01	2.98	2.92	2.88	2.84	2.82	2.77	2.74	2.72	2.70	2.69	2.66	2.65	2.63	2.62	2.62	2.55
24	3.01	2.97	2.93	2.90	2.87	2.81	2.77	2.73	2.70	2.66	2.63	2.60	2.58	2.57	2.55	2.53	2.52	2.51	2.50	2.43
26	2.92	2.87	2.84	2.80	2.77	2.72	2.67	2.64	2.61	2.56	2.53	2.51	2.49	2.47	2.45	2.43	2.42	2.41	2.40	2.33
28	2.84	2.79	2.76	2.72	2.69	2.64	2.59	2.56	2.53	2.48	2.45	2.43	2.41	2.39	2.37	2.35	2.34	2.33	2.32	2.25
30	2.77	2.73	2.69	2.66	2.63	2.57	2.52	2.49	2.46	2.42	2.38	2.36	2.34	2.32	2.30	2.28	2.27	2.26	2.25	2.18
35	2.64	2.60	2.56	2.53	2.50	2.44	2.39	2.36	2.33	2.28	2.25	2.22	2.20	2.19	2.16	2.15	2.13	2.12	2.11	2.04
40	2.55	2.50	2.46	2.43	2.40	2.34	2.30	2.26	2.23	2.18	2.15	2.12	2.10	2.09	2.06	2.05	2.03	2.02	2.01	1.93
45	2.47	2.43	2.39	2.36	2.33	2.27	2.22	2.19	2.16	2.11	2.08	2.05	2.03	2.01	1.99	1.97	1.95	1.94	1.93	1.85
50	2.42	2.37	2.33	2.30	2.27	2.21	2.16	2.13	2.10	2.05	2.02	1.99	1.97	1.95	1.93	1.91	1.89	1.88	1.87	1.79
60	2.33	2.29	2.25	2.22	2.19	2.13	2.08	2.04	2.01	1.96	1.93	1.90	1.88	1.86	1.83	1.81	1.80	1.79	1.78	1.69
70	2.28	2.23	2.19	2.16	2.13	2.07	2.02	1.98	1.95	1.90	1.86	1.84	1.81	1.80	1.77	1.75	1.73	1.72	1.71	1.62
80	2.23	2.19	2.15	2.11	2.08	2.02	1.97	1.94	1.90	1.85	1.82	1.79	1.77	1.75	1.72	1.70	1.68	1.67	1.66	1.57
90	2.20	2.15	2.12	2.08	2.05	1.99	1.94	1.90	1.87	1.82	1.78	1.75	1.73	1.71	1.68	1.66	1.64	1.63	1.62	1.52
100	2.17	2.13	2.09	2.05	2.02	1.96	1.91	1.87	1.84	1.79	1.75	1.72	1.70	1.68	1.65	1.63	1.61	1.60	1.59	1.49
120	2.13	2.09	2.05	2.01	1.98	1.92	1.87	1.83	1.80	1.75	1.71	1.68	1.66	1.64	1.61	1.58	1.57	1.55	1.54	1.43
140	2.11	2.06	2.02	1.99	1.96	1.89	1.84	1.80	1.77	1.72	1.68	1.65	1.62	1.60	1.57	1.55	1.53	1.52	1.51	1.39
160	2.09	2.04	2.00	1.97	1.93	1.87	1.82	1.78	1.75	1.69	1.65	1.62	1.60	1.58	1.55	1.52	1.51	1.49	1.48	1.36
180	2.07	2.02	1.98	1.95	1.92	1.85	1.80	1.76	1.73	1.68	1.64	1.61	1.58	1.56	1.53	1.50	1.49	1.47	1.46	1.34
200	2.06	2.01	1.97	1.94	1.91	1.84	1.79	1.75	1.71	1.66	1.62	1.59	1.56	1.54	1.51	1.49	1.47	1.45	1.44	1.32
∞	1.95	1.90	1.86	1.82	1.79	1.72	1.67	1.63	1.59	1.54	1.49	1.46	1.43	1.40	1.37	1.34	1.31	1.30	1.28	1.00

TABLE 7 Random Numbers

Row	Column													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	13284	16834	74151	92027	24670	36665	00770	22878	02179	51602	07270	76517	97275	45960
2	21224	00370	30420	03883	96648	89428	41583	17564	27395	63904	41548	49197	82277	24120
3	99052	47887	81085	64933	66279	80432	65793	83287	34142	13241	30590	97760	35848	91983
4	00199	50993	98603	38452	87890	94624	69721	57484	67501	77638	44331	11257	71131	11059
5	60578	06483	28733	37867	07936	98710	98539	27186	31237	80612	44488	97819	70401	95419
6	91240	18312	17441	01929	18163	69201	31211	54288	39296	37318	65724	90401	79017	62077
7	97458	14229	12063	59611	32249	90466	33216	19358	02591	54263	88449	01912	07436	50813
8	35249	38646	34475	72417	60514	69257	12489	51924	86871	92446	36607	11458	30440	52639
9	38980	46600	11759	11900	46743	27860	77940	39298	97838	95145	32378	68038	89351	37005
10	10750	52745	38749	87365	58959	53731	89295	59062	39404	13198	59960	70408	29812	83126
11	36247	37850	73958	20673	37800	63835	71051	84724	52492	22342	78071	17456	96104	18327
12	70994	66986	99744	72438	01174	42159	11392	20724	54322	36923	70009	23233	65438	59685
13	99638	94702	11463	18148	81386	80431	90628	52506	02016	85151	88598	47821	00265	82525
14	72055	15774	43857	99805	10419	76939	25993	03544	21560	83471	43989	90770	22965	44247
15	24038	65541	85788	55835	38835	59399	13790	35112	01324	39520	76210	22467	83275	32286
16	94976	14631	35908	28221	39470	91548	12854	30166	09073	75887	36782	00268	97121	57676
17	35553	71628	70189	26436	63407	91178	90348	55359	80392	41012	36270	77786	89578	21059
18	35676	12797	51434	82976	42010	26344	92920	92155	58807	54644	58581	95331	78629	73344
19	74815	67523	72985	23183	02446	63594	98924	20633	58842	85961	07648	70164	34994	67662
20	45246	88048	65173	50989	91060	89894	36063	32819	68559	99221	49475	50558	34698	71800
21	76509	47069	86378	41797	11910	49672	88575	97966	32466	10083	54728	81972	58975	30761
22	19689	90332	04315	21358	97248	11188	39062	63312	52496	07349	79178	33692	57352	72862
23	42751	35318	97513	61537	54955	08159	00337	80778	27507	95478	21252	12746	37554	97775
24	11946	22681	45045	13964	57517	59419	58045	44067	58716	58840	45557	96345	33271	53464
25	96518	48688	20996	11090	48396	57177	83867	86464	14342	21545	46717	72364	86954	55580
26	35726	58643	76869	84622	39098	36083	72505	92265	23107	60278	05822	46760	44294	07672
27	39737	42750	48968	70536	84864	64952	38404	94317	65402	13589	01055	79044	19308	83623
28	97025	66492	56177	04049	80312	48028	26408	43591	75528	65341	49044	95495	81256	53214
29	62814	08075	09788	56350	76787	51591	54509	49295	85830	59860	30883	89660	96142	18354
30	25578	22950	15227	83291	41737	79599	96191	71845	86899	70694	24290	01551	80092	82118
31	68763	69576	88991	49662	46704	63362	56625	00481	73323	91427	15264	06969	57048	54149
32	17900	00813	64361	60725	88974	61005	99709	30666	26451	11528	44323	34778	60342	60388
33	71944	60227	63551	71109	05624	43836	58254	26160	32116	63403	35404	57146	10909	07346
34	54684	93691	85132	64399	29182	44324	14491	55226	78793	34107	30374	48429	51376	09559
35	25946	27623	11258	65204	52832	50880	22273	05554	99521	73791	85744	29276	70326	60251
36	01353	39318	44961	44972	91766	90262	56073	06606	51826	18893	83448	31915	97764	75091
37	99083	88191	27662	99113	57174	35571	99884	13951	71057	53961	61448	74909	07322	80960
38	52021	45406	37945	75234	24327	86978	22644	87779	235753	99926	63898	54886	18051	96314
39	78755	47744	43776	83098	03225	14281	83637	55984	13300	52212	58781	14905	46502	04472
40	25282	69106	59180	16257	22810	43609	12224	25643	89884	31149	85423	32581	34374	70873
41	11959	94202	02743	86847	79725	51811	12998	76844	05320	54236	53891	70226	38632	84776
42	11644	13792	98190	01424	30078	28197	55583	05197	47714	68440	22016	79204	06862	94451
43	06307	97912	68110	59812	95448	43244	31262	88880	13040	16458	43813	89416	42482	33939
44	76285	75714	89585	99296	52640	46518	55486	90754	88932	19937	57119	23251	55619	23679
45	55322	07589	39600	60866	63007	20007	66819	84164	61131	81429	60676	42807	78286	29015
46	78017	90928	90220	92503	83375	26986	74399	30885	88567	29169	72816	53357	15428	86932
47	44768	43342	20696	26331	43140	69744	82928	24988	94237	46138	77426	39039	55596	12655
48	25100	19336	14605	86603	51680	97678	24261	02464	86563	74812	60069	71674	15478	47642
49	83612	46623	62876	85197	07824	91392	58317	37726	84628	42221	10268	20692	15699	29167
50	41347	81666	82961	60413	71020	83658	02415	33322	66036	98712	46795	16308	28413	05417

