STATISTICS & PROBABILITY

for the Australian Curriculum

Years 9&10

Helen MacGillivray & Peter Petocz

Consultants: Michael Evans and Peter Jones

CAMBRIDGE
UNIVERSITY PRESS

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.edu.au
Information on this title: www.cambridge.org/9781107655997

First published 2014

Cover designed by Sean Walsh
Typeset by Kerry Cooke
Printed in China by C & C Offset Printing Co. Ltd.

A Cataloguing-in-Publication entry is available from the catalogue of the National Library of Australia at www.nla.gov.au

ISBN 978-1-107-65599-7 Paperback

Additional resources for this publication at www.cambridge.edu.au/GO

CONTENTS

ISBN 978-1-107-65599-7

arrays. Assign probabilities to outcomes and determine probabilities for events **(ACMSP225)**

Calculate relative frequencies from given or collected data to estimate probabilities of events involving 'and' or 'or' **(ACMSP226)**

Describe the results of two- and three-step chance experiments, both with and without replacements, assign probabilities to outcomes and determine probabilities of events. Investigate the concept of independence **(ACMSP246)**

Use the language of 'if ... then', 'given', 'of', 'knowing that' to investigate conditional statements and identify common mistakes in interpreting such language **(ACMSP247)**

Data representation and interpretation

Determine quartiles and interquartile range **(ACMSP248)**

Construct and interpret box plots and use them to compare data sets **(ACMSP249)**

Compare shapes of box plots to corresponding histograms and dot plots **(ACMSP250)**

Data representation and interpretation

Use scatter plots to investigate and comment on relationships between two numerical variables **(ACMSP251)**

Investigate and describe bivariate numerical data where the independent variable is time **(ACMSP252)**

ISBN 978-1-107-65599-7

Foreword

Statistics and probability are of great importance in our world today. For example, nearly every aspect of government planning is based on the careful statistical analysis of data obtained in the census carried out by the Australian Bureau of Statistics. Statistics is also used extensively in medical and scientific research and planning and forecasting in economics and commerce in general. Indeed, statistics and probability are essential components in the operation of any organisation which has some complexity.

The Australian Curriculum: Mathematics provides an opportunity to change and improve the teaching and learning of statistics and probability. This book is a major step towards achieving this goal. The courses being designed for Years 11 and 12 will require a thorough understanding of the ideas being introduced in this series of two books for Years 7 to 10.

This book emphasises real and everyday contexts engaging and familiar to students. Data concepts and tools are developed holistically using the statistical investigation process with real datasets. The rich experiential approach enables the user to go beyond a mere collection of techniques that has often been introduced at this stage.

The authors are deservedly highly respected both in the Australian and international statistics education communities. This book makes a substantial contribution to the teaching and learning of probability and statistics for students and teachers at this level.

Dr Michael Evans
Australian Mathematical Sciences Institute

ISBN 978-1-107-65599-7

Introduction

Statistics and statistical thinking have become increasingly important in a society that relies more and more on information and demands for evidence. Hence the need to develop statistical skills and thinking across all levels of education has grown and is of core importance in a century which will place even greater demands on society for statistical capabilities throughout industry, government and education.

A natural environment for learning statistical thinking is through experiencing the process of carrying out real statistical data investigations from first thoughts, through planning, collecting and exploring data, to reporting on its features. Statistical data investigations also provide ideal conditions for active learning, hands-on experience and problem solving. Hence the data knowledge and skills developed in this book are embedded in the data investigation and interpretation process and examples of it.

Statistics is the science of variation and uncertainty. Concepts of probability underpin all of statistics, from handling and exploring data to the most complex and sophisticated models of processes that involve randomness. Statistical methods for analysing data are used to evaluate information in situations involving variation and uncertainty, and probability plays a key role in that process. All statistical models of real data and real situations are based on probability models. Probability models are at the heart of statistical inference, in which we use data to draw conclusions about a general situation or population of which the data can be considered randomly representative. Hence the knowledge and skills of chance developed in this book lay the foundations for understanding the processes of modelling probabilities and are integrated with the use of, and applications to, data.

ISBN 978-1-107-65599-7

About the authors

Helen MacGillivray is an Adjunct Professor of Mathematical Sciences at the Queensland University of Technology (QUT). She is currently a vice-president of the International Statistical Institute, joint editor of Teaching Statistics and a past president of the International Association for Statistical Education and of the Statistical Society of Australia. Helen has 40 years experience of teaching and curriculum design in statistics, including many years experience with curriculum and professional development in statistics for secondary school. She was one of the first Australian Senior Learning and Teaching Fellows, a 2011 Australian Citation winner for Outstanding Contributions to Student Learning, and a 2003 Finalist in Australian Universities Teaching Awards. As a consultant to the Australian Mathematical Sciences Institute (AMSI), she is author of 'The Improving Mathematics Education in Schools (TIMES) Modules for Statistics and Probability in the Australian Curriculum: Mathematics'.

Peter Petocz is a lecturer in teaching development in mathematics and statistics at Macquarie University. He has more than 15 years experience with publications in the area of learning and teaching mathematics and statistics, particularly with the preparation and evaluation of learning materials. His teaching awards include a 2006 Citation for Outstanding Contributions to Student Learning at the Carrick Australian Awards for University Teaching, and he was a 2003 Finalist at Australian Universities Teaching Committee National Teaching Awards, Canberra.

About the consultants

Dr Michael Evans is at the Australian Mathematical Sciences Institute (AMSI). He has been involved in the development of the Australian Curriculum: Mathematics and is a key author for the *ICE-EM Mathematics* series and two *Essential Mathematics* textbooks for senior courses.

Peter Jones is a Professor Emeritus and formerly Head of the School of Mathematical Sciences at Swinburne University of Technology. He has been involved in the development of the Australian Curriculum: Mathematics and is the lead author on two *Essential Mathematics* textbooks. His area of expertise is applied statistics.

ISBN 978-1-107-65599-7

How to use this book

To get the best out of this book, supporting resources are provided on the website. The free enrichment activities, and the further investigations and activities in the Teacher Resource Package will enable students to use the statistical inquiry cycle and to integrate the skills and concepts from the textbook into holistic assignments.

The textbook has the following features in each chapter for use in class, for homework or for assignments:

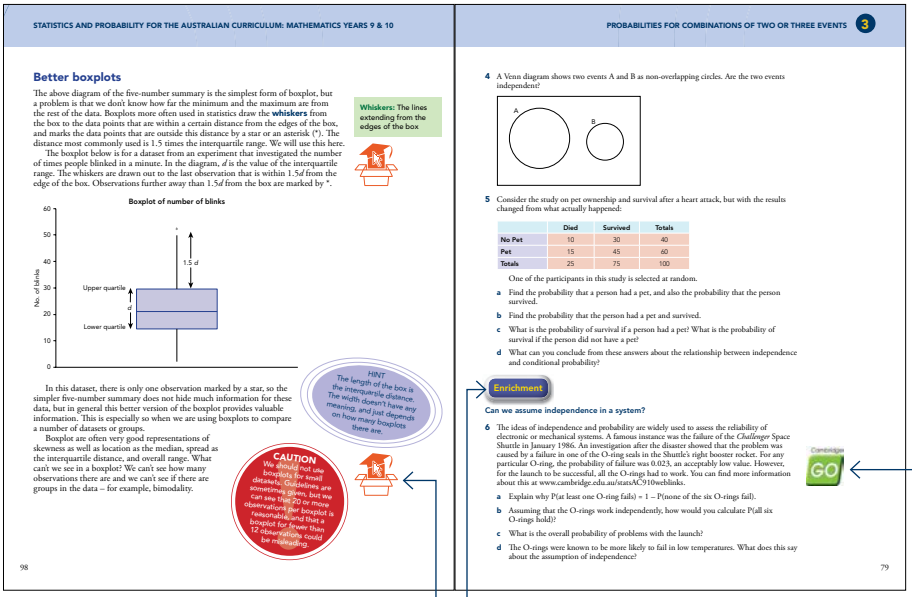
- What you will learn — a list of topics in the chapter
- Chapter introduction: this sets the scene by posing questions that will be addressed in the chapter
- Australian Curriculum linkage for the chapter
- Pre-test — a check of prior knowledge for the chapter
- Terms you will learn — a checklist of new terms to be met in the chapter
- Chapter topic introductions and discussion
- Glossary — words in bold in the text are defined in the margin
- Let's Start — an activity that can be done in groups or individually or as a class discussion
- Key ideas are summarised in boxes
- Examples include full explanations
- Exercises consist of extended response questions
- Hints and cautions are provided in the margin
- Enrichment questions
- Chapter summary
- Multiple choice questions
- Short answer questions
- Extended response questions

At the end of the book you will find:

- A glossary with definitions from the margin organised by chapter for easy revision and reference
- Answers to all questions

ISBN 978-1-107-65599-7

Explanation of icons in the textbook:



GO icon
Information or feature on the website (see the following pages for access details)

Teacher resource package icon
Linked material provided in the Teacher Resource Package available through the website

Enrichment icon
Enrichment activities

Material for students on the website provided with the textbook:

- A digital copy of the textbook with note-taking enabled
- Data sets and graphs in spreadsheets
- Weblinks

Material for teachers included with this textbook

- A syllabus guide
- Updates as required

Resources in the Teacher Resource Package

- Further investigations and activities in worksheet format
- Notes for teachers
- Solutions to questions
- Chapter tests

ISBN 978-1-107-65599-7

Cambridge GO

YOUR GATEWAY ONLINE

Digital resources and support material for schools.

THIS TEXTBOOK IS SUPPORTED BY ONLINE RESOURCES

Additional resources are available free for users of this textbook online at *Cambridge GO* and include:

- the PDF Textbook – a downloadable version of the student text, with note-taking and bookmarking enabled
- activities in Word format
- links to other resources.

Use the unique 16-character access code found in the front of this textbook to activate these resources.

www.cambridge.edu.au/go

For more information or help contact us on 03 8671 1400 or enquiries@cambridge.edu.au

ISBN 978-1-107-65599-7

Access your online resources today at www.cambridge.edu.au/go

1. Log in to your existing *Cambridge GO* user account or create a new user account by visiting: www.cambridge.edu.au/GO/newuser

   - All of your *Cambridge GO* resources can be accessed through this account.
   - You can log in to your *Cambridge GO* account anywhere you can access the internet using the email address and password with which you are registered.

2. Activate *Cambridge GO* resources by entering the unique 16-character access code found in the front of this textbook.

   - Once you have activated your unique code on *Cambridge GO,* it is not necessary to input your code again. Just log in to your account using the email address and password you registered with and you will find all of your resources.

3. Go to the My Resources page on *Cambridge GO* and access all of your resources anywhere, anytime.*

* Technical specifications: You must be connected to the internet to activate your account. Some material, including the PDF Textbook, can be downloaded. To use the PDF Textbook you must have the latest version of Adobe Reader installed.

ISBN 978-1-107-65599-7

Acknowledgements

The author and publisher wish to thank the following sources for permission to reproduce material:

Images: ©Shutterstock – 2013 Used under license from Shutterstock.com/Pressmaster, cover/Pixsooz, cover/ Goodluz, cover/lightpoet, cover/Peshkova, cover/Monkey Business Images, p.1, 86/izf, p.3/John Milnes, p.7/Amma Cat, p.8/Diego Vervo, p.9/Zadrozhni Viktor, p.10/MelBrackstone, p.13/Dan Kosmayer, p.15/Mat Hayward, p.17/SeDmi, p.18/ Olga Danylenko, p.19/Rtimages, p.20/Vacclav, p.22/Sean De Burca, p.23/ LeventeGyori, p.25/FXQuadro, p.26/graph, p.27/Seregam, p.28/Radu Razvan, p.30/Halso Group Production Studio, p.24/LeventeGyori, p.25/FXQuadro, p.26/graph, p.27/Seregam, p.28/ Radu Razvan, p.30/BGSmith, p.31/fotografos, p.32/Ciureu Adrian, p.33/mtkang, p.34/ Konstantin Sutyagin, p.37/tobkatrina, p.38/Constantine Pankin, p.39/Jakub Vacek, p.41/ kitch Bain, p.42/Maiwharn, p.42/racorn, p.44/Tupungato, p.46/ Phase4 Photography, p.47/duckeesue, p.48/Tobias Arhelger, p.51/maxhphoto, p.53/FlashStudio, p.53/Jiri Hera, p.55/Pressmaster, p.55/Rosil Othman, p.56/wavebreakmedia, p.58,64,95/Shawn Hempel, p.59/ evan travels, p.60/Pot of Grass Productions, p.61/Anna Baburkina, p.62/Nikita Rogul, p.63/Kitti Sukhonthanit, p.65/Cheryl Ann Quigley, p.67/auremar, p.68/S_L, p.69/ muzsy, p.72/CURphotography, p.73/315 studio by khunaspix, p.74/ Roman Sotola, p.76/ oksix, p.77/Sirikorn Techatraibhop, p.78/quinky, p.80/thory, p.83/wikimedia commons, p.87/Jor Gough, p.89/Hannamariah, p.90/bikeriderlondon, p.91/ Chen ws, p.93/Ana de Sousa, p.94/K. Miri Photography (L) & McCarthy's PhotoWorks (R), p.92/Andy Dean Photography, p.101/pavla, p.102/CandyBox Images, p.104/ahmad faizal yahya, p.107/ Karlien du Plessis, p.108/Edward Haylan, p.112/ Avava, p.113/Mandy Godbehear, p.114/ Ryan R Fox, p.116/yexelA, p.117/Ilya Andriyanov, p.119/ aerogondo2, p.123/ Cameron Whitman, p.124/ifong, p.125/Genenacom, p.127/ejwhite, p.129/Lasse Kristensen, p.130(t)/"Free material from www.gapminder.org", p.130(b),134(l),134(r),144(b)/jodi Hutchison, p.132/Alfonso de tomas, p.135 /Stephen Rees, p.136/Artazum and Iriana Shiyan, p.138/Instinia, p.140/Diego Barbieri, p.142/Zsolt Biczo, p.143/Tupungato, p.145(t).

Text: "All curriculum content © Australian Curriculum, Assessment and Reporting Authority 2011".

Every effort has been made to trace and acknowledge copyright. The publisher apologises for any accidental infringement and welcomes information that would redress this situation.

ISBN 978-1-107-65599-7

CHAPTER 1

Quantitative data across groups

What you will learn

How well can people judge distances?

Knowing how well people judge distances can be important in general and in a number of areas, for example in designing guidelines, rules and signs for traffic and roads. It is also often said that males are more spatially oriented than females, so are males better at judging distances than females?

This was investigated in a case study for a distance of 5 metres. Male and female subjects between the ages of 16 and 40 years old were randomly selected to help in the case study. The tester would hold a tape measure upside down so that no numbers were visible with the subject holding the end. The tester would walk away until the subject thought the tester had walked 5 metres and this distance was recorded.

What would the investigators be interested in exploring in this dataset? They were interested in how close people's guesses were to 5 metres, how much variation there was, how males compared with females in how close their guesses were to 5 metres, and whether males' and females' guesses had about the same amount of variation. How much people vary is very important, because allowances have to be made for the fact that people are not the same. How the 'average' person reacts is just one bit of information in any situation.

AUSTRALIAN CURRICULUM

Statistics and probability

- Data representation and interpretation
- Identify everyday questions and issues involving at least one numerical and at least one categorical variable, and collect data directly and from secondary sources **(ACMSP228)**
- Construct back-to-back stem-and-leaf plots and histograms and describe data **(first part of ACMSP282)**

ISBN 978-1-107-65599-7

Pre-test

Terms you will learn

back-to-back stem-and-leaf plot
bin
cumulative frequency plot
cumulative histogram
distributed
histogram
placebo
same scale

1 In planning a data investigation, give at least three steps that must be performed before the data collection starts.

2 In investigating pollution in a river, the amount of one type of nutrient in milligrams in water samples of 100 millilitres was used to measure pollution. The water samples were taken at three locations along the river on Mondays and Fridays of 12 weeks.

a There are four variables in this dataset. What are they?

b Give the types of each of these four variables.

c On the data recording sheet or spreadsheet for this investigation, the four variables would each have a column. What would the rows of the sheet correspond to?