Source: Abridged from W. H. Beyer, ed., *CRC Standard Mathematical Tables*, 26th ed. (Boca Raton: CRC Press, 1981). Reproduced by permission of the publisher. Copyright CRC Press, Inc., Boca Raton, Florida.

TABLE 8(a) Critical Values for the Durbin-Watson Statistic, $\alpha = 0.05$

<i>n</i>	<i>k = 1</i>		<i>k = 2</i>		<i>k = 3</i>		<i>k = 4</i>		<i>k = 5</i>	
	<i>D_L</i>	<i>D_u</i>								
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

Source: From J. Durbin and G. S. Watson, 'Testing for Serial Correlation in Least Squares Regression, 11/' *Biometrika* 30 (1951): 159–78. Reproduced by permission of the Biometrika Trustees.

TABLE 8(b) Critical Values for the Durbin-Watson Statistic, $\alpha = 0.01$

<i>n</i>	<i>k = 1</i>		<i>k = 2</i>		<i>k = 3</i>		<i>k = 4</i>		<i>k = 5</i>	
	<i>D_L</i>	<i>D_u</i>								
15	0.81	1.07	0.70	1.25	0.59	1.46	0.49	1.70	0.39	1.96
16	0.84	1.09	0.74	1.25	0.63	1.44	0.53	1.66	0.44	1.90
17	0.87	1.10	0.77	1.25	0.67	1.43	0.57	1.63	0.48	1.85
18	0.90	1.12	0.80	1.26	0.71	1.42	0.61	1.60	0.52	1.80
19	0.93	1.13	0.83	1.26	0.74	1.41	0.65	1.58	0.56	1.77
20	0.95	1.15	0.86	1.27	0.77	1.41	0.68	1.57	0.60	1.74
21	0.97	1.16	0.89	1.27	0.80	1.41	0.72	1.55	0.63	1.71
22	1.00	1.17	0.91	1.28	0.83	1.40	0.75	1.54	0.66	1.69
23	1.02	1.19	0.94	1.29	0.86	1.40	0.77	1.53	0.70	1.67
24	1.04	1.20	0.96	1.30	0.88	1.41	0.80	1.53	0.72	1.66
25	1.05	1.21	0.98	1.30	0.90	1.41	0.83	1.52	0.75	1.65
26	1.07	1.22	1.00	1.31	0.93	1.41	0.85	1.52	0.78	1.64
27	1.09	1.23	1.02	1.32	0.95	1.41	0.88	1.51	0.81	1.63
28	1.10	1.24	1.04	1.32	0.97	1.41	0.90	1.51	0.83	1.62
29	1.12	1.25	1.05	1.33	0.99	1.42	0.92	1.51	0.85	1.61
30	1.13	1.26	1.07	1.34	1.01	1.42	0.94	1.51	0.88	1.61
31	1.15	1.27	1.08	1.34	1.02	1.42	0.96	1.51	0.90	1.60
32	1.16	1.28	1.10	1.35	1.04	1.43	0.98	1.51	0.92	1.60
33	1.17	1.29	1.11	1.36	1.05	1.43	1.00	1.51	0.94	1.59
34	1.18	1.30	1.13	1.36	1.07	1.43	1.01	1.51	0.95	1.59
35	1.19	1.31	1.14	1.37	1.08	1.44	1.03	1.51	0.97	1.59
36	1.21	1.32	1.15	1.38	1.10	1.44	1.04	1.51	0.99	1.59
37	1.22	1.32	1.16	1.38	1.11	1.45	1.06	1.51	1.00	1.59
38	1.23	1.33	1.18	1.39	1.12	1.45	1.07	1.52	1.02	1.58
39	1.24	1.34	1.19	1.39	1.14	1.45	1.09	1.52	1.03	1.58
40	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
45	1.29	1.38	1.24	1.42	1.20	1.48	1.16	1.53	1.11	1.58
50	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
55	1.36	1.43	1.32	1.47	1.28	1.51	1.25	1.55	1.21	1.59
60	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
65	1.41	1.47	1.38	1.50	1.35	1.53	1.31	1.57	1.28	1.61
70	1.43	1.49	1.40	1.52	1.37	1.55	1.34	1.58	1.31	1.61
75	1.45	1.50	1.42	1.53	1.39	1.56	1.37	1.59	1.34	1.62
80	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
85	1.48	1.53	1.46	1.55	1.43	1.58	1.41	1.60	1.39	1.63
90	1.50	1.54	1.47	1.56	1.45	1.59	1.43	1.61	1.41	1.64
95	1.51	1.55	1.49	1.57	1.47	1.60	1.45	1.62	1.42	1.64
100	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65

Source: From J. Durbin and G. S. Watson, 'Testing for Serial Correlation in Least Squares Regression, 11/' *Biometrika* 30 (1951): 159–78. Reproduced by permission of the Biometrika Trustees.

Glossary

A

acceptance region

Values of the test statistic for which the null hypothesis H_0 is not rejected.

addition rule for mutually exclusive events

The rule that allows us to calculate the probability of the union of two mutually exclusive events.

addition rule

The rule that allows us to calculate the probability of the union of two events.

alternative (or research) hypothesis

The proposition, denoted H_A , that will be accepted if the null hypothesis H_0 is rejected.

autoregressive model

A model based on the belief that there is correlation between consecutive residuals.

average of relative price index

A simple average of the n price relatives, multiplied by 100.

B

bar chart

A chart in which vertical bars represent data in different categories.

bell-shaped

Symmetric in the shape of a bell (or mound-shaped).

Bernoulli trial

A random experiment that has only two possible outcomes.

bimodal histogram

A histogram with two modes.

binomial experiment

An experiment consisting of a number of repeated Bernoulli trials.

binomial probability distribution

The probability distribution of the binomial random variable.

binomial random variable

The number of successes in the n trials of a binomial experiment.

bivariate distribution

Joint probability distribution of two variables.

bivariate methods

Techniques involving or dependent on two variables.

box plot

Pictorial display showing the five summary statistics, the three quartiles and the largest and smallest values of the data.

C

central limit theorem

The sampling distribution of the sample mean \bar{X} will be approximately normal when $n > 30$.

centred moving average

A technique for centring the moving averages when the number of time periods used to calculate the averages is an even number.

Chebyshev's theorem

The proportion of observations that lie within k standard deviations of the mean is at least $(1 - 1/k^2)$.

chi-squared distribution

A continuous nonsymmetric distribution used in statistical inference.

chi-squared goodness-of-fit test

This test is commonly used to test whether a data set comes from a normal population under a multinomial framework.

chi-squared random variable

A random variable that is chi-squared distributed.

chi-squared test of a contingency table

Test used to determine whether there is enough evidence to infer that two nominal variables are related, and to infer that differences exist between two or more populations of nominal variables.

classes

Non-overlapping intervals of numerical data.

classical approach

Assigning equal probabilities to all the elementary events or outcomes.

class relative frequency

Percentage of data in each class.

cluster sample

A simple random sample of clusters, or groups, of elements.

coefficient of correlation (Pearson)

A measurement of the strength and direction of a linear relationship between two numerical variables.

coefficient of correlation (Pearson)

A measurement of the strength and direction of a linear relationship between two numerical variables.

coefficient of determination adjusted for degrees of freedom (adjusted R^2)

A goodness-of-fit measure of the relationship between the dependent and independent variables, adjusted to take into account the number of independent variables in a multiple regression model.

coefficient of determination (R^2)

The proportion of the variation in the dependent variable that is explained by the variation in the independent variable(s).

coefficient of determination

The proportion of the amount of variation in the dependent variable that is explained by the independent variable.

coefficient of variation

Standard deviation divided by the mean.

complement of an event

The set of all outcomes not in the event but in the sample space.

complement rule

The probability of an event equals one minus the probability of its complement.

conditional probability

The probability an event will occur, given that another has occurred or will occur.

confidence interval estimator

An interval estimator in which we have a certain degree of confidence that it contains the value of the parameter.

consistent estimator

Approaches the value of the parameter it is estimating as the sample size increases.

Consumer Price Index (CPI)

An economic indicator that measures the changes in the total price of a basket of goods and services.

contingency table

(or cross-classification table) A table with the expected values calculated contingent on the assumption that the null hypothesis is true.

continuity correction factor

A correction factor that allows for the approximation of a discrete random variable by a continuous random variable.

continuous random variable

A random variable that can assume an uncountable number of values in an interval.

continuous random variable

A random variable that can assume an uncountable number of values in an interval.

covariance

A measure of how two variables are linearly related.

critical region

Another name for the rejection region.

critical value

Value that separates the acceptance and rejection regions.

cross-classification table

A first step in graphing the relationship between two nominal variables.

cross-sectional data

Data measured across a population (or a sample) at one point in time.

cumulative probability

The probability that a random variable X is less than or equal to x ,

$$P(X \leq x)$$

cumulative relative frequency

Percentage of observations less than or equal to the upper limit of a class.

cycle

A wave-like trend in time series data resulting in a cyclical effect.

D**data**

Observations of the variables of interest.

decision rule

A statement specifying the condition for the rejection of H_0 .

dependent events

Events in which the occurrence of one does affect the probability the other will occur.

descriptive statistics

Methods of organising, summarising and presenting data in ways that are useful, attractive and informative to the reader.

deseasonalising

A method of removing seasonal effects from a time series, resulting in a seasonally adjusted series.

deterministic model

An equation in which the value of the dependent variable is completely determined by the values of the independent variable(s).

deviation

Difference between an observation and the mean of the set of data it belongs to.

discrete random variable

A random variable that can assume only a countable number of values (finite or infinite).

Durbin–Watson test

A test for autocorrelation.

E**empirical rule**

When the distribution is bell shaped, the percentage of observations that fall within 1, 2 and 3 standard deviations (SDs) from the mean are 68%, 95% and 99.7% respectively.

equal variances t-test for $\mu_1 - \mu_2$

This test assesses the significance of the difference between two populations when the variances are unknown but are expected to be equal.

equal variances t-test statistic for $\mu_1 - \mu_2$

The statistic used to test for the equality of means when both variances of two populations are unknown and presumed, or tested, to be equal.

error of estimation

The absolute difference between the statistic and the parameter.

estimate

An approximate value of a parameter based on a sample statistic.

event

A set of one or more simple events or outcomes.

exhaustive

Covers all possible outcomes.

expected frequency

The frequency of each outcome we expect to observe if the null hypothesis is true.

expected value

The sum of all possible values a random variable can take times the corresponding probabilities.

exponential distribution

A continuous distribution with probability density function $f(x) = \lambda e^{-\lambda x}, x \geq 0$.

exponential random variable

A random variable that is exponentially distributed.

exponential smoothing

A technique for smoothing a time series in a way that includes all data prior to time t in the smoothed time series in period t .

F**failure**

The non-success outcome of a Bernoulli trial.

F-distribution

A continuous distribution used in statistical inference.

Fisher price index

A geometric mean of the Laspeyres price index and the Paasche price index.

frequency distribution

Method of presenting data and their counts in each category or class.

G**geometric mean**

The n th root of the product of all observations.

graphical deception

Presentation of a distorted impression using graphs.

graphical excellence

Application of graphical techniques that are informative and concise and that impart information clearly.

H**heteroscedasticity**

The condition under which the variance of the error variables is not constant.

histogram

Graphical presentation of a frequency distribution of numerical data.

homoscedasticity

The condition under which the variance of the error variables is constant.

hypothesis

A proposition or conjecture that the statistician will test by a means called hypothesis testing.

I**independent events**

Events in which the occurrence of one does not affect the probability the other will occur.

index numbers

Measure the changes over time of particular time-series data.

inferential statistics

Methods used to draw conclusions about a population based on information provided by a sample of the population.

interquartile range (IQR)

The difference between the first and third quartiles.

intersection

A joint event consisting of common outcomes from two events.

interval estimator

Estimates the value of an unknown parameter which reflects the variability in the sample, sample size and the level of confidence using an interval.

J**joint probability**

The likelihood of occurrence of a joint event.

K**Known variances z-test of $\mu_1 - \mu_2$**

This test assesses the significance of the difference between two populations when the variances are known.

known variances z-test statistic of $\mu_1 - \mu_2$

The statistic used to test for the equality of means when both population variances are known.

L**Laspeyres price index**

Ratio (in percentages) of current price expenditure to the base price expenditure of purchasing a base period basket of goods.

least squares method

A method of deriving an estimated linear equation (straight line) which best fits the data.

linear relationship

One in which two variables move proportionately.

line chart

A chart showing the movement of a variable over time.

lower confidence limit (LCL)

The lower bound of the confidence interval.

lower fence

$Q_1 - 1.5(IQR)$

M**marginal probability**

The probability of an event irrespective of any other event.

matched pairs experiment

One in which each observation from one sample can be matched with an observation in another sample.

mean absolute deviation (MAD)

A measure of variation, which is the average of absolute deviations.

mean (arithmetic mean)

The sum of a set of observations divided by the number of observations.

mean of the population of differences

The mean of the paired differences in a matched pairs experiment.

median

The middle value of a set of observations when they are arranged in order of magnitude.

modal class

The class with the largest number of observations.

mode

The most frequently occurring value in a set of data.

moving average

The arithmetic average of a point in a time series with nearby points.

multicollinearity

Condition in which the independent variables in a regression model are correlated.

multimodal histogram

A histogram with two or more peaks.

multinomial experiment

An extension of the binomial experiment, in which there are two or more possible outcomes per trial.

multiplication rule for independent events

The rule that allows us to calculate the probability of the intersection of two independent events.

multiplication rule

The rule that allows us to calculate the probability of the intersection of two events.

mutually exclusive

Outcomes that cannot occur at the same time.

N**negatively skewed histogram**

A histogram with a long tail to the left.

negative relationship

A relationship in which the variables move in opposite directions to each other.

nominal data

Observations are categorical or qualitative.

normal distribution

The most important continuous distribution. The curve is bell-shaped, and describes many phenomena that occur both in nature and in business.

normal random variable

A random variable that is normally distributed.

null hypothesis

The proposition about which a decision is to be made in testing a hypothesis, denoted H_0 .

numerical data

Observations are real numbers.

O**observed frequencies**

The number of observations of each outcome in the experiment.

ogive

Line graph of cumulative relative frequency.

one-tail test

A test with the rejection region in only one tail of the distribution.

operating characteristic (OC) curve

A plot of the probability of making Type II error against the values of the parameter.

ordinal data

Ordered nominal data.

ordinal scale

A scale applied to ranked data.

origin

The initial node of the experiment.

outlier

An observation more than $Q_3 + 1.5(\text{IQR})$ or less than $Q_1 - 1.5(\text{IQR})$.

P**Paasche price index**

Ratio (in percentages) of current price expenditure to the base price expenditure of purchasing a current period basket of goods.

parameter

A descriptive measure of a population.

percentage of trend

The amount of trend produced by a given effect.

percentile

The p th percentile is the value for which $p\%$ of observations are less than that value and $(100 - p)\%$ are greater than that value.

pie chart

A circle subdivided into sectors representing the share of data in different categories.

point estimator

Estimates the value of an unknown parameter using a single value.