3 A company wanted to compare their new sunscreen with their current one; both are rated SPF30+. The company had a number of male and female volunteers of different ages (in years) and skin types (fair, medium and dark). For each volunteer, the current sunscreen was put on one arm, and the new sunscreen on the other arm. The volunteers then stayed in the sun until their skin started to redden and the time until the start of reddening was recorded for each arm for each volunteer. Then the difference between arms of time to start of reddening was calculated for each volunteer.

a What type of investigation was this: survey, experiment or observational study?

b There were five variables in this investigation. What were they?

c Give two practical aspects of this investigation that would require care.

4 In a public transport investigation, the numbers of passengers getting off each bus arriving at a city bus station were recorded during the two periods 7.30–8.30 am and 10–11 am each Tuesday and Wednesday for three weeks.

a What type of investigation was this: survey, experiment or observational study?

b There were four variables. What were they and what were their types?

5 In the investigation in question 4, a random sample of passengers from each bus were asked how often they used public transport each week.

a What type of investigation was this: survey, experiment or observational study?

b What population was being sampled?

c What type of sampling is this called?

6 The times in seconds between the first 20 phone calls arriving at an office were recorded one morning, giving

20 58 7 1 35 57 104 2 43 53 92 189 14 41 13 108 138 69 4 55

a Use a stem-and-leaf plot to graph these observations.

b Find the mean, median and range of these data.

c On the stem-and-leaf plot, how many modes are there and what are their values? How many modes are there in the original data above?

ISBN 978-1-107-65599-7

1-1 Investigating to compare

Questions that compare groups are some of the most common in all areas in which variation happens and data are collected. In medicine, is a new drug effective or more effective than a current one? Does it have more side effects for some groups of patients than others? In engineering, is a new process more efficient? Is one maintenance program better than another? In psychology, how does the effectiveness of a program to improve memory vary across age groups? In government, does the public opinion on a policy vary across regions and age groups? In agriculture, how does the yield of a crop vary for different combinations of fertilisers and soil conditions? In science, how do the results of a chemical reaction vary if other chemicals are present or not?

In Year 6 Data strand, and in Years 7–8 Chance strand, you have used side-by-side column graphs and two-way tables to explore how the data for one categorical variable varies over the categories of another categorical variable. For example, how does the support for a government policy vary across age groups? Is the opinion amongst teenagers on getting a suntan different across states? Now we consider how to investigate and explore how quantitative data varies across groups.

CAUTION
Remember that groups are the categories of a categorical variable or a combination of categorical variables.

Planning and collecting data

Remember how important it is to plan the collection of primary data, or to know how data were collected in using secondary data. Remind yourself of the data investigation process which can be represented in diagrams like the one following.

Cambridge GO

Planning a primary data collection

Suppose we investigate reaction times of students by measuring how far down a ruler they catch it when it is dropped vertically from a height. We need a random sample of students, and we need to be very careful to give the same explanation to each student, to use the same ruler dropped from the same height and to make the measurements to the same accuracy, for example, to the nearest 0.5 centimetre. We might want to compare boys and girls or we might want to compare different age groups; we therefore might choose students at random from different age groups.

Using secondary data

Suppose we are interested in how long students spend on Facebook and we find a report of a survey on this. We need to know how the survey selected the subjects, and how the question was asked. For example, were students asked how long approximately in a day, or a week, or were they asked about a particular day – this would have to be the same day.

LET'S START Judging distances

In the investigation described at the beginning of this chapter, 70 people between the ages of 20 and 40 were randomly chosen for the test. As described, each person watched as the tester walked away from them holding a tape measure upside down, and called out when they thought the tester had walked 5 metres. Measurements were made correct to centimetres. The investigation recorded whether the subjects were male or females, and whether they wore glasses or contact lenses at all. If they did, they were asked to wear what they would normally in outside conditions.

ISBN 978-1-107-65599-7

Below are the guesses of the 25 males and the 19 females who did not wear glasses or contact lenses:

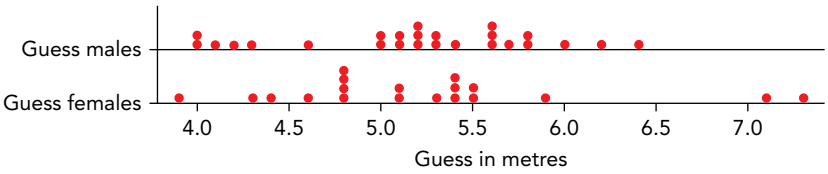
Males

5.06 5.78 4.95 5.40 5.26 5.61 4.61 5.56 5.17 5.76 4.13 5.31 5.21 5.57
4.96 5.98 3.98 6.44 4.22 4.32 5.09 6.19 5.21 5.66 4.01

Females

5.38 7.11 3.87 5.11 5.06 5.40 5.40 5.88 4.44 4.75 4.75 4.31 5.34 5.52
5.48 4.80 4.62 7.32 4.77

The experiment was carefully carried out with all measurements being made in the same way, so we can compare the guesses of the males and females. What have you used to explore measurement data – dotplots and stem-and-leaf plots? For either, we must use the **same scale** to be able to compare the two groups of data. Below are dotplots on the same scale.



Back-to-back stem-and-leaf plots

We can draw two stem-and-leaf plots on the same scale and using the same leaf interval, and put them side by side. A good way of placing them side by side is to put them back to back, which gives us the name **back-to-back stem-and-leaf plot**.

Same scale: Refers to plots having the same range of values on the *x*-axis and the same distances between these values … *see glossary*

Back-to-back stem-and-leaf plot: Two stem-and-leaf plots placed side by side with a common stem … *see glossary*

ISBN 978-1-107-65599-7

Leaf unit = 0.1

Female guesses		Male guesses
8	3	9
	4	01
3	4	23
4	4	
7776	4	6
8	4	99
10	5	001
33	5	2223
5444	5	455
	5	6677
8	5	9
	6	1
	6	
	6	4
	6	
	6	
1	7	
3	7	

From this plot, we can find what we want for each group and see how the two compare. For example, there were $\frac{6}{25}$ males who guessed between 4.8 and 5.2 m, and $\frac{3}{19}$ females.

Key ideas

- Comparing quantitative data across groups involves a continuous variable and a categorical variable.
- For both primary and secondary data, the quantitative data must be collected in the same way and to the same accuracy across groups.
- In using plots to compare quantitative data across groups, we must use the same scale.
- Back-to-back stem-and-leaf plots can be used to compare quantitative data across two groups.

Example 1: Is coral density close to the coast different from that away from the coast?

The density, in gram per cubic centimetre to the nearest 0.01, of the heads of a type of coral in the Great Barrier Reef was measured by scientists at a number of reefs that are different distances from the coastline. Below, the measurements are split into two groups – at reefs less than 20 kilometres from the coast and at reefs more than 20 kilometres from the coast.

ISBN 978-1-107-65599-7

Density at reefs < 20 km from coast

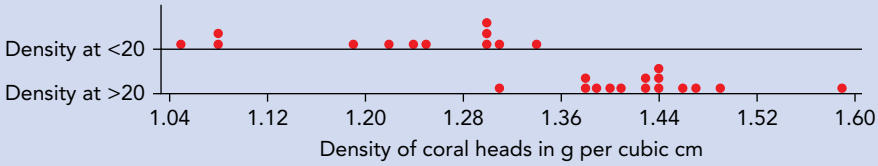
1.34 1.22 1.31 1.05 1.08 1.08 1.24 1.19 1.30 1.25 1.30 1.30

Density at reefs > 20 km from coast

1.38 1.38 1.31 1.44 1.49 1.47 1.39 1.44 1.43 1.41 1.40 1.43 1.44 1.59 1.46

Question: How do the densities compare?

We can explore how the densities compare by either dotplots on the same scale or back-to-back stem-and-leaf plots, as below.



Leaf unit = 0.01

More than 20 km		Less than 20 km
	10	588
	11	
	11	9
	12	24
	12	5
1	13	00014
988	13	
4443310	14	
976	14	
	15	
9	15	

In both types of plots, we can see that all but one of the reefs less than 20 kilometres from shore have densities of coral heads less than those for reefs more than 20 kilometres from shore.

Exercise 1A

1 A 'force platform' can be used to measure balance. Subjects stand on it with bare feet and it measures the amount of sway either sideways or backwards and forwards. The platform automatically measures the amount of sway in each direction in millimetres. An experiment is to be conducted to investigate the effects on balance of concentrating on something other than balancing. Subjects are asked to stay as still as possible on the platform and to react as quickly as possible to a sudden noise that could come at any time. They react by pressing a button as soon as they hear the noise. The investigators are interested in the effects on balance, and in comparing males and females and different age groups.

a What measurements should the investigators take in this experiment?

ISBN 978-1-107-65599-7

b How should the measurements be used in exploring the data?

c Briefly describe what plots could be used and how they could be used to investigate these data.

2 Two exercise scientists are arguing about who are fitter – cricketers or tennis players. They decide to take pulse rates after a short burst of intense exercise as their measure of fitness. The scientists choose a group of first-grade cricketers and tennis players and ask each to undertake the same intense exercise.

a What measurements should the scientists take and what should they use for comparison?

b Give at least one aspect requiring care in conducting the experiment.

3 The lengths of rivers in kilometres (to the nearest km) in the South Island of New Zealand were obtained from information in books and on the internet. They were divided into those which flow into the Tasman Sea and those which flow into the Pacific Ocean. The lengths are:

Into Tasman Sea

76 64 68 64 37 32 32 51 56 40 64 56 80 121 177 56 80 35 72 72
108 48

Into Pacific Ocean

209 48 169 138 64 97 161 95 145 90 121 80 56 64 209 64 72 288 322

a Plot the lengths on dotplots, using the same scale.

b Plot the lengths on a back-to-back stem-and-leaf plot.

c What is the main comparison that shows up in these plots? Can you think of a reason for this comparison?

4 A music fan claims that Alternative rock songs tend to be longer than Indie songs. The fan collects data from the internet for the Triple J top 100 list in a year in which Indie was popular. The lengths in seconds of the songs on this chart for these two music genres are given below.

Lengths of Indie songs

219 200 204 199 203 275 226 186 278 237 208 200 232 250 190 288
233 227 233 226 197 332 239 234 192 182 226 255 130 257 219 248
221 216 254 258

Lengths of Alternative rock songs

499 191 201 200 485 181 406 258 326 298 197 213 181 152 188 220
252 213 362 275 234 250 194

a Plot these lengths on a back-to-back stem-and-leaf plot.

b Check that the median of the lengths of Indie songs is 226 s.

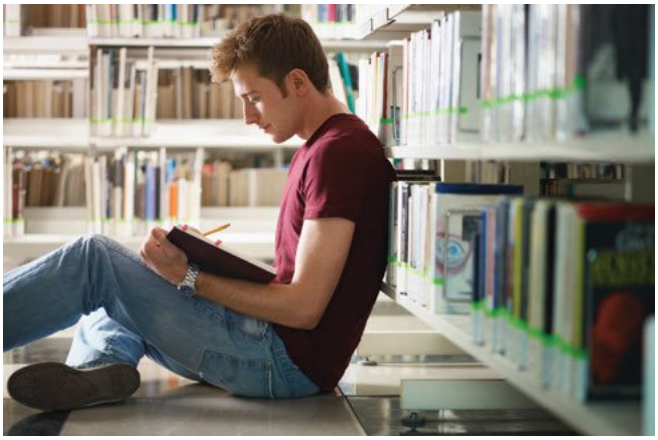
c What proportion of Alternative rock songs have lengths greater than the Indie median?

ISBN 978-1-107-65599-7

Enrichment

Does reading alter perception of time differently for men and women?

5 An experiment was conducted to investigate if people's perception of time is affected by focusing on an activity such as reading. A random sample of people aged between 20 and 40 years were asked to guess when 20 seconds had passed when not reading and when reading. Below are their guesses to the nearest 0.1 s. Note that the observations for not reading and for reading are in the same order of people. That is, the first observation in the list for females reading is for the same person as the first observation in the list for females not reading, and so on.



Females not reading

19.0 13.3 24.5 20.3 23.2 16.4 23.3 21.3 27.2 20.1 18.9 19.6 19.1 22.5
18.6 21.1 21.2 23.1 22.0 20.2 20.6 19.8 21.2

Females reading

19.4 14.6 17.6 19.4 21.4 27.3 23.0 29.4 20.5 26.4 16.1 22.8 25.5 18.8
21.4 21.4 23.4 24.0 19.8 18.7 21.5 18.6 21.2

Males not reading

16.4 23.0 21.1 19.5 24.3 24.2 17.3 20.5 22.6 20.2 18.4 21.8 25.6 28.2
19.5 23.6 21.3 20.5 18.9 18.6 21.3 22.1 24.0 21.6

Males reading

21.3 17.3 19.1 16.3 31.4 19.3 25.4 18.6 24.5 19.2 20.3 23.8 25.2 25.4
22.6 20.1 20.6 21.0 22.1 19.5 18.6 19.7 22.7 20.2

a What could you randomise in this experiment?

b Give an experimental condition that should stay the same.

c What quantities would you use to explore the data for this investigation?

d Use either dotplots or back-to-back stem-and-leaf plots to explore the data for this investigation.

e Comment on at least one feature of the data that you can see in the plots.

ISBN 978-1-107-65599-7

1-2 Histograms

You have seen that quantitative data, particularly data from a continuous variable – like measurement data – is best displayed as frequencies in intervals. That is, the range of the data is divided into intervals and the numbers of observations in those intervals are plotted. This is because data from continuous variables usually have many (if not all!) different values, so unless we collect them into intervals, it is difficult to see how the data are behaving – how the data are **distributed** over the range of values.

Dotplots usually have only a small amount of collecting of observations into intervals. For stem-and-leaf plots, we can choose the size of the interval but we are restricted to 1, 2, 5 or 10 intervals for each digit in the stem. For example, if the leaf unit is 1, then the stem-and-leaf intervals must be one of the following for each digit in the stem:

1 interval: leaf digits are 0, 1, …, 9
2 intervals: leaf digits are 0, 1, …, 5 in first interval and 6, 7, …, 9 in second
5 intervals: leaf digits are 0, 1 in first interval; 2, 3 in second interval; …; 8, 9 in 5th interval
10 intervals: each leaf digit is in a different interval.

You have seen how useful stem-and-leaf plots are, but, as well as having some restrictions on the intervals, the digits in the leaves can also be a bit distracting. And once we start to put stem-and-leaf plots back-to-back, it takes a bit of looking to see how each dataset is distributed and how they compare. And it becomes very difficult if we have three or more groups to compare.

Another plot for quantitative data that works by collecting observations into intervals is a **histogram**. Like the stem-and-leaf plot, we divide a suitable range of values that include all our observations, into a number of equal intervals. These intervals are

Distributed: How the data are spread over the range of values

CAUTION
Have a look at the examples of back-to-back stem-and-leaf plots in section 1-1. Notice that you have to look carefully at them to see how the data are distributed in each group and in comparing the two groups.

Histogram: A (simple) histogram is a graph of frequencies of quantitative data grouped into equal intervals which cover the range of the data … *see glossary*

called **bins**; when we collect the data into the intervals, we are putting them into bins – or binning them! A 'box' then sits on each interval, with the height of the 'box' giving the frequency of observations in that interval. Note that a histogram is *not* a column graph or a bar chart. Histograms are for data that are collected into intervals; the bins must touch each other and the rectangles on the bins must have edges in common. Column graphs or bar charts are for data that take distinct values and the columns must be separated. The only reason for using 'boxes' in a column graph or bar chart is that single dots giving the frequencies at each value are too hard to see!

Bin: Interval of a histogram

As you will see below, the good aspect of histograms is easily seeing how the data are distributed. The bad aspect that you must remember is that the appearance of histograms for the same data can change a lot because we can choose both the starting point and the number of intervals. This flexibility also has its downside!