Poisson experiment

An experiment with rare outcomes.

Poisson probability distribution

The probability distribution of the Poisson random variable.

Poisson random variable

The number of successes that occur in a period of time or an interval of space in a Poisson experiment.

pooled proportion estimate

The weighted average of two sample proportions.

pooled variance estimate

The weighted average of two sample variances.

population mean

The average of all its possible values, denoted by μ .

population

The set of all items of interest.

positively skewed histogram

A histogram with a long tail to the right.

positive relationship

A relationship in which the variables move in the same direction.

power

The probability of correctly rejecting a false null hypothesis.

prediction interval

The confidence interval for a predicted value of the dependent variable for a given value of the independent variable.

probabilistic model

A model that contains a random term.

probability density function (pdf)

A function $f(x)$ such that (1) $f(x)$ is non-negative, (2) the total area under $f(x)$ is 1, (3) the area under $f(x)$ between the lines $x = a$ and $x = b$ gives the probability that the value of X is between a and b , where X is a continuous random variable.

probability distribution

A table or graph that assigns probability values to all possible values or range of values that a random variable can assume.

probability of an event

Sum of the probabilities of the simple events in the event.

probability

The likelihood of the occurrence of an outcome or a collection of outcomes.

probability tree

A depiction of events as branches of a tree.

p-value

Smallest value of α that would lead to the rejection of the null hypothesis.

Q**quartiles**

The 25th (Q_1), 50th (Q_2 , or median), and 75th (Q_3) percentiles.

R**random experiment**

A process that results in one of a number of possible different outcomes.

random variable

A variable whose values are determined by outcomes of a random experiment.

random variation

Irregular changes in a time series.

range

The difference between largest and smallest observations.

regression analysis

A technique that estimates the relationship between variables and aids forecasting.

rejection region

The range of values of the test statistic that would cause us to reject the null hypothesis.

relative frequency approach

Expresses the probability of an outcome as the relative frequency of its occurrence based on past experience.

relative frequency distribution

Frequency distribution giving the percentage each category or class represents of the total.

relative frequency histogram

A histogram with the vertical axis representing relative frequencies.

residual

The difference between the predicted value of the dependent variable and its actual value.

response surface

The graphical depiction of the regression equation when there is more than one independent variable; when there are two independent variables, the response surface is a plane.

rule of five

A rule specifying that each expected frequency should be at least 5 in order for the chi-square test to be sufficiently well approximated.

S**sample**

A set of data drawn from the studied population.

sampled population

The actual population from which the sample has been drawn.

sample mean

The arithmetic mean of sample data.

sample size required to estimate the population mean μ

The sample size required to estimate μ given the confidence level of the interval estimator and the size of the error bound.

sample space

The set of all possible simple events or outcomes.

sampling distribution of the sample mean

A relative frequency distribution of various values of the sample mean obtained from a number of different samples selected from the same population.

sampling distribution of the sample proportion

A relative frequency distribution of various values of the sample proportion using a number of samples from the same population.

scatter diagram

A plot of points of one variable against another which illustrates the relationship between them.

seasonal indexes

Measures of the seasonal effects in time series data.

seasonal variations

Short-term seasonal cycles in time series data.

significance level

Probability of rejecting the null hypothesis when the null hypothesis is true.

simple aggregate price index

The ratio (in percentages) of the sum of the prices of n commodities in the current period 1 divided by the sum of the prices of the same n commodities in the base period 0.

simple linear regression model

Also called the first-order linear model, this is a linear regression equation with only one independent variable.

simple price index

The ratio (in percentages) of the price of a commodity in current period 1 divided by its value in some base period 0.

simple random sample

One in which each element of the population has an equal chance of appearing.

skewed to the left

Negatively skewed, has a long tail extending off to the left.

skewed to the right

Positively skewed, has a long tail extending off to the right.

skewness

The degree to which a graph differs from a symmetric graph.

standard deviation

Square root of the variance.

standard error of estimate

An estimate of the standard deviation of the error variable, which is the square root of the sum of squares error (SSE) divided by the degrees of freedom.

standard error of the sample mean

The standard deviation of the sampling distribution of the sample mean, σ / \sqrt{n} .

standard error of the sample proportion

The standard deviation of the sampling distribution of the sample proportion, $\sqrt{pq / n}$.

standard normal distribution**(z-distribution)**

Normal distribution with a mean of 0 and a standard deviation of 1.

standard normal random variable

Labelled Z, a normal random variable with a mean of 0 and a standard deviation of 1.

statistic

A descriptive measure of a sample.

statistically significant

There is enough evidence to reject the null hypothesis.

stem-and-leaf display

Display of data in which the stem consists of the digits to the left of a given point and the leaf the digits to the right.

stratified random sample

One in which the population is separated into mutually exclusive layers, or strata, from which simple random samples are drawn.

Student t distribution, or t distribution

A continuous distribution used in statistical inference when the population variance is not known.

subjective approach

Assigns probability to an outcome based on personal judgement.

success

An arbitrary label given to one of the outcomes of a Bernoulli trial.

sum of squares error (SSE)

The sum of squared residuals in a regression model.

symmetric histograms

Histograms in which, if lines were drawn down the middle, the two halves would be mirror images.

T**target population**

The population about which we want to draw inferences.

test statistic

The statistic used to decide whether or not to reject the null hypothesis.

time series

A variable measured over time in sequential order.

time series

A variable measured over time in sequential order.

time-series chart

Line chart in which the categories are points in time.

time-series data

Data measured on the same variable at different points of time.

trend

A long-term pattern in a time series.

t-statistic for μ

Standardised value of the sample mean when σ is unknown and replaced by s .

two-tail test

A test with the rejection region in both tails of the distribution, typically split evenly.

Type I error

The act of rejecting a null hypothesis when it is true.

Type II error

The act of not rejecting a null hypothesis when it is false.

U**unbiased estimator**

Has an expected value that equals the parameter being estimated.

unequal variances t-test of $\mu_1 - \mu_2$

This test assesses the significance of the difference between two population means when the variances are unknown but are not expected to be equal.

unequal variances t-test statistic of $\mu_1 - \mu_2$

The statistic used to test for the equality of two population means when the variances of the two populations are unknown and presumed, or tested, to be unequal.

uniform distribution

A continuous distribution with probability density function $f(x) = 1/(b - a)$ for values of x in the interval a to b .

unimodal histogram

A histogram with only one mode.

union

An event consisting of all outcomes from two events.

upper confidence limit (UCL)

The upper bound of the confidence interval.

upper fence

$Q_3 + 1.5(\text{IQR})$

V**variable**

Any characteristic of a population or sample.

variance

A measure of variability of a numerical data set.

variance of a population

Sum of the squared deviations (from the population mean) of all of the items of data in the population divided by the number of items.

variance of a sample

Sum of the squared deviations (from the sample mean) of all the items of data in a sample, divided by one less than the number of items.

W**weighted aggregate price index**

The weighted sum of the n price relatives, multiplied by 100, where the weights are non-negative and sum to 1.

weighted mean

The sum of a set of observations multiplied by their corresponding weights.

Index

A

acceptance region 470
addition rule
for mutually exclusive events 235
probability 235–6
alternative hypothesis 468, 469, 471, 472, 474, 475
nominal data 469
numerical data 468
analysis of variance (ANOVA), and
t-tests (regression models) 703
ANOVA table, simple regression model 653
applications
in finance 296–9, 654
in marketing 417–20
arithmetic mean 136–8, 143–4, 145
asset allocation 296–9
Australian Bureau of Statistics 1, 18–19, 800, 812, 821
Australian Consumer Price Index (CPI) 800, 812–23
composition of CPI basket 813
construction 813–15
definition 800
using to deflate GDP 817–19
using to deflate wages 816–17
what is it? 812–13
autocorrelation 671–2, 714, 733
first-order 726–8
see also Durbin–Watson test
autoregressive model 792–4
average compounding rate of return 146–9
average of relative price index 806–8

B

bar charts 46–54
component 60–1
for time-series data 108–9
Bayes, Thomas 244
Bayes' law 244–6
bell-shaped histogram 96
bell-shaped normal distribution 316
Bernoulli trial 275
bias 40
biased estimator 376–7
big data 24–5
bimodal distribution 140
bimodal histogram 96
binomial distribution 275–83
mean and variance 281–3
normal approximation to the 344–7
binomial experiment 275–6
binomial probability
approximate, using a normal
distribution 344–7
tables of 843–6
using the binomial table to find 280
using the computer to find 281
binomial probability distribution 276–9
binomial random variable 275
mean and variance 281–3
binomial table 280
bivariate distributions 290–4

describing 291–3
marginal probabilities 290–1
sum of two variables 293–4
bivariate methods 68, 111
box plots (box-and-whisker plots) 172–5
bubble chart 117–18
business data analytics 10
business intelligence 10

C

categorical data 21, 23
cause-and-effect relationship (linear regression) 654
census 1, 18–19, 37–8
central limit theorem 359
centred moving averages 757–8
Chebyshev's theorem 162–3
chi-squared distribution 622–3
critical values 852
chi-squared goodness-of-fit test 583–91, 614
chi-squared goodness-of-fit test statistic 585
chi-squared random variable 623
chi-squared test of a contingency table 583, 593–605
data formats 600–2
degrees of freedom 600
rule of five 602–3
test statistic 594–600
chi-squared test for normality 608–13
choosing class intervals 610–12
interpreting the results 612–13
test for a normal distribution 608–10
using the computer 611–12
chi-squared tests 583–616
of a contingency table 583, 593–605
goodness-of-fit test 583–91, 614
multinomial experiment 583–91
for normality 608–13
chi-squared values
determining manually 622–3
determining using the computer 623
class intervals 91, 610–12
class relative frequency 92
classes 90
determining number of 90–1
classical approach (probabilities) 214
cluster sampling 36
coefficient of correlation 179, 179–85, 292, 664–7
Pearson 179–85, 664
Spearman rank 667
testing 664–7
coefficient of determination 179, 182, 191–2, 650–3, 693
adjusted for degrees of freedom 693
coefficient of variation 163–5
coefficients, estimating (multiple regression) 691–5, 698–701
coefficients, estimating (simple linear regression) 628–35
fitness of the regression line 631–5
collinearity 714

complement of an event 216–17
complement rule (probability) 234–5
component bar charts 60–1
Compustat 27
conditional probability 224–30
using Bayes' law 244–6
confidence interval estimator
for difference between two population means when population variances are known 433–4
for difference between two population means when population variances are unknown and equal 442–4
for difference between two population means when population variances are unknown and unequal 439–42
for difference between two population proportions 455–6
of expected value 660
and hypothesis testing 488–9
interpreting 383–5
mean of the population difference 450
for population mean when population variance is known 379–80
for population mean when population variance is unknown 394–5
of population proportion 404
and prediction intervals 662
slope and intercept (simple linear regression) 646–7
to determine required sample size 410–11
total number of successes in a large finite population 398, 408
and the width of the interval 385–6
confidence level 6
consistent estimator 377
Consumer Price Index (CPI) *see* Australian Consumer Price Index (CPI)
contingency table 68
chi-square test of 583, 583–605
continuity correction factor 346
omitting 346–7
continuous probability distributions 309–39
exponential distribution 336–9
normal distribution 316–32
probability density functions 310–13
summary of formulas 341
uniform distribution 313–15
continuous random variables 263, 310
correlation, interpreting 186
correlation coefficient *see* coefficient of correlation
counting rule 277
covariance 179–81, 291–2
CPI basket 812, 813
CPI population group 812
criminal trials 471
critical region 470
critical values 470
cross-classification table 68, 593
cross-sectional data 24, 106

cross-tabulation table 68
 cubic model for long-term trend 763
 cumulative probability 279
 cumulative relative frequency 100–2
 cumulative standardised normal probabilities 849–50
 cycle 751
 cyclical effect, measuring 768–71
 cyclical variation 751–2

D

data 20
 acquisition, errors 40
 experimental 27, 642
 general guidelines on the exploration of 193–4
 grouped 206–8
 hierarchy of 22–3
 missing 399–400, 428
 observational 27, 642
 published 27
 recoding 428
 data collection methods 26–9
 data dashboards 10–11
 data formats 72, 544–5, 600
 data mining techniques 11
 data queries 1
 data types 20–5
 big 24–5
 cross-sectional 24
 nominal 21, 22, 23
 numerical 20, 21, 23, 86
 ordinal 21, 22, 23
 panel 24
 and problem objectives 25, 351–2
 time-series 24, 106–8
 decision rule 470
 degrees of freedom 399
 for contingency table 600
 dependent events 225
 dependent samples (matched pairs) 449–52, 551–9
 dependent variable 352, 625
 descriptive analytics 10–11
 descriptive statistics 3–4, 45
 deseasonalising a time series 777–8
 determination, coefficient of see coefficient of determination
 deterministic models 626
 deviation 154
 mean absolute (MAD) 155, 781
 mean squared 156
 see also standard deviation
 direct observation 27
 direction (scatter diagrams) 115
 discrete probability distributions 263–6
 discrete random variables 262
 drug testing 471–2
 dummy variables, forecasting seasonal time series with 786, 790–1
 Durbin–Watson statistic
 $\alpha = 0.01$ 863
 $\alpha = 0.05$ 862
 Durbin–Watson test 714, 726–33, 792

E

empirical rule 161–2
 equal-variances t-interval estimator of difference between population means 444
 equal-variances t-test for the difference between population means 543
 equal-variances t-test statistic 532
 error
 in data acquisition 40
 of estimation 411
 non-response 40
 non-sampling 40
 sampling 31, 39
 Type I 470, 471, 472, 703
 Type II 470, 471, 472, 510–16, 521
 error variable
 non-independence 671–2
 required conditions 641–2
 estimate 375
 estimation 20
 concepts 375–8
 error of 411
 expected value 659–60
 estimation (single population) 374–423
 concepts 375–8
 determining the required sample size 410–15
 population mean using sample median 386–7
 population mean when population variance is known 378–87
 population mean when population variance is unknown 391–400
 population proportion 403–8
 three-stage solution process 380–2
 see also confidence interval estimator
 estimation (two populations) 430–58
 difference between two population means when the population variances are known: independent samples 431–7
 difference between two population means when the population variances are unknown: independent samples 439–45
 difference between two population means with matched pairs experiments: dependent samples 449–52
 difference between two population proportions 453–8
 violations of the required conditions 445
 estimators 375
 biased 376–7
 confidence interval 379–80, 383–6, 394–5
 consistent 377
 point and interval 375
 quality of 376–8
 relative efficiency 377–8
 unbiased 376, 377–8
 event, probability of an 214–15
 Excel instructions 11, 15–17
 Data Analysis/Analysis ToolPak 12, 16–17
Data Analysis Plus 12, 13
 formula bar and insertion function f_x 17
 importing data files 16
 inputting data 16
 manipulating data 581
 missing data 428
 online resources 13
 opening Excel 15
 performing statistical procedures 16
 recoding data 428
 saving workbooks 17
 spreadsheets 12–13
 statistical 12
 testing population mean when the variance is known 529
 workbook and worksheet 15
 exhaustive outcomes 213
 expected frequency 584
 for a contingency table 594
 expected value 269–72, 659–62
 laws of 272–3, 294
 experimental data 27, 642
 experiments 27
 matched pairs (dependent samples) 449–52
 exponential distribution 336–9
 exponential probabilities
 calculating 336–8
 using a computer to find 338
 exponential probability density function 336
 exponential random variable 336, 337
 exponential smoothing 758–61
 time series forecasting with 783–5