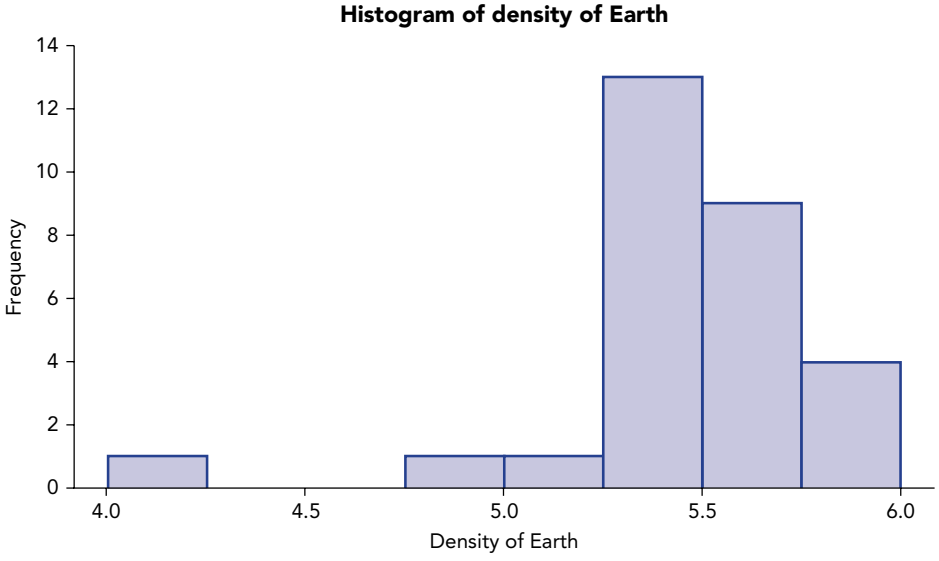
LET'S START A histogram of the famous Cavendish data

In 1798, Henry Cavendish estimated the density of Earth by using a torsion balance. The 'Cavendish' dataset contains his 29 measurements of the density of Earth, presented as a multiple of the density of water. Here are the data.

5.50	5.57	5.42	5.61	5.53	5.47	4.88	5.62	5.63	4.07
5.29	5.34	5.26	5.44	5.46	5.55	5.34	5.30	5.36	5.79
5.75	5.29	5.10	5.86	5.58	5.27	5.85	5.65	5.39	

Question: How do we draw a histogram?

The smallest value is 4.07 and the largest is 5.86, so 4 to 6 will cover all the values. How many intervals should we divide this into? This is never an easy or straightforward decision, as too few will clump the data too much, and too many will spread it too thinly. Let's see what starting at 4 and going to 6 with 8 bins will look like. The bins will be of length 0.25, and the histogram is given below.

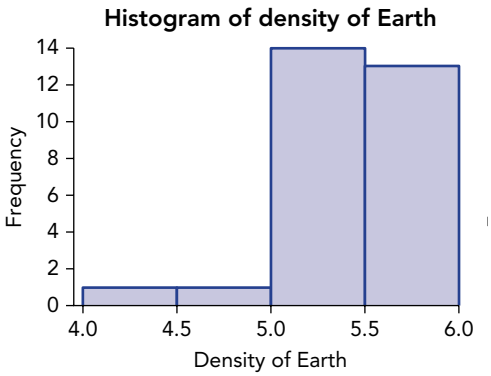


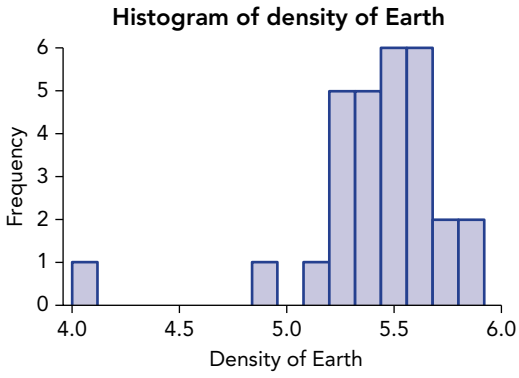
ISBN 978-1-107-65599-7

What happens to the observations that are exactly on a boundary? There is one here: 5.50. The convention is that it is put in the bin on the right of the boundary. If you count up the number of observations between 5.25 and 5.50, you'll see that there are 13, so 5.50 has been put into the next bin on the right.

Let's see what happens if we have 4 intervals or if we have 16 intervals. Because these data are fairly evenly spread between 5.25 and approximately 5.6, the two histograms below do not distort the picture of the data too much. Which do you prefer?

CAUTION
Generally there is no 'best' number of bins because how well a histogram 'pictures' the data depends on the range and spread of the data.





Key ideas

- A histogram is another type of plot for data from a continuous variable.
- The bins of a histogram divide a total interval covering all the data values into intervals of equal length.
- The heights of the rectangles standing on the bins give the frequencies of observations in the bins.
- Not enough bins can clump the data so much that it's difficult to see its features. Too many bins can make it look ragged – like a broken comb – and also hard to see its features.

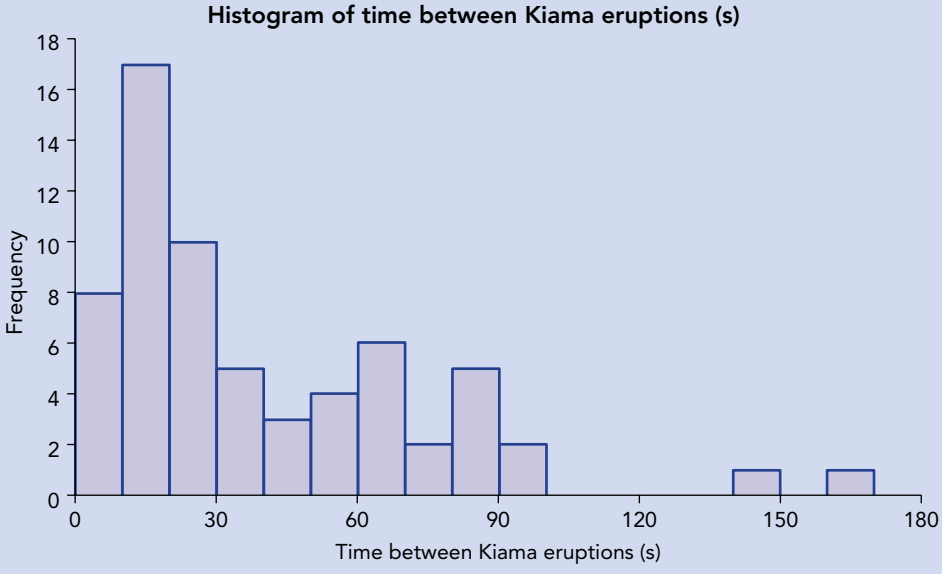
ISBN 978-1-107-65599-7

Example 2: How variable is the Kiama blowhole?

A hole in the cliff at Kiama, about 120 kilometres south of Sydney, NSW, is known as the 'blowhole' because the waves breaking on the cliff cause the water to blow up through the hole. The time between eruptions can vary a lot and depend on the combination of tide and winds and sizes of waves. An investigator collected 64 observations of times (in seconds) between eruptions. The data and a histogram with intervals of 10 seconds are below.

83	51	87	60	28	95	8	27	15
10	18	16	29	54	91	8	17	55
10	35	47	77	36	17	21	36	18
40	10	7	34	27	28	56	8	25
68	146	89	18	73	69	9	37	10
82	29	8	60	61	61	18	169	
25	8	26	11	83	11	42	17	
14	9	12						

Although the histogram below has a few 'jagged' parts, the picture is smooth enough to see that there are a lot of eruptions reasonably close together – up to 30 seconds apart – and then another group approximately 60 to 90 seconds apart, with the possibility of some being over 2 minutes apart.



ISBN 978-1-107-65599-7

Example 3: How different can histograms of the same data look?

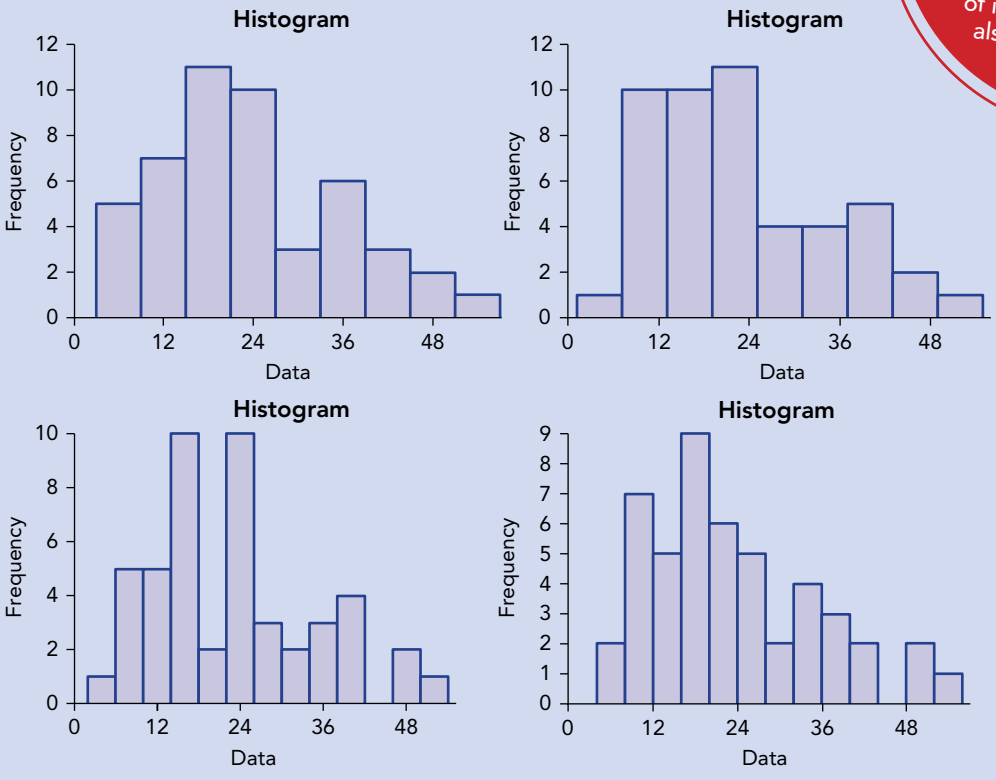
For the same dataset, the appearance of a histogram depends on where we start and on the number of bins. For very large datasets, different histograms will not look too different but for smaller datasets, the pictures provided by histograms can look very different. Below is a simple dataset of 48 observations arranged from smallest to largest for convenience.

4	7	8	8	8	9	11	11	11	12	12	14	15	15	16	16	16	16
16	17	17	19	19	22	24	23	23	23	23	23	24	24	25	27	28	29
33	33	34	35	37	38	39	40	40	48	48	53						

CAUTION
Remember that various histograms of a dataset can look quite different unless the dataset is very large. Dotplots, stem-and-leaf plots and histograms are all useful pictures of quantitative data but be careful not to interpret them as if they are fixed – they *all* involve choice of intervals. Histograms also involve choice of starting points.

Question: How different can histograms of these data look?

Below are four histograms of these data.



The first two have 9 bins each but slightly different starting points. The second two have 13 bins each and slightly different starting points.

Exercise 1B

1 Draw another histogram of the time between eruptions for the Kiama blowhole, again with bins of length 10 s, but starting at 5 s.

- **a** What is a main difference, if any, between this histogram and the one in Example 2?
- **b** From this histogram, estimate the probability that the time between eruptions is at least 15 s but no more than 25 s.
- **c** From the histogram in Example 2, can you estimate the probability that the time between eruptions is between 15 and 25 s?
- **d** From the histogram in Example 2, estimate the probability that the time between eruptions is at least 10 s but no more than 20 s.
- **e** Use the original data to estimate the probabilities for parts **b** and **d** above. Are they the same as in parts **b** and **d**? Say why or why not.

2 Draw two histograms as described below of the dataset in section 1-1 of the distance guesses of 5 m by the males.

- **a** For one histogram, choose bins of length 0.4 m, starting at 3.8 m.
- **b** For the other histogram, choose bins of length 0.3 m, starting at 3.9 m.
- **c** What are the main differences, if any, between the two histograms?
- **d** Using one of the histograms, estimate the probability that a male guess is at least 4.2 m but no more than 5.4 m. Why will this be the same for the other?
- **e** From the original data, estimate the probability that a male guess is at least 4.2 m but no more than 5.4 m. How is this different from the estimate in part **d**? Why?

3 In checking the suitability of a list of words to be used to check hearing, 30 people with normal hearing were each tested in a situation with moderate background noise. Words were read to the subjects who then said back the word they thought they heard. The list consisted of 25 words. Below are the scaled scores out of 50 for the 30 people.

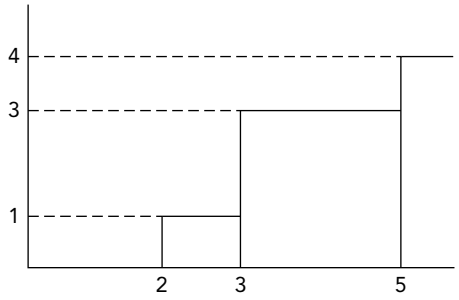
28 24 32 30 34 30 36 32 48 32 32 38 32 40
28 48 34 28 40 18 20 26 36 40 20 16 38 20
34 30

- **a** Draw a histogram of these data, identifying your choice of bin size and starting point.
- **b** What type of data are these?
- **c** Draw a column graph/bar chart of these data.
- **d** Why can we draw either a histogram or a column graph/bar chart for these data?
- **e** Which graph do you prefer for these data? Why?

ISBN 978-1-107-65599-7

4 A **cumulative histogram** is another type of histogram but it does not change – it does not depend on choices such as bin size and starting point. It adds up the number of observations we have accumulated as we move along the x-axis. So it draws a 'staircase' that takes a step up every time we come to another observation. If two observations are equal, it takes two steps up when we come to that value. For example, if the first few values in a dataset are 2, 3, 3, 5, then the cumulative histogram would start like this.

Cumulative histogram: A graph of quantitative data that gives the number of observations less than or equal to the values on the horizontal axis … *see glossary*



a Order the data in question 3 (the scaled scores out of 50) from smallest to largest. Then draw a cumulative histogram by drawing a staircase that takes its first step at the smallest value, then continues to step up at each value in the data. If there are two observations the same, then the staircase steps up by 2.

b Another picture that is used for a cumulative histogram marks the steps up with a dot and then joins the dots up. So the joining up goes through the edges of the staircase and there are no vertical or horizontal lines. This is sometimes called a **cumulative frequency plot**. Use the same data to draw a cumulative frequency plot.

Cumulative frequency plot: Joins up the tops of the steps in a cumulative histogram

Enrichment

Do Indie songs tend to be longer than Alternative rock songs?

5 Use the data given in question 4 of Exercise 1A.

a Draw histograms of the data for the two types of songs, using the same scale. A suggestion is to use bin size 50 s and start at 100 s.

b Compare your histograms with the back-to-back stem-and-leaf plots you produced in question 4 of Exercise 1A. Which do you prefer for these data? Give one advantage and one disadvantage of the histograms compared with the stem-and-leaf plots.

c The heights of the rectangles in the histograms in part **a** are the frequencies of observations in the bin. The bin size is the width of the rectangle and the bin sizes are equal. If the heights of the rectangles are changed to be the relative frequencies, what would change in the histogram? What would stay the same?

d If we wanted to draw the histograms so that the total area of the rectangles is equal to 1, what would the heights of the rectangles need to be?

e Histograms are almost always drawn with equal bin sizes. However, occasionally it is decided to have unequal bin sizes. Usually this is when one bin collects together a small number of the largest (or smallest) observations that are spread rather widely. Look at the histogram for the Alternative rock songs. Suppose we collect together in one bin all the

ISBN 978-1-107-65599-7

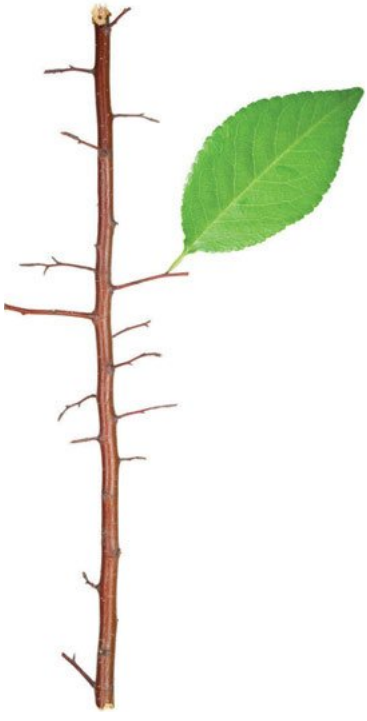
songs of length at least 300 s. Then the last bin will be from 300 s to 500 s, so its width is 200 s. If we want the total area of the rectangles in this histogram to be equal to 1, what will the height of the last rectangle (the one with base from 300 s to 500 s) need to be? How is this different from the heights of the other rectangles?

f How would you reply to the music fan who claims that Alternative rock songs tend to be longer than Indie songs?

ISBN 978-1-107-65599-7

1-3 Comparing plots

We are seeing more and more examples where it is desired to compare quantitative data across groups. And we are also seeing that the comparison is not necessarily simple or simple to describe. We now have three types of plots we can use for quantitative data – dotplots, stem-and-leaf plots and histograms. All have advantages and disadvantages and which type gives a good picture of the data can depend on the data.



The three types of plots

Dotplots tend to have the least collecting of data and are often useful in a first look at the raw data. Because they do little collecting of the data, they tend to be very bumpy. Stem-and-leaf plots collect data into intervals that divide the stem digits into 1, 2, 5 or 10 intervals. They are useful in giving a reasonable picture of how the data are distributed as well as retaining numerical information. However, the digits can be distracting. Histograms collect data into bins whose width and positioning can be chosen. They give the best overall picture of how the data are distributed, but different choices of bins can give a variety of appearances unless the dataset is very large – so caution is required. All three plots present the frequencies, but the histogram can be adapted to present relative frequencies.

Same scale

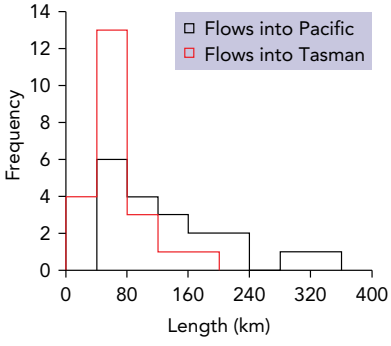
To compare across groups using any of these plots, the same scale *must* be used. Dotplots are readily placed above each other for comparisons, but stem-and-leaf plots need to be placed beside each other. Back-to-back stem-and-leaf plots can be used to compare two groups (e.g. males and females) but comparing more than three groups using stem-and-leaf plots is difficult. Histograms on the same scale can be placed beside each other or above each other, but they are fairly bulky pictures. There is a way to put two histograms on the same plot as we see below, but, as for stem-and-leaf plots, this is only practical for comparing two groups.

LET'S START Combining histograms

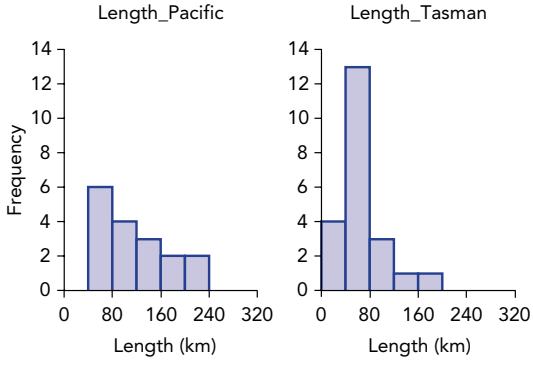
Histograms on the same scale (both axes) can be placed beside each other or above each other and two histograms can be placed on top of each other on the same plot. To prevent some rectangles hiding each other, the rectangles can be just outlined or can be shaded differently. Below, the lengths of New Zealand South Island rivers flowing into the Tasman Sea or the Pacific Ocean are graphed using histograms on the one graph and side by side on the same scale.

CAUTION
For histograms, remember to use the same scale on the *y*-axis as well as the *x*-axis.

Histogram of length



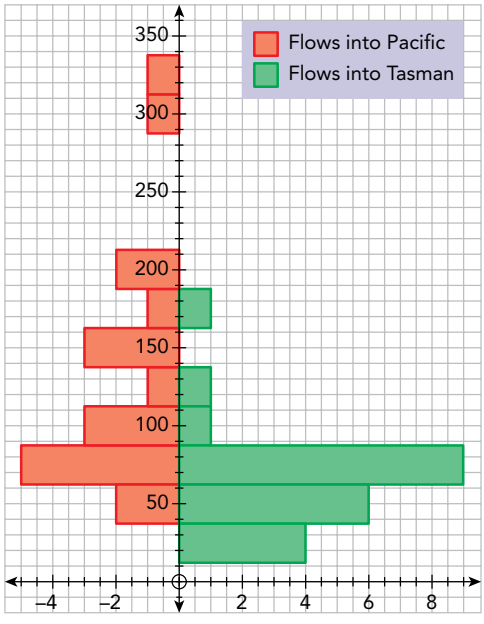
Histograms of length–Pacific, length–Tasman



Histograms can also be placed side by side like back-to-back stem-and-leaf plots.

From all the plots we see that there are quite a number of rivers of similar lengths flowing both ways, but generally the rivers flowing into the Tasman Sea tend to be shorter than those flowing into the Pacific Ocean. All the rivers flowing into the Pacific Ocean are more than 40 kilometres long, but the most commonly occurring lengths are between 40 and 80 kilometres, then 80 and 120 kilometres, and then there are fewer and fewer rivers as the lengths increase.

Which plot do you think shows the data best?



ISBN 978-1-107-65599-7

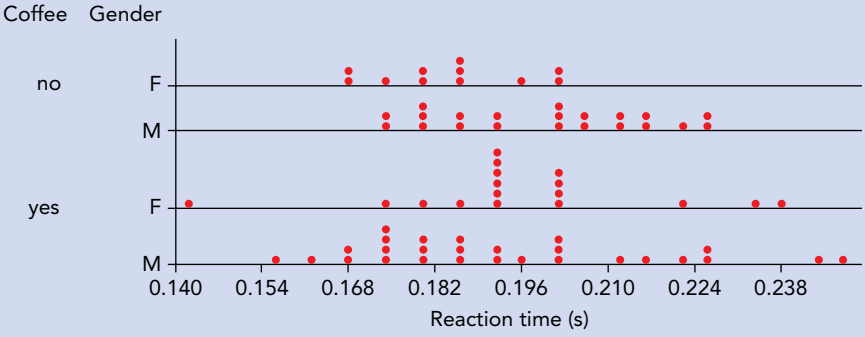
Key ideas

- Dotplots, stem-and-leaf plots and histograms:
  - can all be used to present quantitative data
  - all have advantages and disadvantages.
- The same scale must be used in plots that are comparing quantitative data across groups.
- There is no plot that is best for all datasets; choosing which type of plot to use may require exploration.

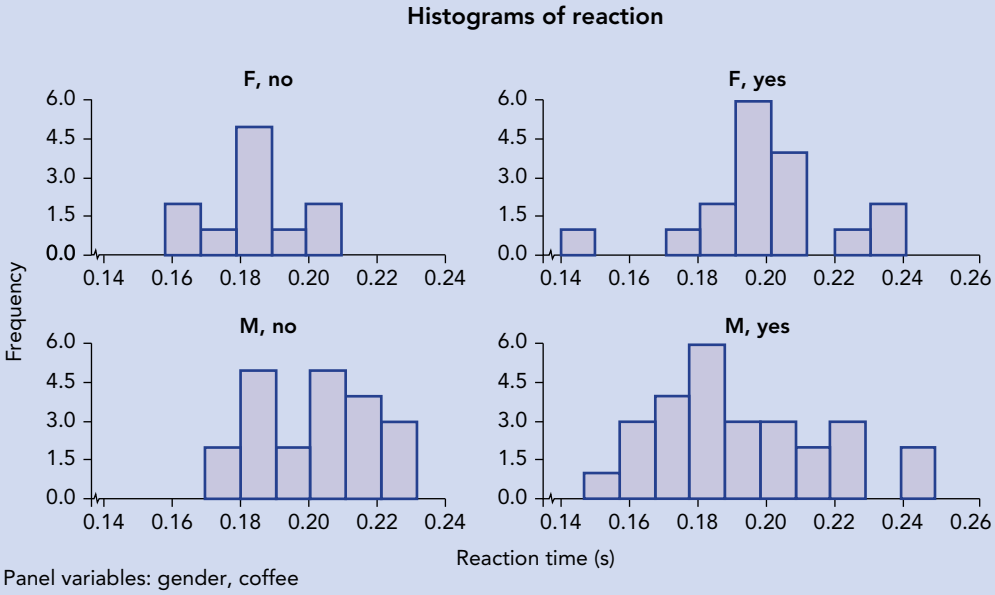
Example 4: Do males react faster than females? Does coffee affect reactions?

An experiment to measure speed of reaction was carried out on a random selection of males and females aged at least 17 years. Reactions were measured by catching a ruler dropped from a given height, and the measurements were converted to time in seconds. Subjects were also asked if they had drunk coffee in the previous hour.

The data file called *Reactions* is available on Cambridge GO. Below are dotplots and histograms on the same scale of the reaction times, split by both gender and whether they had drunk coffee or not.



ISBN 978-1-107-65599-7



We see that there's not much difference between the reactions of the males and females, except that drinking coffee seems to have the opposite effect on males to females! The reactions of the females who had drunk coffee were generally slightly longer, while those of the males who had drunk coffee were generally slightly faster. Notice that the variation of reaction times tends to be greater for those who had drunk coffee, but there were more people in the group who had drunk coffee, and generally the more people involved in an investigation, the greater the variation.

Exercise 1C

1 Refer to the data on guesses of 5 m distance by males and females in section 1-1.

- **a** Draw histograms of these data for males and females using the same scale.
- **b** Use the plots in section 1-1 and the histograms to comment on the comparison between male and female guesses.
- **c** Which plots do you think give the best presentation of these two datasets? Why?

2 Refer to the data in Example 1 on coral density close to and away from the coastline.

- **a** Draw histograms of these data using the same scale.
- **b** Use the plots in Example 1 and the histograms to comment on the comparison of coral density in the two groups.
- **c** Which plots do you think give the best presentation of these two datasets? Why?

ISBN 978-1-107-65599-7

3 Do motorists speed up if the traffic lights are amber? In a pilot study to investigate this question, an intersection with traffic lights was chosen where there were no right-hand turns or other possible obstacles to traffic flow. At a time and a day with light to moderate traffic, the speed (in km/h and correct to 0.1 km/h) at which motorists travelled the last 50 m before the lights was measured for both green and amber lights. The speeds for 50 observations for each colour of lights are given below. The dataset called *Traffic lights* is available on Cambridge GO.

Speeds amber lights

65.5	62.1	50.7	50.8	75.6	63.6	86.5	46.6	62.9	65.5	48.0	54.5
51.4	55.2	49.3	50.1	69.0	37.7	61.9	45.7	59.8	58.1	50.8	52.0
55.2	45.3	48.1	45.3	43.7	54.1	38.3	51.3	62.1	64.7	48.3	47.7
47.6	57.1	36.4	53.9	47.4	58.1	60.2	61.0	45.9	45.5	59.8	43.1
43.5	57.7										

Speeds green lights

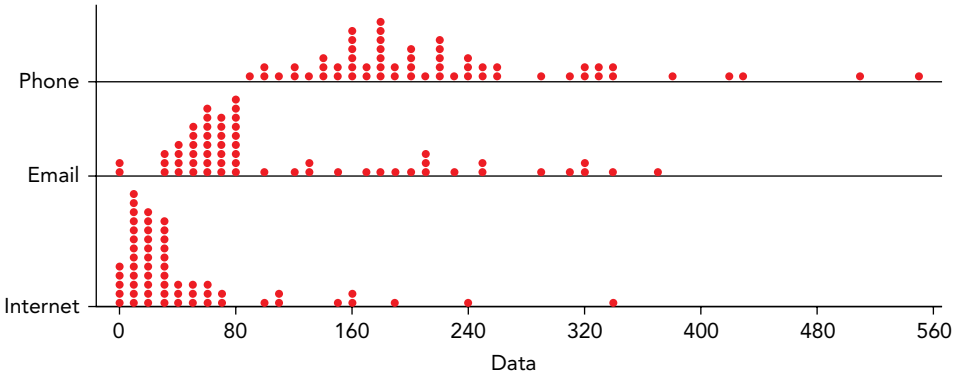
34.8	51.1	43.9	46.2	43.0	48.0	37.5	38.8	38.7	49.3	42.4	49.6
51.4	46.2	44.8	54.5	39.8	54.4	38.0	33.9	41.7	40.8	44.0	41.7
40.0	46.2	48.5	58.1	48.4	51.1	39.0	37.7	46.0	44.7	88.7	46.5
47.7	48.3	43.5	31.6	43.6	40.4	32.7	40.0	26.4	37.9	56.6	49.9
53.6	34.0										

a Choose a plot to compare the speeds of approach to amber and green lights.

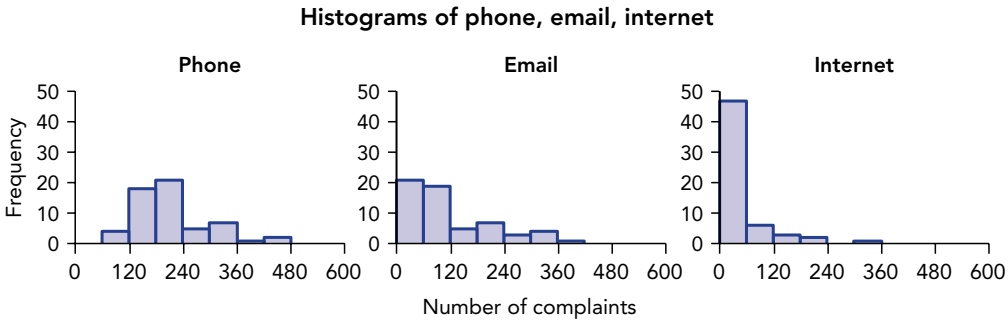
b The speed limit on this street was 60 km/h. Comment on the speeds.

c What are your comments on the question of whether motorists speed up when approaching amber lights?

4 The third runway at Sydney airport has been the cause of many complaints over the years about noise, despite a curfew imposed in 1995 by federal government legislation. Complaints about noise may be made by phone, email, letter, internet or callback. The numbers of complaints per month by these different ways were obtained from information provided by the Sydney airport over a 5-year period. Below are histograms and dotplots of the numbers of complaints per month made by phone, email and internet.



ISBN 978-1-107-65599-7



a Comment on the comparisons between the number of phone, email and internet complaints per month.

b For these data, which plots do you think make the comparison most clear? Why?

c There is a major difference between this (whole) dataset and others in this chapter. Think of the spreadsheet or data collection sheet for this investigation. What do the rows correspond to? Hence, can you identify why this dataset is different from others in this chapter where two groups are being compared?

Enrichment

How well do people estimate time?

5 An experiment aimed to discover how well people can estimate a given length of time, and how this varies over males and females. The study also considered two age groups, but in this question we just consider those aged between 20 and 40 years. Random groups of 30 males and 30 females in this age group were chosen, and were asked to guess 5 and 10 second intervals. The guessing was done by a subject saying 'stop' when he or she thought the time was reached. The guesses were measured in seconds accurate to 0.1 s. The following table gives the data.