F

F-distribution 743–5
 values: $A = 0.005$ 859–60
 values: $A = 0.01$ 857–8
 values: $A = 0.025$ 855–6
 values: $A = 0.05$ 853–4
F-probabilities 745
F-tests, and *t*-test in simple linear regression model 703
F-values
 determining manually 743–5
 determining using the computer 745
 failure (Bernoulli trial) 275
 finance, applications in 296–9, 654
 finite population
 large, estimating the total number of successes 398, 408
 sampling from a 364
 finite population correction factor 364, 398, 408
 first-order autocorrelation 726–8
 first-order linear model 627
 Fisher price index 811–12
 flowchart, of graphical and numerical techniques 210
 forecasting 749, 780–2
 autoregressive model 792–4
 with seasonal indexes 786–9
 seasonal time series with dummy variables 786, 790–1
 time series with exponential smoothing 783–5
 time series with regression 785–94
 time series with trend and seasonality 786
 formulas, summary of 195, 252, 303–4, 341, 371, 423, 461, 524, 574, 616, 678, 737, 796, 824
 frequency distribution 46, 90
 and classes 90
 number of classes 90–1
 frequency polygons 100

G

GDP, deflating using CPI 817–19
 geometric mean rate of return 147–9
 glossary 864–8
 goodness-of-fit test (chi-squared test) 583–91
 graphical deception 124–9
 graphical descriptive techniques (nominal data) 44, 45–61, 208
 bar and pie charts 46–54
 component bar charts 60–1
 flowchart 210
 relationship between two nominal variables 68–72
 selecting a chart 54–60
 which is best? 54–60
 graphical descriptive techniques (numerical data) 85–130, 208
 flowchart 210
 frequency distribution and classes 90–1
 frequency polygons 100
 graphical deception 124–9
 graphical excellence 123–4
 graphing the relationship between two numerical variables 111–17
 graphing the relationship between three numerical variables 117–18
 heat map 119–20
 histograms 91–4, 95–100
 numerical variables, relationship between 111–18
 ogives 100–2
 oral presentations 130
 relative frequency histograms 92–3
 scatter diagrams 111–17
 shapes of histograms 95–100
 stem-and-leaf displays 94–5
 time-series data 106–9
 written reports 129–30
 graphical descriptive techniques (ordinal data) 61–2, 210
 graphical excellence 123–4
 graphical techniques 208
 group mean 206
 group variance 206
 grouped data, approximate descriptive measures for 206–7

H

heat map 119–20
 heteroscedasticity 670–1, 718
 histograms 91–4, 95–100
 bell-shaped 96
 negatively (left) skewed 95
 number of modal classes 96
 positively (right) skewed 95
 relative frequency 92–3
 shapes of 95–100
 skewness 95–6
 symmetric 95, 96
 with unequal class widths 93
 homoscedasticity 670
 hypothesis 467
 hypothesis testing (single population) 467–521
 applications in other disciplines 471–2
 calculating the probability of a Type II error 510–16
 components of the tests 468–70

concepts of 467–75

and confidence interval estimators 488–9
 discussions and conclusions 481–2
 judging the test 512–14
 level of significance 470–1
 one-tail test 472–3, 485–8
 operating characteristic curve and the power curve 515–16
p-value of a test 491–502
 population proportion 517–21
 power of a test 515
 six-step process 473–5
 and statistical concepts 489
 testing the population mean when the population variance is known 476–89
 testing the population mean when the population variance is unknown 504–8
 two-tail test 472
 hypothesis testing (two populations) 531–70
 checking the required condition 544
 data formats 544–5
 six-step process 531–4
 and statistical concepts 545, 559
 testing the difference between two population means: independent samples 531–45
 testing the difference between two population means: matched pairs experiment 551–9
 testing the difference between two population proportions 562–70
 violation of the required condition 544

I

independent events 225–6, 230, 237
 independent samples
 difference between two population means when the variances are known 431–7
 difference between two population means when the variances are unknown 439–45
 experiment 552–3
 testing the difference between two population means 531–45
 vs matched pairs – which experimental design is better? 557–8
 independent variables 352, 625
 index numbers 800–24
 Australian Consumer Price Index (CPI) 800, 812–21
 changing the base period of an index number series 821–3
 constructing unweighted 801–8
 constructing weighted 808–12
 definition 800
 summary of formulas 824
 see also price index
 inferential statistics 4–5, 45, 351–7
 influential observations (regression analysis) 673–4
 intercept coefficient estimator 646
 intercorrelation 714
 interquartile range (IQR) 172
 intersection 216
 interval data 20, 23
 interval estimators 375, 461–2, 646
 see also confidence interval estimator
 interviews 28

J

joint probabilities 224–30, 251

K

known variances z-test of the differences in population means 534
 known variances z-test statistic 532

L

large finite population, estimating the total number of successes 398, 408
 Laspeyres price index 808–10, 811
 laws of expected value 272–3, 294
 laws of variance 272–3, 294
 least squares estimators 629
 least squares method 114, 179, 186–7, 629
 least squares regression line 629–31
 line charts 106–8
 linear model for long-term trend 763
 linear relationship 112, 114
 estimating the 179, 186–91
 interpreting 115
 linear trend model 763
 linearity 114
 location of a percentile 170–1
 log transformation 718
 long-term trend 750–1
 lower confidence limit (LCL) 380
 lower fence 173

M

managers, use of statistics 9–11
 marginal probabilities 224–30, 290–1
 market model 654
 market segmentation 417–20
 marketing, applications in 417–20
 matched pairs experiments
 dependent samples 449–52
 estimating the mean difference 450–2
 recognising 451
 testing the difference between two population means 551–9
 violation of the required condition 452, 559
 vs independent samples – which experimental design is better? 557–8
 mean 136–8, 143–4
 of binomial distribution 281–3
 of Poisson random variables 288
 of the population of differences (matched pairs experiments) 449–51
 of portfolio of *k* shares 299
 of portfolio of two shares 297
 of the sample proportion 367
 of the sampling distribution 357
 see also geometric mean; weighted mean
 mean absolute deviation (MAD) 155, 780, 781
 mean square error (MSE) 377
 mean squared deviation 156
 measures of association 136, 179–92
 coefficient of correlation 179, 181–5
 coefficient of determination 179, 182, 192–3
 covariance 179–81
 estimating linear relationship 179, 186–91
 interpreting correlation 186
 measures of central location 3, 136–49
 arithmetic mean (mean) 137–8

geometric mean 146
 geometric mean rate of return 147–9
 mean, median, mode: which is best? 143–4
 median 138–9
 mode 140–2
 for ordinal and nominal data 144
 relationship between mean, median and mode 144–5
 weighted mean 145–6
 measures of relative standing 169–72, 175–7
 interquartile range (IQR) 172
 outliers 172, 175–7
 percentiles 169–71
 and variability for ordinal data 177
see also box plots
 measures of variability 3, 153–65
 coefficient of variation 163–5
 for ordinal and nominal data 165, 177
 range 153–4
 standard deviation 158–63
 variance 154–8
 median 22, 138–9, 143–4
 Microsoft Excel 11
see also Excel instructions
 midpoint of the modal class 140
 missing data 399–400, 428
 modal class 96, 140
 mode 140–2, 143–4
 models
 deterministic 609
 first-order linear 627
 linear trend 763
 multiple regression 687–733
 polynomial trend 763
 probabilistic 626–7
 simple linear regression 627–62
 time series 753
 moving averages 753–6
 centred 757–8
 multicollinearity 703, 714–18
 multimodal histogram 96
 multinomial experiment (chi-squared test of goodness-of-fit) 583–91
 expected frequencies 584
 observed frequencies 584
 properties 584
 rule of five 590–1
 test statistic and decision rule 585–90
 testing hypotheses 584
 multiple regression models 687–733
 Durbin–Watson test 703, 714–18
 estimating the coefficients and assessing the model 688–707
 interpreting the results 702–3
 model and requirement conditions 687–8
 multicollinearity 703, 714–18
 regression diagnostics – II 714–22
 regression diagnostics – III (time series) 726–33
 regression equation 701–2
 six steps of regression analysis 689–702
t-tests and the analysis of variance 703
 multiplication rule
 for independent events 237
 probability 236–7
 multivariate methods 111
 mutually exclusive events 213, 230, 235