ISBN 978-1-107-65599-7

5 s guess	10 s guess	Gender	5 s guess	10 s guess	Gender	5 s guess	10 s guess	Gender	5 s guess	10 s guess	Gender	5 s guess	10 s guess	Gender
3.6	6.7	female	2.0	9.2	female	3.8	8.3	female	6.0	9.4	male	2.9	10.7	male
5.1	9.5	female	6.2	12.6	female	3.6	7.8	female	4.6	10.1	male	3.8	11.7	male
4.2	5.7	female	4.4	8.2	female	3.6	8.3	female	4.0	9.5	male	5.1	9.7	male
4.3	7.3	female	3.0	5.9	female	5.6	8.1	female	4.2	9.7	male	4.8	9.3	male
4.0	8.2	female	1.9	6.3	female	5.4	10.6	female	5.2	12.1	male	4.4	9.4	male
4.4	9.4	female	4.9	8.7	female	5.0	8.2	female	5.2	9.4	male	4.7	9.1	male
1.9	5.5	female	4.2	9.3	female	4.5	11.1	female	5.1	10.2	male	3.6	8.7	male
4.9	9.2	female	6.0	9.4	female	5.3	10.4	female	4.7	11.4	male	3.5	8.9	male
3.8	11.7	female	5.4	9.4	female	4.6	10.1	female	3.7	8.1	male	2.5	9.2	male
3.3	7.7	female	5.1	10.2	female	5.2	9.4	female	4.3	9.4	male	5.3	10.2	male
4.6	8.0	male	2.7	8.8	male	3.4	7.7	male	4.8	8.9	male	5.7	9.6	male
4.1	8.1	male	4.4	9.9	male	3.1	9.9	male	4.4	10.0	male	6.0	10.1	male

a Choose a type of graph to plot the following guesses for males and females:

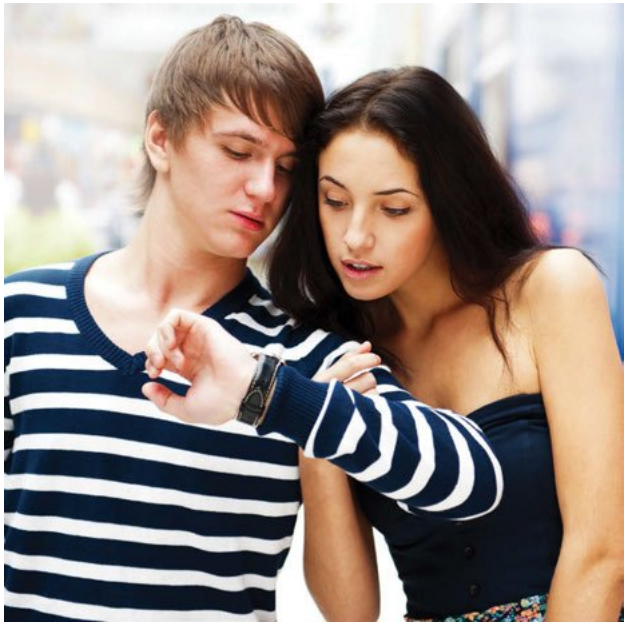
i 5 s **ii** 10 s.

b Did the subjects tend to underestimate or overestimate the time intervals?

c Comment on the comparisons between males and females in guessing 5 s and 10 s intervals.

d Using the plots drawn in part **a**, does the comparison between males and females look different for the two time intervals? In what way?

e Think of a way of comparing the guesses of males' and females for both 5 s and 10 s in one set of plots. Does it add anything to your comments in part **d** above?



ISBN 978-1-107-65599-7

Chapter summary

Investigating quantitative data across groups

- Comparing quantitative data across groups involves a continuous variable and a categorical variable
- The quantitative data must be collected in the same way and to the same accuracy across groups
- In using plots to compare quantitative data across groups, we must use the same scale
- Back-to-back stem-and-leaf plots can be used to compare quantitative data across two groups.

Histograms

- The bins of a histogram divide the data range into intervals of equal length
- The heights of the rectangles give the frequencies of observations in the bins
- Choice of bin size and starting point can change the appearance of a histogram.

Comparing plots

- Dotplots, stem-and-leaf plots and histograms are all plots for quantitative data
- The same scale must be used in comparing quantitative data across groups
- Plot choice may depend on data and may need exploration.

Multiple-choice questions

Questions 1–7 refer to the following situation.

A data recording sheet with 50 rows has one column headed 'Resting pulse rate' recorded as beats per 15 s, another headed 'Gender' and another headed 'Age group' with entries 1, 2 or 3 depending on whether the subject is under 25 years, between 25 and 55 years, or older than 55 years. The values of the resting pulse rates range from 15 to 35.

1 Bar charts or column graphs can be used to plot

A Resting pulse rate **B** Age group
C Both A and B **D** Neither A nor B

2 It is appropriate to use stem-and-leaf plots or histograms for resting pulse rate data because

A The values are numbers **B** They are quantitative with a number of different values
C They are data **D** All of these

3 A back-to-back stem-and-leaf plot is used to compare pulse rates across males and females irrespective of age. The digits in the leaves are

A The numbers of males or females with the pulse rate given in the stem **B** The numbers 1, 2 or 3 depending on age group
C The tens values of the pulse rates **D** The units values of the pulse rates

4 A back-to-back stem-and-leaf plot is used to compare pulse rates across males and females irrespective of age. The total number of digits in the leaves is

A The number of different pulse rates recorded **B** 25
C 50 **D** Unknown

ISBN 978-1-107-65599-7

5 Histograms are used to compare pulse rates across males and females. The bins of the two histograms must

A Start at the same value
B Be the same for the two histograms
C Both A and B
D Neither A nor B

6 The rectangles of the histograms in question 5 above

A Share sides
B Have heights equal to the pulse rates of the subjects in the corresponding bins
C Have equal areas
D All of these

7 Histograms are used to compare pulse rates across males and females and the different age groups. How many histograms are there?

A 2
B 3
C 5
D 6

Short-answer questions

1 In which of the Let's Start questions, examples and exercise questions in sections 1-1, 1-2 and 1-3 above were the data obtained by the investigators secondary data?

2 An experiment was carried out to investigate if a dose of vitamin C assists muscular endurance of young men. Each volunteer squeezed a dynamometer (an instrument that measures grip strength) three times before taking either a dose of vitamin C or a **placebo**. Their maximum initial grip strength was taken as the reference for that person. After the dose of vitamin C, each volunteer squeezed the dynamometer for 3 s, and repeated this until the instrument registered 50% of the volunteer's maximum initial strength. The measure of endurance was the number of repetitions until 50% of the initial strength was reached for the first time.

Placebo: A dummy treatment; a treatment in which no treatment is given but may pretend to be one; in health, placebos have no active ingredients

a What would need care in carrying out this experiment?

b Can we use a stem-and-leaf plot or histogram for the numbers of repetitions? Why or why not?

c What should we plot to investigate the research question?

d What must we make sure of in the plots in part **c** to investigate the research question?

3 An experiment to investigate if pain thresholds are different for blondes and brunettes was conducted. Below are the pain threshold scores of male and female volunteers in a sensitivity test – the higher the score, the higher the person's pain tolerance.

Brunette:	42	50	41	37	32	39	51	30	35	
Blonde:	62	60	71	55	48	63	57	52	41	43

ISBN 978-1-107-65599-7

a Draw dotplots (on the same scale) of these data.

b Draw a back-to-back stem-and-leaf plot to compare scores for blondes and brunettes.

c What do the plots indicate?

4 In the data on the numbers of complaints about Sydney airport noise made per month by phone in question 4 of Exercise 1C, it was also recorded whether each month included school holidays or not. The data are below.

No school holidays

157	553	146	215	184	161	164	104	202
217	159	146	87	181	165	189	205	236
158	100	136	334	234	235	260	199	248
179	424	512	307	335	317	220	215	

Includes school holidays

127	186	172	200	108	182	286	216	160
135	184	201	123	140	120	317	256	236
177	180	431	379	337	247	326		

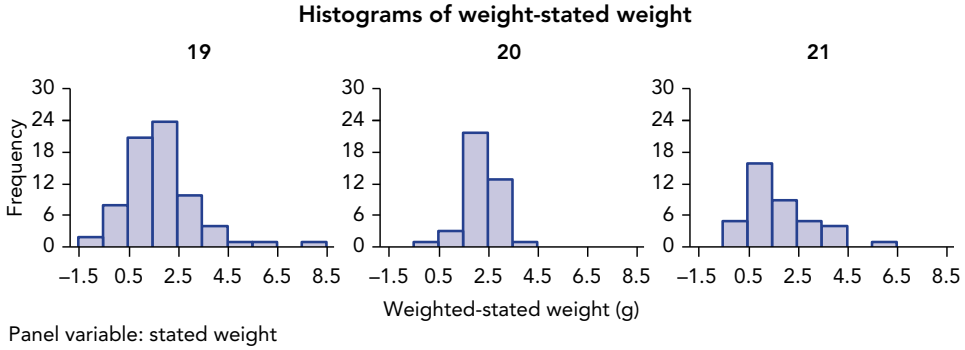
a Draw a back-to-back stem-and-leaf plot to compare the numbers of monthly phone complaints in months that do, or do not, include school holidays.

b Draw histograms on the same scale to compare the numbers of monthly phone complaints in months that do, or do not, include school holidays.

c What do the plots indicate?

d Which type of plot do you prefer for these data?

5 An investigation was carried out into the quality of small packets of crisps (chips). As part of the investigation the weights in grams of unopened packets of three brands of crisps were obtained. These weights included the weights of the (empty) packets. Each brand gave a different stated weight of crisps in the packet – 19, 20 and 21 grams. Below are histograms of the weights minus the stated weights for the three brands.



a Why are the data changed to weight-stated weight before plotting?

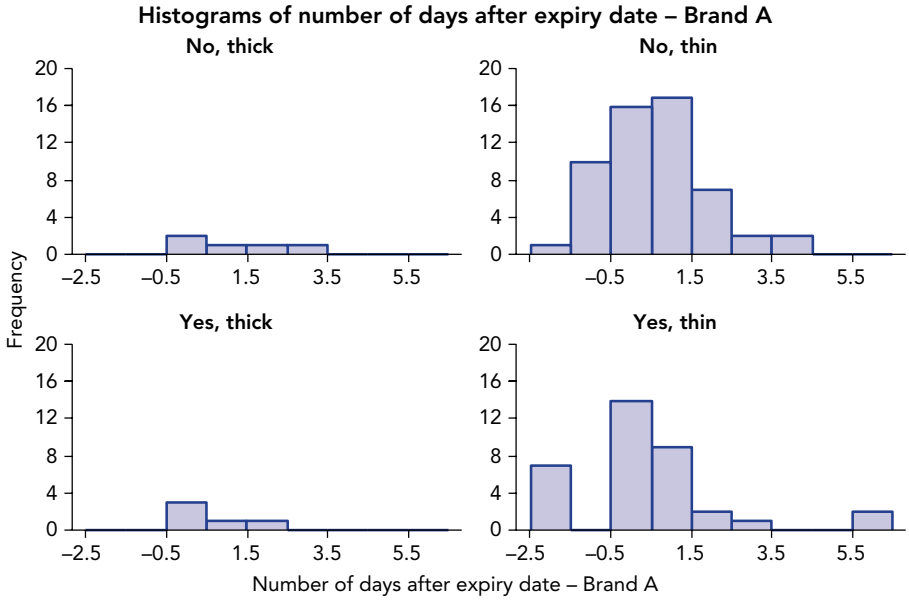
b Comment on the plots.

ISBN 978-1-107-65599-7

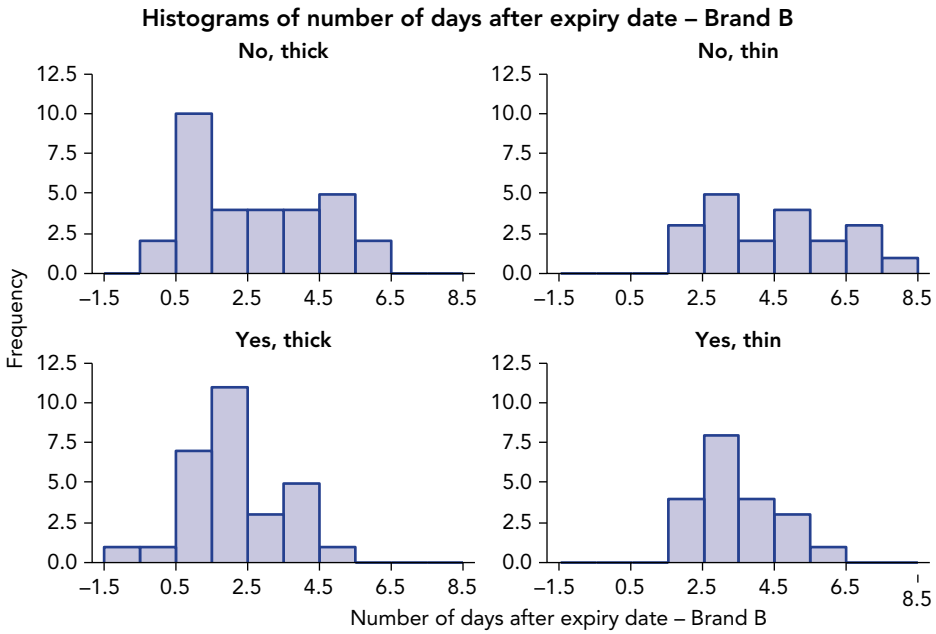
Extended-response question

6 An investigation was carried out into the extended storage life of sliced bread. The extended storage life was defined as the number of days after the expiry date on the packet before mould first appeared. In the data considered here, the results for white bread of two brands are presented. Thin and thick sliced breads were included in the investigation. Some of the breads were left in sunlight and some were not.

Below are histograms on the same scale for brand A and separately for brand B, split by thickness of the bread and whether it was left in sunlight or not.



ISBN 978-1-107-65599-7



Panel variables: Sunlight_Brand B, thick or thin_Brand B

a What do the negative values mean?

b Comment on the plots for brand A.

c Comment on the plots for brand B.

d What is one problem with the data for brand A?

e Compare brands A and B.

f What in the plots would make it easier to compare brands A and B?

g Why would doing the above comparisons be more difficult using stem-and-leaf plots?

ISBN 978-1-107-65599-7

CHAPTER

2

Quantitative data shapes

What you will learn

How much do pedestrian speeds vary?

How fast do people walk along a footpath or walkway when they are going about their normal business? How much variation is there? Is there a difference overall in walking speed between males and females? Does their type of footwear matter? For example, if the people using the walkway are mostly business people, they are unlikely to be wearing flip-flops. Does it make a difference to walking speed if people wear flip-flops?

An observational study was carried out on a section of city footpath that was used by many people as a thoroughfare – that is, it was mostly office buildings so that people were generally just walking past. Subjects were chosen 'randomly' and timed over a distance of approximately 15 metres between two set points. Subjects who exited or stopped between the two points were not included.

Information about walking speeds when people are going about their normal activities can help in designing and coordinating traffic and pedestrian lights. There would be particular interest in the variation. How much variation is there and what does it look like?

AUSTRALIAN CURRICULUM

Statistics and probability

- Data representation and interpretation
- Describe data, using terms including 'skewed', 'symmetric' and 'bimodal' **(second part of ACMSP282)**
- Compare data displays using mean, median and range to describe and interpret numerical datasets in terms of location (centre) and spread **(ACMSP283)**

ISBN 978-1-107-65599-7

PRE-TEST

Terms you will learn

asymmetric
bimodal
multimodal
skewed to the left
skewed to the right
sub-groups
symmetric
tails of the data
unimodal

1 Each of the following lists is part of recorded data in an investigation. What is a problem with the recording of the data in each case?

a The lengths of a sample of timber

170 162 164 154 146 144 164 17.2 15.9 16.6 16.6 15.0 15.8

b The heights in centimetres of a sample of school children

156.5 163.2 160 154.25 158 170 163.35

c The amount of protein in grams for different cereals

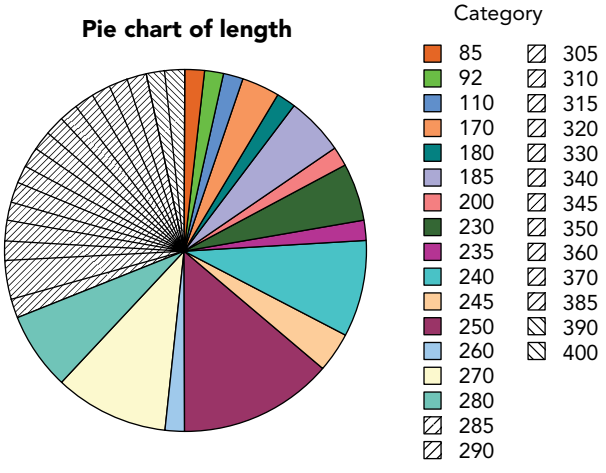
15.00 11.60 6.70 6.00 7.80 7.10 5.40 6.10 6.40 8.10

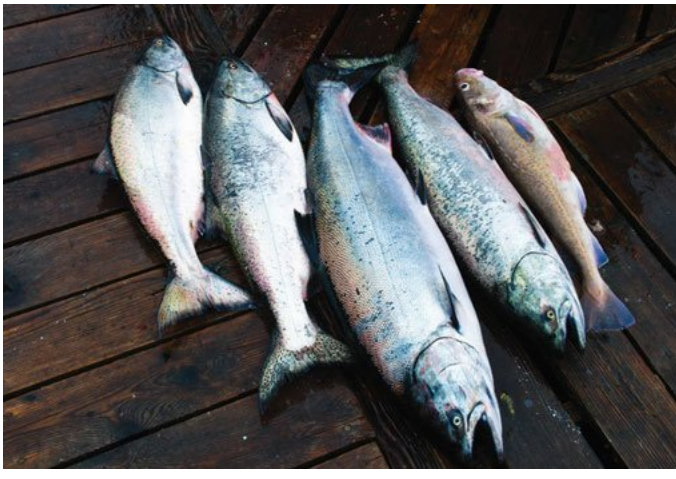
2 The differences between actual and scheduled times of departures of planes were recorded at three different Australian airports. The overall recording sheet of the data could be arranged to have three columns or two columns.

a What would be in the columns if it had two columns? What would the rows correspond to?

b There are two possible ways of recording the data using three columns. What are they and what would be in the columns in each case?

3 Below is a plot of the lengths of fish caught on a fishing trip.





a Why is this plot incorrect?

b How could the data be plotted?

ISBN 978-1-107-65599-7

4 In an experiment investigating how often people blink when they are inside or outside, the following data were recorded giving the number of blinks per minute.

Inside

13 8 24 26 15 25 19 42 53 34 35 16 34 13 27 22 23 38 32 25 19 17
23 39 25

Outside

14 38 17 11 17 16 11 37 40 16 21 23 19 15 23 33 23 17 19 30 22 24
33 23

a Draw a dotplot of the inside and outside data on the same scale.

b Draw a back-to-back stem-and-leaf plot.

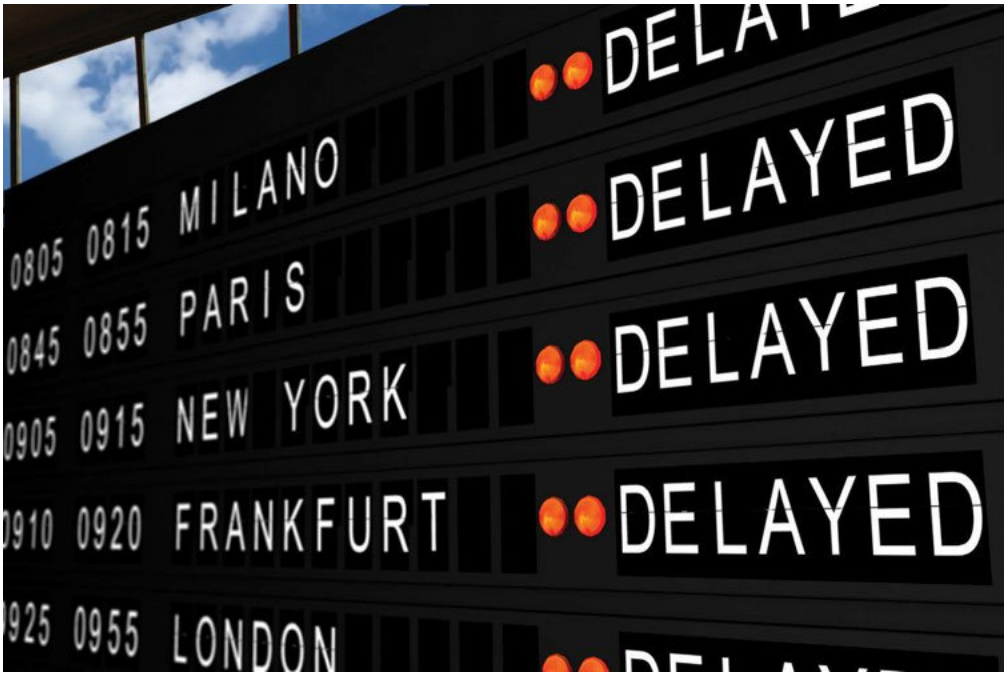
c Draw histograms of the inside and outside data on the same scale.

d Find the data mean, median and range for each group.

ISBN 978-1-107-65599-7

2-1 Skewness

Think about comparing arrival times of planes with their scheduled arrival times. We would want to know whether the actual arrivals tended to be around the scheduled times – that is, is the average or median arrival time close to the scheduled time? What would be of interest in describing the variation? We would want to know how much variation there is around the average, median or scheduled arrival time.



Skewed to the right: Refers to quantitative data in which the variation of the data for values greater than the 'middle' is more than for values less than the 'middle'

Skewed to the left: Refers to quantitative data in which the variation of the data for values greater than the 'middle' is less than for values less than the 'middle'

Asymmetric: Data that are not **symmetric**; some data are asymmetric but cannot be readily described as skewed to the left or to the right

But we would also want to know how the arrival times vary when the planes are early and when they are late. Perhaps the variation of the arrival times when the planes are late is much greater than the variation when the planes are early. This is comparing the variation on the right of the 'middle' of the data with the variation on the left of the 'middle' – the 'middle' might be taken as the mean or the median.

If the variation of the data for values greater than the 'middle' is more than for values less than the 'middle', we say the data are **skewed to the right**. If the variation of the data for values greater than the 'middle' is less than for values less than the 'middle', we say the data are **skewed to the left**. If the data are not symmetric, we say they are **asymmetric**. Sometimes data are not symmetric but cannot be said to be skewed to the right or to the left. In that case, we just say they are asymmetric.

What is meant by no skewness? If the shape of the data for values greater than the 'middle' is about the same as for values less than

HINT
Notice that 'no skewness' requires the shape of the data to be (close to) the same on either side of the middle. In this case the data mean and median will be (close to) equal and give the middle of the data.

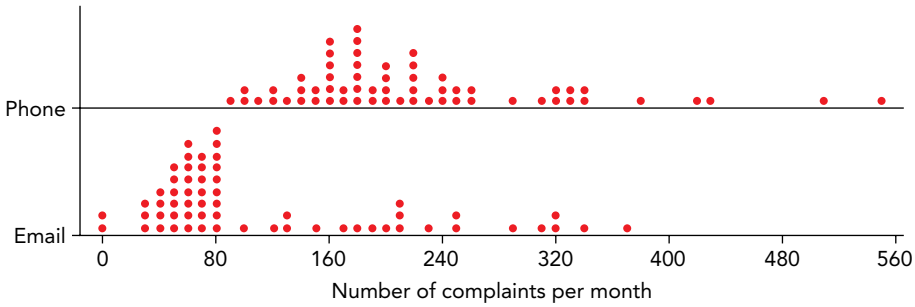
ISBN 978-1-107-65599-7

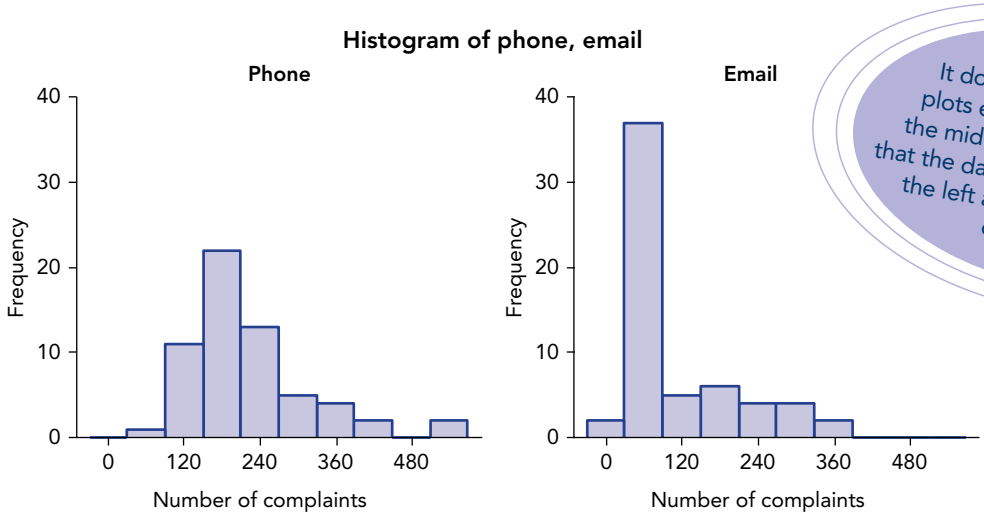
the 'middle', we say the data are (approximately) **symmetric**. This is almost always approximate because we have to allow for sampling variation. For example, perhaps the variation in the time to dissolve a (soluble) tablet looks about the same on either side of the average time to dissolve.

Symmetric: Refers to quantitative data in which the variation of the data for values greater than the 'middle' is the same as for values less than the 'middle' … *see glossary*

LET'S START How skewed is skew?

In question 4 of Exercise 1C, the number of complaints about noise per month made in different ways to Sydney airport over a number of years was considered. Below are dotplots and histograms of the number of monthly complaints made by phone and by email.





HINT
It doesn't matter in these plots exactly where we take the middle to be – we can see that the data are more clumped on the left and more spread out on the right.

Notice how the numbers of email complaints are much more clumped for the lower values and much more spread out as we move to the right – to the larger values. The numbers of email complaints are very skewed to the right. The numbers of phone complaints are also skewed to the right, but the contrast between right and left values is not quite so dramatic.

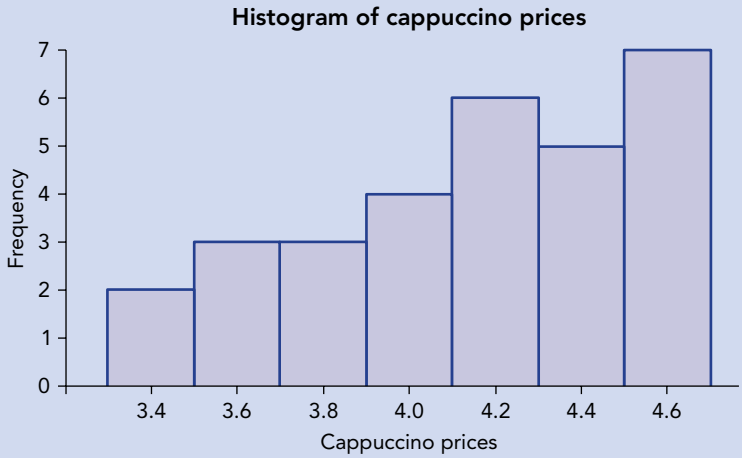
ISBN 978-1-107-65599-7

Key ideas

- When data are more spread out over the larger values than the smaller values, we say the data are skewed to the right. When data are more spread out over the smaller values than the larger values, we say the data are skewed to the left.
- When the data mean and median are (almost) equal, and the data have (close to) the same shape on each side of this 'middle' value, we say the data are (close to) symmetric.
- Data that are not symmetric are said to be asymmetric. Some data are asymmetric but cannot be readily described as skewed to the left or to the right.

Example 1: How much for a cappuccino?

The prices of cappuccinos were recorded in cafes chosen randomly in a city and are plotted below in a histogram.



CAUTION
Note that the midpoints of the bins are marked in this histogram. Marking either the midpoints or the edge points of the bins is acceptable as the others can always be obtained by looking at the plot. For example, the edge points in this histogram are 3.3, 3.5, 3.7 and so on.

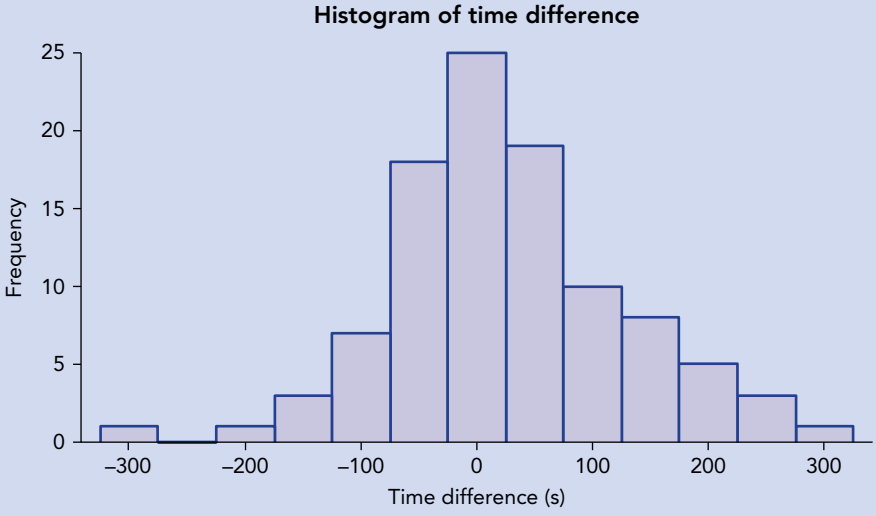
HINT
Real data are more often skewed to the right than to the left as you will see. Can you think of a reason why these data are skewed to the left?

Question: Are the data symmetric, asymmetric, skewed to the left or skewed to the right?

We know we must allow for differences in histogram pictures of data depending on the choices of starting point and bin size. However, these data are definitely skewed to the left, and it is unlikely that any change in the choice of starting point and bin size would change this.

Example 2: How much do people's watches vary from the correct time?

A survey was carried out of people randomly chosen in a city mall to investigate how close or not people's watches were to the correct time. Some people deliberately keep their watches slightly fast (or slow!) and this was allowed for. The following plot is a histogram of the difference in seconds = time on watch – actual time.



Question: Are the data symmetric, asymmetric, skewed to the left or skewed to the right?

As in Example 1, a different choice of starting point and bin size might produce a different impression of the data, but in the above histogram, the data look close to symmetric. Also what we are really interested in, is how close people's watches in general are to the correct time. Therefore we are looking at this as a sample of the general population randomly represented by the sample with respect to their watch time. So we also need to allow for sampling variation when we look at pictures of the data. These reasons are why we usually only say that the data are 'close to symmetric'.

HINT
In these examples, the skewness to the right or to the left is fairly strong. In many real datasets, skewness is often not as clear. So often we might say that a plot or graph indicates that the data are skewed to the right or to the left, or are asymmetric.

Exercise 2A

1 Consider question 6 of Chapter 1 Pre-test on the time between the arrival of phone calls to an office. Are the data skewed? If so, are they skewed to the right or to the left? How strongly are they skewed?

2 Consider Example 1 of Chapter 1 on the coral densities of reefs. Describe each of the two groups of observations in terms of their symmetry, asymmetry or skewness.

ISBN 978-1-107-65599-7

3 Consider Example 2 in Chapter 1 on the time between eruptions of the Kiama blowhole.

- **a** Are the data skewed?
- **b** If so, are they skewed to the right or to the left?
- **c** Do you think your answers to parts **a** and **b** would change if different choices were made for the starting point and bin size in the histogram?