N

negative linear relationship 115
 negative relationship 112
 negatively (left) skewed histogram 95
 negatively skewed distribution 144
 nominal data 21, 22, 23
 alternative hypothesis 469
 chi-squared test of a contingency table 583, 593–605
 chi-squared test of goodness-of-fit 583–91, 614
 comparing two of more sets of 71
 estimating the difference between two population proportions 453–8
 graphical techniques 44, 45–61
 measures of central location 144
 measures of variability 165
 statistical concepts 615
 statistical techniques 614
 summary of tests 614–15
 z-test of the difference between two population proportions 563–70, 614
 z-test of population proportion 517–20, 614
 nominal variables 24
 relationship between two 68–71
 non-independence of the error variable 671–2
 non-linear relationship 115
 non-normality 669–70, 718
 non-response error 40
 non-sampling error 40
 nonparametric statistics, Spearman rank correlation coefficient 667
 normal approximation to the binomial distribution 344–7
 normal distribution 316–22
 approximate binomial probabilities using 344–7
 calculating normal probabilities 317–25
 finding values of Z 325–8
 z_A and percentiles 329
 normal probabilities
 calculating 317–25
 using the computer to find 328–9
 normal random variable 317
 normality, chi-squared test for 608–13
 null hypothesis 468, 469, 471, 474, 475
 numerical data 20, 21, 23, 86
 alternative hypothesis 469
 numerical descriptive measures 135–95, 209
 approximating descriptive measures for grouped data 206–7
 flowchart of techniques 210
 general guidelines on exploration of data 193–4
 measures of association 17–82, 136
 measures of central location 136–49
 measures of relative standing and box plots 169–77
 measures of variability 153–65
 summary of formulas 195–6
 summation notation 203–5
 numerical techniques 209
 numerical variables 24
 graphing the relationship between two 111–17
 graphing the relationship between three 117–18

O

observational data 27, 642
 observed frequencies 584
 ogives 100–2
 one-tail tests 472–3
 setting up the hypothesis 485–8
 online resources 13
 operating characteristic (OC) curve 515–16
 oral presentations 130
 ordinal data 21, 22, 23
 graphical techniques 61–2
 measures of central location 144
 measures of relative standing 177
 measures of variability 165, 177
 ordinal scale 21
 ordinal variables 24
 origin 240
 outliers 172, 175–7, 672–3

P

p-value of a test of hypothesis 491–502
 calculation of 493
 describing 496
 drawing conclusions 495
 interpreting 493–5
 and rejection region methods 495, 496–502
 Paasche price index 810–11
 panel data 24
 parameters 5, 19, 136
 parametric tests 468
 Pearson coefficient of correlation 179–85, 664
 percentage of trend 768
 percentiles 169–70
 locating 170–1
 and z_A 329
 personal interview 28
 pie charts 46–54
 point estimator 375
 of a population proportion 403
 Poisson distribution 284–8
 Poisson experiment 285
 Poisson probability
 table of 847–8
 using the computer to find 288
 using Poisson table to find 287
 Poisson probability distribution 285–6
 Poisson random variables 285
 mean and variance 288
 Poisson table 286–7
 polynomial trend models 763
 pooled proportion estimates 564
 pooled variance estimate 442, 445
 population 4, 5, 19, 45
 and probability distributions 267
 population coefficient of correlation 181, 664, 665
 population mean 137, 269–70, 355, 374
 estimating the difference between two population means when the variances are known: independent samples 431–7
 estimating the difference between two population means when the variances are unknown: independent samples 439–45
 estimating the difference between two population means when the variances are unknown and unequal 439–42

estimating the difference between two population means when the variances are unknown but equal 442–4
 estimating when population variance is known 378–87
 estimating when population variance is unknown 394–400
 estimation using the sample median 386–7
 sample size required to estimate 411–13
t-test of 508
 testing the difference between two population means: independent samples 531–45
 testing the difference between two population means: matched pairs experiment 551–9
 testing when the population variance is known 476–89, 529
 testing when the population variance is unknown 504–8
 violations of required conditions (two populations) 445
 z -interval estimator 436
 z -test of 489
 population parameters 19, 136, 181, 212, 354, 364, 370
 population proportion(s) 366, 374
 estimating the difference between two 453–8
 estimating the 403–8
 hypothesis testing 517–21
 point estimator of 403
 probability of a Type II error 521
 sample size required to estimate 413–15
 sampling distribution of the difference between two 454, 562–3
 selecting sample size to estimate difference between two 456–8
 testing the difference between two 562–70
 z -statistic for 517, 564
 population standard deviation 270, 355
 population variance 156–7, 270, 355
 difference between two population means when the population variances are known 431–7
 difference between two population means when the population variances are unknown 439–45
 estimating population mean when population variance is known 378–87
 estimating population mean when population variance is unknown 394–7
 testing population mean when the population variance is known 476–89, 529
 testing population mean when the population variance is unknown 504–8
 portfolio diversification 296–9
 with more than two shares 299
 in practice 298
 with two shares 297–8
 positive linear relationship 115
 positive relationship 112
 positively (right) skewed histogram 95
 positively skewed distribution 144
 posterior probabilities 244
 power curve 516
 prediction interval 659, 662

predictive analytics 11
 prescriptive analytics 11
 presenting statistics 129–30
 price index
 average of relative 806–8
 Fisher 811–12
 Laspeyres 808–10, 811
 Paasche 810–11
 simple 801–4
 simple aggregated 804–7
 weighted aggregated 808
 see also Australian Consumer Price Index (CPI)
 primary data 27, 28
 probabilistic models 626–7
 probability 212–52
 addition rule 235–6
 of an event 214–15
 of an outcome 213
 assigning probabilities to events 212–20
 Bayes' law 244–6
 classical approach 214
 complement rule 234–5
 conditional 224–30, 244–6
 cumulative 279
 identifying the correct method 251
 independent events 225–6, 230
 interpreting 220
 joint 224–30, 251
 marginal 224–30, 290–1
 multiplication rule 236–8
 relative frequency approach 214
 requirements 213
 rules 234–8
 subjective approach 214
 summary of formulas 252
 three approaches to assigning 214
 trees 239–42
 of Type II error (hypothesis testing) 510–16
 of Type II error (population proportion) 521
 probability density functions 310–13
 probability distributions 263
 discrete 263–6
 Poisson 285–6
 and populations 267
 probability trees 239–42
 problem objectives and data type 25, 351–2
 published data 27

Q

quadratic model for long-term trend 763
 qualitative data 21, 23
 quantitative data 20, 23
 quartiles 169–70, 172
 questionnaire design 28–9

R

random experiment 212–13
 random numbers 861
 random variables 261–2
 discrete and continuous 262–3
 random variation 753
 range 3, 153–4
 range approximation of the standard deviation 162
 ranked data 21, 22, 23
 reciprocal transformation 719

recoding data 428
 rectangular probability distribution 313
 regression, time series forecasting with 785–94
 regression analysis 625
 examples 625–6
 six steps for 689–702
 see also multiple regression; simple linear regression model
 regression diagnostics – I 667–75
 heteroscedasticity 670–1
 influential observations 673–4
 non-independence of the error variable 671–2
 non-normality 669–70
 outliers 672–3
 procedure 674–5
 residual analysis 667–9
 regression diagnostics – II 714–22
 informal diagnostic procedures 714
 multicollinearity 714–18
 remedying violations of required conditions using transformations 718–22
 regression diagnostics – III (time series) 726–33
 Durbin–Watson test 726–33
 regression equation 659–62, 701–2
 regression line, fitness of the 631–5
 rejection region method 470
 and *p*-value tests 495, 496–502
 relative efficiency 377–8
 relative frequency approach (probabilities) 214
 relative frequency distribution 46
 relative frequency histograms 92–3
 research hypothesis 468
 residual analysis 667–9
 residuals 631
 response surface 687–8
 rule of five
 contingency table 602–3
 multinomial experiment 590–1
 rules of probability 234–8

S

sample 4, 5, 19, 45
 variance of a 157
 sample average 20
 sample coefficient of correlation 181, 664
 sample mean 137, 399
 sampling distribution of 354–64
 standard error of 359
 sample median, estimating population mean using 386–7
 sample proportion(s)
 mean and standard deviation of 367
 sampling distribution 366–9, 404, 517
 sampling distribution of the difference between two 454
 standard error of the 367
 sample size 36
 determining the required 410–15
 required to estimate population mean 411–13
 see also sample proportion
 sample size required to estimate population proportion 413–15
 to estimate the difference between two population means 436–7
 to estimate the difference between two population proportions 456–8