4 Consider question 3 of Exercise 1A and Let's start of section 1-3 on the lengths of New Zealand South Island rivers.

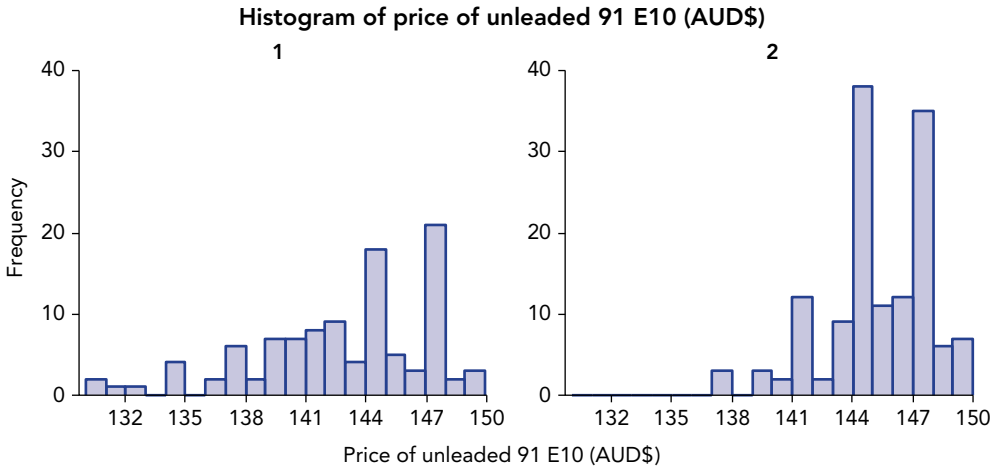
- **a** Are the data skewed for both groups of rivers?
- **b** If so, are they skewed to the left or the right?
- **c** Compare the skewness of the data for lengths of rivers flowing into the Pacific Ocean with those flowing into the Tasman Sea.

5 Refer to the data on the guesses of males and females of 5 metres length in section 1-1 and question 1 of Exercise 1C. Describe the skewness, or otherwise, in the data in each group.

Enrichment

How much do fuel prices vary?

6 Data were collected on the price of unleaded 91 E10 fuel from a number of randomly chosen service stations each day for 5 weeks in an Australian city. The histograms below are of the prices in two groups: 1 = Monday–Wednesday, 2 = Thursday–Sunday.



- **a** Describe each group of data in terms of skewness, symmetry or asymmetry.
- **b** Compare the two groups with respect to your answers to part **a**.
- **c** Why do you think the histograms of these data are particularly 'bumpy'?

ISBN 978-1-107-65599-7

2-2 Sub-groups and bimodalities

Categorical data

You have seen in categorical data that the category with the highest frequency (or relative frequency) is valuable information as it gives the most commonly occurring category. This is called the mode. Similarly the mode of count data where there are not too many different counts – such as number of children in a family – is also informative. If ordinal (or count) data have two categories (or count values) that have greater frequencies than neighbouring categories (or count values), the data are said to have two modes, or be **bimodal**. If there is only one category (or count value) with greater frequency than its neighbours, the data are said to be **unimodal**.

CAUTION
For nominal data – categorical data with no natural ordering of categories – the categories can be arranged in any order we like. So it doesn't make sense to talk of neighbouring categories.

Quantitative data

You have also seen in Years 7–8 that the concept of a mode is difficult and can be quite misleading for data from a continuous variable. If the data are not collected into intervals, then all the observed values may occur once only or there may be many observed values, each occuring a small number of times. This can depend on the situation and the measurement accuracy. When the data are collected into intervals for plots, the commonly occurring intervals can be changed – sometimes very much – by the choice of intervals.

Sometimes stem-and-leaf plots and histograms have two regions that have higher frequencies than neighbouring ones even for different choices of intervals. This means there may be two 'clumps' of data. If it is clear that there are two 'clumps' of data, it may be appropriate to say that the data look bimodal. This sometimes suggests there are two groups in the data that may be different. But we must be careful in using the word 'bimodal' for quantitative data unless we have a very large dataset.

Bimodal: Ordinal data (or ungrouped count data) with two categories (or count values) with higher frequencies than their neighbours … *see glossary*

Unimodal: Ordinal data (or ungrouped count data) with only one category (or count value) with greater frequency than its neighbours … *see glossary*

CAUTION
Remember that for any data, modes are not necessarily measures of centre. For data from continuous variables, modes are difficult to define and commonly occurring intervals can be quite different for different plots.

Sub-groups

So for all types of data, if clumps of data occur in distinct places, this can be important information. For example, this might indicate that there are **sub-groups** in the data that are different. Such clumps could provide clues to investigators to look for sub-groups or another variable or variables that are affecting the data.

If there are clearly more than two clumps of data, we sometimes say that the data look **multimodal**. If the data tend to be most clumped in one region, we sometimes say the data look unimodal.

Sub-groups: Groups of observations within data which are different and may refer to different categories of a categorical variable

Multimodal: When quantitative data have more than two clearly indicated clumps of data, this may indicate that there are different **sub-groups** within the data.

LET'S START How loud is your commuting environment?

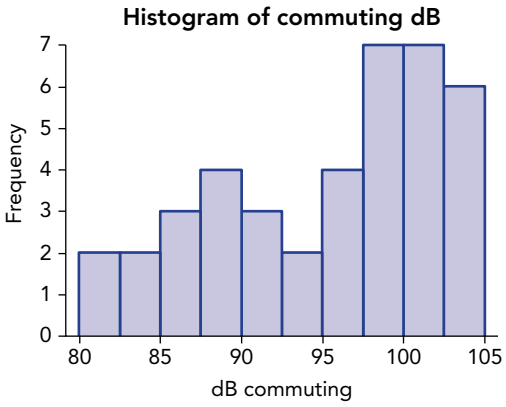
A study into the noise levels of different environments recorded the sound level in decibels (dB) at a variety of randomly chosen times in places that could be generally classified as weekend home, recreation or commuting. In some cases, the sources of sound could be identified but in others there were mixed sources.

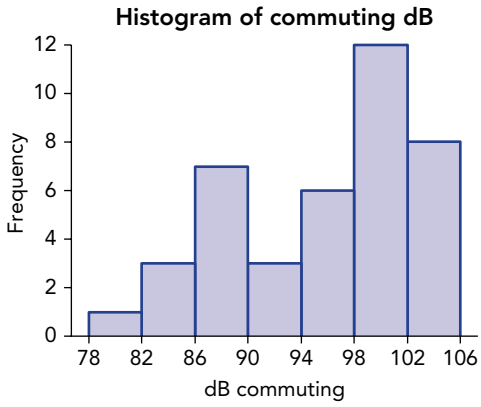
CAUTION
We need to be careful in looking for clumps of data in histograms and stem-and-leaf plots. But if the clumps are sufficiently large and separated, they will show in these plots with sensible interval choices that smooth the data without collecting too much together.

On the next page are two histograms of the decibels recorded in the commuting environment. The first has 10 bins and the second has 7 bins. In both histograms, we can see two 'clumps' of observations. The one with the most observations is centred somewhere around 100 dB, while the other, with fewer observations, is centred somewhere between 85 dB and 90 dB. The gap between the two groups seems to be approximately between 90 dB and 94 dB.

If we just had the first histogram, we might think there was some indication of two groups or bimodality. The benefit of seeing the second histogram as well is that we can be more confident that there are two groups. What we can see in the first histogram is not just because of a particular choice of intervals.

There's no information in these plots as to what might be causing the two groups or bimodality, but there may be information in the full dataset.





CAUTION

Because stem-and-leaf plots, histograms and dotplots are seldom smooth, and are often quite 'bumpy', you can see that very great care is needed before calling data 'multimodal'.

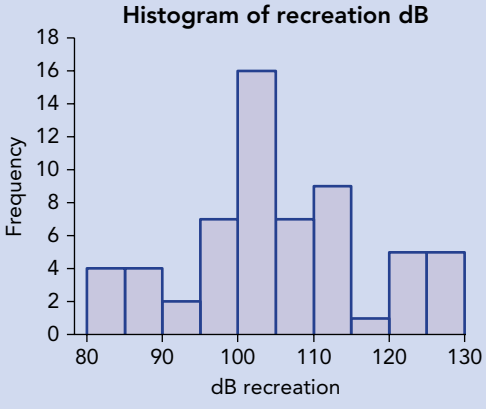
Key ideas

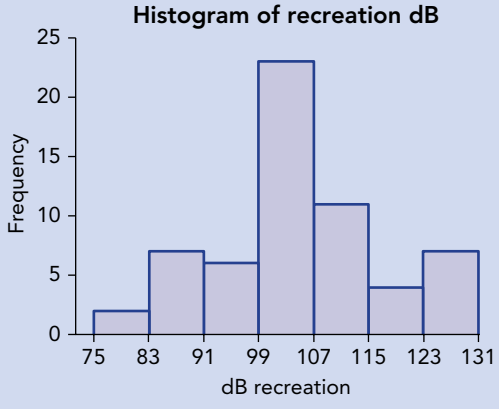
- Ordinal data (or count data that are not collected into intervals) may have two categories (or count values) with higher frequencies than their neighbours. The data may be said to be bimodal.
- Clumps of data in plots of quantitative data may indicate that there are different sub-groups within the data. If plots clearly indicate two clumps, we say the data tend to be bimodal. If plots clearly indicate more than two clumps, we say the data tend to be multimodal.
- Stem-and-leaf plots and histograms collect quantitative data into intervals. Because the appearance of the plots can depend on the choice of intervals, we need to be careful in saying that a plot looks bimodal.

Example 3: How loud is your recreational environment?

Below are two histograms of the decibels of sound recorded in various recreational environments.

Question: Do the data on recreational sounds in dB look bimodal or multimodal?





ISBN 978-1-107-65599-7

Looking at the first histogram, we might wonder if there are small sub-groups in the **tails of the data**. However, the second histogram with fewer bins tends to indicate that there are no sub-groups. The data have a very large range, from 80 dB to 130 dB, and there may be lots of different sources contributing to the variation, but we couldn't say that these data are bimodal or multimodal.

Tails of data: Refers to spreading out of data for the smaller or larger data values. A data tail has low frequencies spread over a range of small (left tail) or large (right tail) data values

CAUTION
Care is needed in looking at the tails of data as there can be a small number of observations there and different interval choices may affect them more than where there are more observations.

Exercise 2B

1 Refer to the data on the guesses of males and females of 5 metres length in section 1-1 and question 1 of Exercise 1C.

- **a** Do you think the data indicate bimodality in the females' guesses?
- **b** Do you think the data indicate bimodality in the males' guesses?
- **c** If you answered yes to part **a** or part **b**, do you think there may be a reason for the bimodality or is it just due to variation in a small sample?

2 Refer to the data of Example 1 in Chapter 1 on the coral densities of reefs.

- **a** If the data in these two groups were combined and plotted using a stem-and-leaf plot or histogram, do you think the plot would indicate bimodality?
- **b** Explain your answer to part **a** and say where you think lumps of data might occur in the plot.

ISBN 978-1-107-65599-7

3 Consider Example 3 of Chapter 1.

a Which of the histograms do you think might mislead people most in thinking that the data are bimodal or even multimodal?

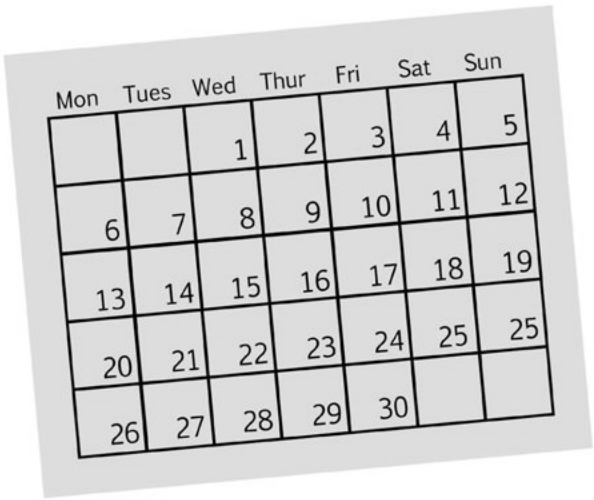
b Why is it important to know there are only 48 observations when looking at these histograms?

4 Refer to question 3 of Exercise 1A and Let's start of section 1-3 on the lengths of New Zealand South Island rivers.

a If all the rivers were considered together and their lengths plotted, the plots would not indicate bimodality. Why?

b Do you think the combined dataset would be skewed? If so, would they be skewed to the right or the left? If not, would they be symmetric or asymmetric?

5 In question 6 of Exercise 2A, there are two distinct commonly occurring intervals in the histogram for the Thursday to Sunday prices (the right-hand histogram). The same two intervals are also the most commonly occurring in the Monday to Wednesday prices (the left-hand histogram). They are unlikely to correspond to particular days. Can you think of a reason for the high frequencies in these two intervals? (Hint: remember these are prices collected over a 5-week interval.)



Mon	Tues	Wed	Thur	Fri	Sat	Sun
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	25
26	27	28	29	30		

Enrichment

Are pedestrian walking speeds bimodal?

6 The case study described at the beginning of this chapter recorded times for pedestrians walking a 15-metre section of city footpath that was used by many people as a thoroughfare. Subjects were chosen randomly and any who exited or stopped in the section were not included.

ISBN 978-1-107-65599-7

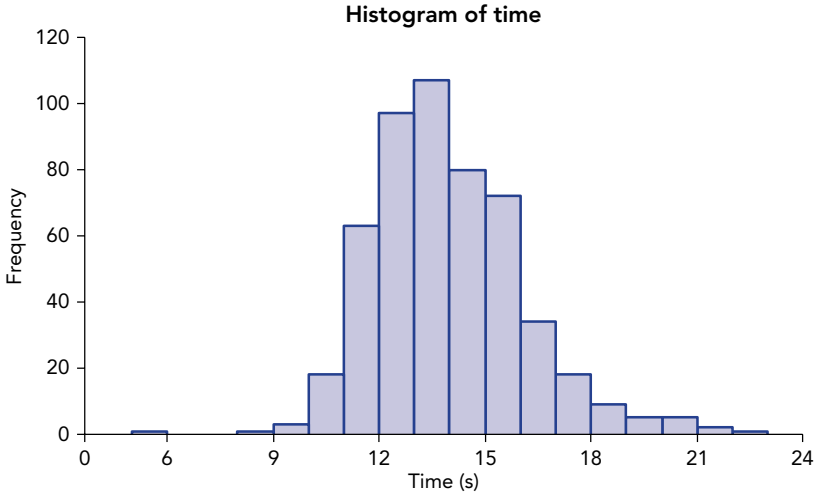
a Below is a sample of 30 times (in seconds) selected at random from the full dataset consisting of 500 times.

17.9 17.0 13.9 17.7 14.8 13.0 16.9 11.6 16.5 14.0 14.6 13.3 14.3 12.0
12.4 16.3 14.2 12.7 15.8 13.9 19.8 14.8 14.9 12.8 15.1 12.5 13.0 12.2
12.2 12.7

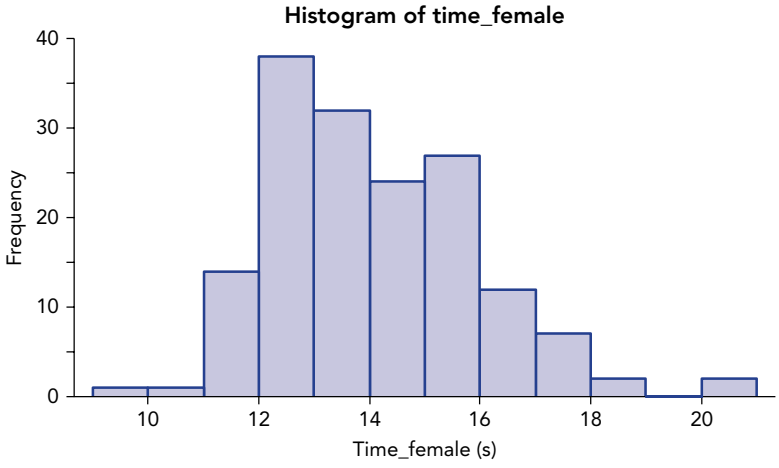
i Draw a histogram of these data with 9 bins starting at 11 s. Does the histogram indicate that the data are bimodal?

ii Now draw a histogram of the same data with 6 bins starting at 11 s. Does the histogram indicate that the data are bimodal?

b Below is a histogram of the whole dataset. Do you think the data are bimodal?



c As part of the study, various variables were recorded for each pedestrian, including their gender and whether they were wearing flip-flops or other footwear. In the whole dataset of 516 pedestrians recorded over almost 3 hours, there were 160 females wearing shoes other than flip-flops. Below is a histogram of these 160 times (in seconds). Do you think the histogram indicates that the data are unimodal or bimodal?