- sample space 213
 sample statistic 136, 399
 sampled population 30–1
 sampling 18–19, 30–1
 cluster 36
 error 31, 39
 from a finite population 364
 simple random 32–4
 stratified random 35–6
 sampling distribution 354
 creating empirically 360
 of the difference between two sample means 431–2
 of the difference between two sample proportions 454, 562–3
 linear model 648–9
 linked to statistical inference 351, 369–70
 of the sample mean 354–64
 of the sample proportion 366–9, 404, 517
 summary of formulas 371
 of the *t*-statistic 648
 sampling plans 32–8
 scatter diagrams 111–13, 182, 625, 673–4
 direction 115
 interpreting a strong linear relationship 115
 linearity 114
 seasonal effect, measuring 772–9
 seasonal indexes 772
 alternative method to estimate S_t
 R_t 775–7
 deseasonalising a time series 777–8
 estimating 772–5
 forecasting with 786–9
 seasonal variations 752
 forecasting with dummy variables 790–1
 secondary data 27
 secular trend 750
 selection bias 40
 self-administered survey 28
 self-selected samples 31
 significance level 6, 470–1
 simple aggregate price index 804–7
 simple event 214
 simple linear regression models 627–62
 ANOVA table 653
 applications in finance: market model 654
 assessing the model 643–56
 cause-and-effect relationship 654
 coefficient of determination 650–3
 error variable: required conditions 641–2
 estimating the coefficients 628–36
 F-test and *t*-test 703
 fitness of the regression line 631–5
 interval estimates 646–7
 least squares regression line 629–31
 regression diagnostics 667–75
 regression equation 659–62
 sampling distribution 648–50
 standard error of estimate 644–5
 sum of squares for error 631, 643–4
 summary of formulas 678
 test statistic 648–50
 testing the slope 647–50
 simple price index 801–4
 simple random sampling 32–4
 skewed distribution 144
 skewed histograms 95–6
- slope coefficient estimator 646
 smoothed relative frequency polygons 153
 smoothing techniques 753–61
 centred moving averages 757–8
 exponential smoothing 758–61
 moving averages 753–6
 time series forecasting with exponential smoothing 783–5
 Spearman rank correlation coefficient 667
 spreadsheets 12–13
 square-root transformation 719
 squared transformation 718
 stacked data 544, 545
 stacking data 581
 standard deviation 154, 158–63
 Chebyshev's theorem 162–3
 interpreting 161–3
 of the *i*th residual 669
 of population 270
 range approximation of the 162
 of the sample proportion 367
 of the sampling distribution of the sample mean 357
 standard error of the difference between two means 432, 545
 standard error of estimate 644–5, 693
 standard error of the sample mean 359
 standard error of the sample proportion 367
 standard normal distribution 317, 379
 standard normal probability table 317–19
 standard normal random variable (Z) 317, 319–20
 finding values of 325–8
 statistic 5
 statistical analysis 10
 statistical applications in business 6–9
 statistical concepts 5–6, 378, 399–400, 489, 545, 559, 653, 733
 statistical inference 4, 5–6, 351–71
 linked to sampling distribution 351, 369–70
 systematic approach 352–3
 statistical tables
 binomial probabilities 843–6
 critical values of the chi-squared distribution 852
 critical values for the Durbin–Watson statistic 862–3
 critical values of the Student *t* distribution 851
 cumulative standardised normal probabilities 849–50
 Poisson probabilities 847–8
 random numbers 861
 values of the *F*-distribution 853–60
 statistical techniques, three-stage solution process 353, 380–2
 statistically significant 496
 statistics 2–6
 and the computer 11–12
 descriptive 3–4, 45
 inferential 4–5, 45
 introduction to 2–5
 key concepts 5–6
 managers use of 9–11
 and missing data 399–400, 428
 presenting 129–30
 stem-and-leaf displays 94–5
 stochastic models 627
- stratified random sampling 35–6
 Student *t* distribution 391–4, 399–400
 checking the required conditions 397–8
 critical values 851
 using the computer 393–4
 Student *t* random variable 394
 subjective approach (probability) 214
 success (Bernoulli trial) 275
 sum of squares for error (SSE) 631, 643–4
 sum of squares for forecast error (SSFE) 780, 781
 sum of two variables 293–4
 laws of expected value and variance 294
 summary of tests for nominal data 614–15
 summation notation 203–5
 rules of 205
 terminology and notation 203–4
 surveys 28–9
 symbols 195, 303, 341, 371, 422, 616, 677, 736, 796
 symmetric histograms 95, 96

T

- t* distribution 391
t-interval estimator of difference between population means
 equal variances 444
 unequal variances 442
t-statistic 297, 391, 399
 sampling distribution 648
t-test
 and analysis of variance (regression analysis) 703
 and estimator of the mean of the population of differences 559
 and *F*-test (simple linear regression model) 703
 of population mean 508
t-test of the difference between population means
 equal-variances 543
 unequal variances 539
 target population 30, 31
 telephone interview 28
 test statistic 469–70, 489
 chi-squared goodness-of-fit 585–90
 chi-squared test of a contingency table 594–600
 of the difference between two population proportions 564
 and factors that identify their use 532
 formulas 524, 574
 for the mean of the population of differences 559
 multiple regression model 698
 population coefficient of correlation 665
 population mean when population variance is unknown 504
 population proportion 517
 simple linear regression model 648–50
 time series 749
 components 749–53
 deseasonalising 777–8
 measuring the cyclical effect 768–71
 measuring the seasonal effect 772–9
 models 753
 smoothing techniques 753–61
 trend analysis 763–7

time-series chart 106
 time-series data 24, 106–8
 Durbin–Watson test 726–33
 which chart is better? 108–9
 time-series forecasting
 with exponential smoothing 783–5
 with regression 785–94
 seasonal time series with dummy variables 786, 790–1
 with trend and seasonality 786
 trade, Australian 2
 transformations 718–19
 trend 750
 percentage of 768
 trend analysis 763–7
 two-tail test 472
 Type I error 470, 471, 472, 703
 Type II error 470, 471, 472
 probability of (hypothesis testing) 510–16
 probability of (population proportion) 521

U

unbiased estimator 376, 377–8
 unequal class widths 91, 93
 unequal-variances *t*-interval estimator of difference between population means 442
 unequal-variances *t*-test of the difference between population means 539
 unequal-variances *t*-test statistic 532
 uniform distribution 313–15
 uniform probability density function 313
 uniform probability distribution 313
 unimodal distribution 140

unimodal histogram 96
 union 216–17
 univariate methods 111
 unstacked data 544, 545
 unstacking data 581
 unweighted index numbers 801–8
 upper confidence limit (UCL) 380
 upper fence 173

V

values 20
 variables 20
 dummy 786, 790–1
 sum of two 293–4
 variance 154–8, 269–72
 of binomial distribution 281–3
 of the difference between two sample means 432
 laws of 272–3, 294
 of Poisson random variables 288
 of population 156–7, 270
 of portfolio of k shares 299
 of portfolio of two shares 297
 of a sample 157
 of the sampling distribution of the sample mean 357
 see also population variance
 variation, coefficient of 163–5
 Venn diagram 215–16

W

wages, deflating using CPI 816
 weighted aggregated price index 808

weighted index numbers 808–12
 weighted mean 145–6
 written reports 129–30

X

XLSTAT 11, 12

Y

Yearbook of National Accounts Statistics 27

Z

Z 317, 319–20
 finding values of 325–8
 z-distribution 317
 z-interval estimator of the difference between two population means 436
 z-test
 for the difference between two population proportions 563–70, 614
 of the differences in population means, known variances 534
 of the population mean 489
 for a population proportion 517–20, 614
 z-test statistic
 equal variances 532
 known variances 552
 unequal variances 532
 z_A 325–8
 and percentiles 329

This is an electronic version of the print textbook. Due to electronic rights restrictions, some third party content may be suppressed. Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. The publisher reserves the right to remove content from this title at any time if subsequent rights restrictions require it. For valuable information on pricing, previous editions, changes to current editions, and alternate formats, please visit www.cengage.com/highered to search by ISBN, author, title, or keyword for materials in your areas of interest.

Important notice: Media content referenced within the product description or the product text may not be available in the eBook version.