ISBN 978-1-107-65599-7

d Below are two datasets of 20 each, obtained by random sampling of two subsets of the female pedestrians who were not wearing flip-flops: in one subset the pedestrians were walking by themselves (single) and in the other, they were walking in a group (group). Note that if a group walked by, the time for the group was recorded; that is, a pedestrian classified as 'group' was chosen at random from the group and was the only person recorded for that group. Draw histograms of each group and use these to comment on the histogram in part **c** above.

Sample group time (s)

16.4 16.4 14.4 12.0 16.3 13.5 15.2 12.1 14.5 16.6 18.0 14.5 13.3 20.1
15.2 14.5 18.1 12.9 17.9 13.3

Sample single time (s)

11.9 14.9 13.8 12.4 13.0 12.8 14.9 12.7 11.7 13.0 14.9 18.6 11.3 15.5
12.4 11.7 12.9 13.3 15.8 11.9

ISBN 978-1-107-65599-7

2-3 Commenting on quantitative data features

We have seen how important plots are for exploring quantitative data. We have also seen the need to be careful in interpreting them because of the interval choices involved, particularly for smaller datasets. We have also seen that statistics, such as the sample mean, the sample median and the sample range, are useful as numerical summaries. The sample mean and median are measures of the general location or centre of the data, and the sample range is an indication of the spread of the data.

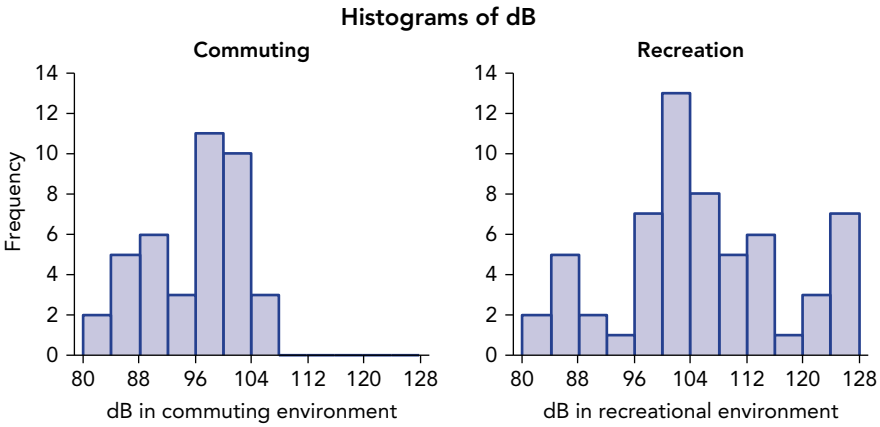
But we have also seen that considering only the values of the sample mean, median and range does not tell us much about what the data look like, and can be misleading.

Sections 2-1 and 2-2 have considered the shape of data – how the data are spread over the lower range of values compared with the upper range, and whether the data indicate that there may be sub-groups that are different. Now we have ways of describing the shape of data as well as the general size (location or centre) and the spread of the data. To describe data, and especially in comparing datasets, we should look at both plots and summary statistics.

Note that in section 2-2, if data look bimodal, it can be because of two sub-groups in the dataset that tend to have different locations or centres.

LET'S START How does recreational noise compare with commuting noise?

Below are histograms on the same scale for the data of section 2-2 on the decibels recorded for commuting and recreational environments.



We have these graphs in section 2-2 but not on the same scale. From the graphs, we can see that the noise levels in commuting environments tend to be less variable than those for recreation, which can also be much greater than in commuting environments. The commuting noise levels also tend to be skewed to the left, whereas those for recreational environments are less skewed – if anything, they may be said to be slightly skewed to the right.

ISBN 978-1-107-65599-7

The sample means, medians and ranges of the two groups are:

Variable	Type	Mean	Median	Range
dB	Commuting	95.08	97.00	24.00
	Recreation	104.65	103.50	45.00

We see that the mean and median commuting noise levels are less than for the recreation noise levels, but, as in the graphs, this is mainly due to the recreation noise levels greater range. For commuting environments, the sample mean is less than the sample median, which often happens with data skewed to the left. For the recreational environments, the sample mean is slightly greater than the sample median, which often happens with data slightly skewed to the right.

In both histograms, there may be indications of sub-groups – which would not be surprising, considering the different types of noises in the two environments – but the datasets are not large enough to say anything more than 'maybe'.

Key ideas

- To describe and compare quantitative datasets, we should consider plots (on the same scale if comparing data) and summary statistics.
- Features of quantitative data can be thought of as location or centre, spread and shape.
- Sample means and medians help to describe location or centre and sample range helps to describe spread.
- Shape can be described in terms of skewness, and whether there are indications of sub-groups in the data.
- If there is skewness to the right, the mean may tend to be larger than the median. If there is skewness to the left, the mean may tend to be smaller than the median.

ISBN 978-1-107-65599-7

Example 4: Does coffee speed up reaction times?

In Example 4 of Chapter 1, there are plots of the reaction times of males and females, and whether they had drunk coffee in the previous hour. Below are the means, medians and ranges of these four groups.

Females

Variable	Coffee	N	Mean	Median	Range
Reaction	No	11	0.185	0.186	0.033
	Yes	17	0.196	0.192	0.096

Males

Variable	Coffee	N	Mean	Median	Range
Reaction	No	21	0.200	0.202	0.051
	Yes	27	0.194	0.186	0.091

Question: How do the reaction times compare between males and females, and whether they've recently drunk coffee or not?

As we could see in the graphs, the females who had drunk coffee, tended to have slightly longer reaction times, as measured by both mean and median, and greater variability. (Perhaps they felt they needed the coffee!) The males who had drunk coffee tended to have slightly faster reaction times, and also greater variability. For the males who had drunk coffee, the plot shows skewness to the right, and the mean time is greater than the median. For the other three groups, there is little skewness, and the means and medians are fairly similar to each other. The group sizes are too small to be able to say if there are any indications of sub-groups.

ISBN 978-1-107-65599-7

Example 5: Comparing fuel prices

In question 6 of Exercise 2A, plots show the data collected on the price of unleaded 91 E10 fuel from a number of randomly chosen service stations each day for 5 weeks, in two groups: 1 = Monday–Wednesday, 2 = Thursday–Sunday. Below are the sample means, medians and ranges.

Variable	Group	N	Mean	Median	Range
Price of unleaded 91 E10	1	105	143.22	143.90	19.10
	2	140	145.64	145.80	12.10

Question: How do the fuel prices in the two groups of days compare?

Group 1 has the lower mean and the lower median, as we would expect looking at the graphs. There is also quite a difference in the range of prices – group 1 has a greater range than group 2. In both groups, the mean and median are very similar, with the mean slightly less than the median, but by less than $1. This may be surprising looking at the plots, which look skewed to the left, but shows that we can't guess the shapes of plots just by looking at the mean and median of the data. The bumpiness of the data makes it difficult to guess what the means and medians will be.

Exercise 2C

1 Example 1 shows a histogram of the prices of cappuccinos at 30 cafes randomly chosen in a city. The data are below.

4.20 3.50 4.20 3.90 3.60 3.90 3.70 3.90 3.60 3.50 4.20
3.70 3.60 3.70 4.20 4.40 4.70 4.65 4.00 4.45 4.40 4.10
4.70 4.40 4.20 4.50 4.50 4.50 4.50 4.50

a Obtain the mean, median and range of these data, and use these plus the histogram in Example 1, to comment on the features of these data.

b From the histogram, would you expect the difference between the mean and median to be greater? What is the effect of the number of observations?

2 Example 2 shows a histogram of the data from a survey on how close or not people's watches are to the correct time. The histogram is of difference in seconds = time on watch – actual time.

Summary statistics for these differences are:

Variable	N	Mean	Median	Range
Time difference	101	25.9	15.0	588.0

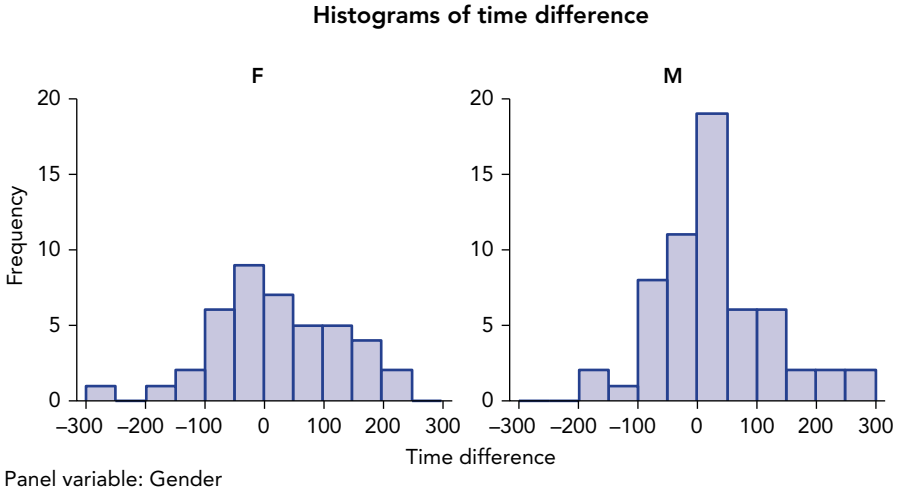
a Interpret the values of the mean, median and range of the data in the context of the times on people's watches.

ISBN 978-1-107-65599-7

b Use the summary statistics plus the histogram in Example 2 to comment on the features of the data.

c Is there a feature of the summary statistics that could be regarded as surprising when looking at the histogram? What would you like to see that might help you better interpret the summary statistics?

3 The data for question 2 include other information, including the gender of the watch wearer. Below are histograms and summary statistics of the time differences for females and for males.



Panel variable: Gender

Variable	Gender	N	Mean	Median	Range
Time difference	F	42	23.6	9.0	543.0
	M	59	27.6	23.0	474.0

a Use the summary statistics and the histograms to comment on these data and the comparison between males and females.

b Do these graphs and summary statistics help in interpreting the overall dataset?

c Is there any indication that the data (overall or the male and female data) are bimodal? Why or why not?

4 Question 4 of Exercise 1C considers the number of complaints about noise per month made in different ways to Sydney airport over a number of years. The plots there and in section 2-1 show the contrast between the monthly numbers of complaints by phone, email and internet. (Notice that a month's internet record is missing.) Below are the sample means, medians and ranges.

Variable	N	Mean	Median	Range
Phone	60	221.8	199.5	466.0
Email	60	114.9	75.0	371.0
Internet	59	47.15	27.00	335.00

a What type of variables are these? Comment on the values of the means, medians and ranges of the data.

ISBN 978-1-107-65599-7

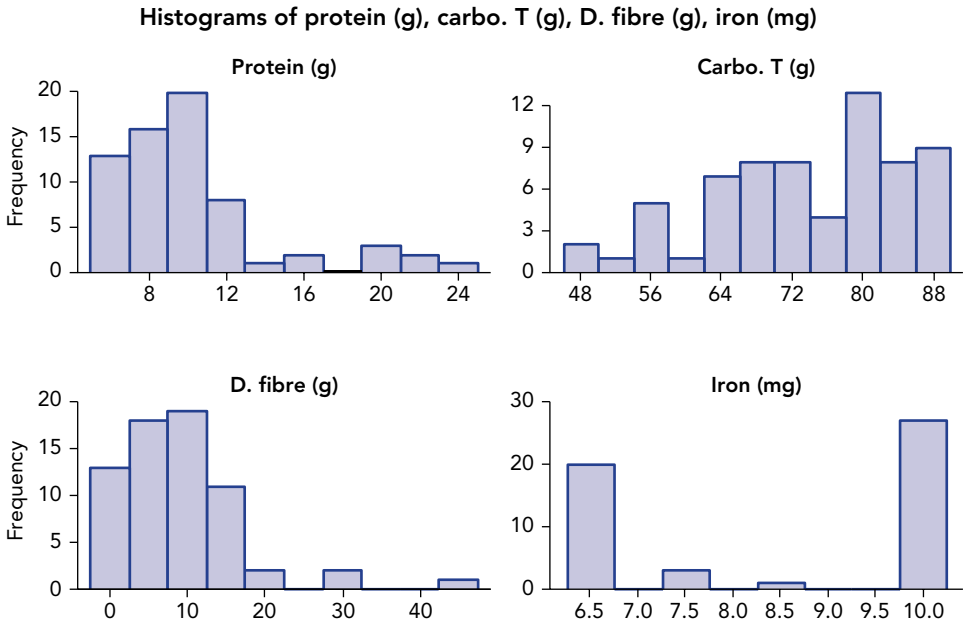
b For each type of complaint, use the summary statistics and the graph to comment on the data features.

c Use the summary statistics and the graphs to compare the numbers of complaints for the different types of complaints.

Enrichment

What's in your cereal?

5 The nutritional information on packets of breakfast cereal was collected for a wide variety of cereals across brands; 66 observations were obtained. The amounts of protein, carbohydrate and dietary fibre in g per 100 g of cereal, and of iron in mg per 100 g of cereal were calculated. Below are histograms and summary statistics of these data. (N* = number of missing observations)



Variable	N	N*	Mean	Median	Range
Protein (g)	66	0	10.070	9.295	17.7
Carbo. T (g)	66	0	73.40	74.20	40.7
D. Fibre (g)	66	0	8.680	8.300	45.0
Iron (mg)	51	15	8.516	10.000	3.7

ISBN 978-1-107-65599-7

a Why have the data been converted to amounts per 100 g of cereal?

b Why aren't the above histograms on the same scale?

c To what accuracy are the data? That is, to how many decimal places have the data been calculated? How can you tell?

d For each of the four variables, use the graphs and the summary statistics to comment on the features of the data.

e What do these four variables have in common?

f Do any of the variables appear to have bimodal data? If so, is there a possible reason that could be investigated?

g What would be important information that we cannot obtain from the above plots and summary statistics?

ISBN 978-1-107-65599-7

Chapter summary

Skewness

- When data are more spread out over the larger (smaller) values than the smaller (larger) values, we say the data are skewed to the right (left)
- When the data have (close to) the same shape on each side of the 'middle', we say the data are (close to) symmetric
- For (close to) symmetric data, the data mean and median are (close to) equal
- Data that are not symmetric are said to be asymmetric.

Sub-groups and bimodalities

- If ordinal data (or count data plotted using bar charts) have two categories (or count values) with higher frequencies than their neighbours, the data may be said to be bimodal
- If plots of quantitative data clearly have two (more than two) clumps of data, the data may be said to be bimodal (multimodal); this may indicate that there are different sub-groups within the data
- Because the appearance of stem-and-leaf plots and histograms depends on the choice of intervals, we need to be careful in saying that a plot looks bimodal.

Commenting on quantitative data features

- To describe and compare quantitative datasets, we should consider plots and summary statistics
- Quantitative data features refer to location or centre, spread and shape
- Shape can be described in terms of skewness, and whether there are indications of sub-groups in the data
- If there is skewness to the right (left), the mean may tend to be larger (smaller) than the median.

Multiple-choice questions

1 Graphs of data are said to be skewed to the right if

A The data are clumped more closely on the left than the right
B The data are more spread out on the right than the left
C There is a long tail on the right
D All of A, B and C

2 If data are asymmetric,

A The data are skewed to the right or to the left
B The data mean and median are not equal
C The data are not symmetric
D All of A, B and C

3 If data are skewed to the right,

A The data mean is always greater than the data median
B The data mean is often greater than the data median
C The data mean is always less than the data median
D The data mean is often less than the data median

4 Data on adult incomes are likely to be

A Symmetric
B Skewed to the left
C Skewed to the right
D All of these

ISBN 978-1-107-65599-7

5 Participants in a charity event can choose to run or walk the route. The overall times for participants are likely to be

A Skewed to the left
B Unimodal
C Bimodal
D Unable to be plotted

6 To compare the resting pulse rates of footballers and golfers, we need to look at

A The data means and medians
B The data ranges
C Histograms or stem-and-leaf plots
D All of these

7 In a report comparing footballers and golfers, histograms of the resting pulse rates of footballers and golfers need to be on the same scale

A To take up less room in the report
B So that the report looks good
C To calculate means and medians
D None of these

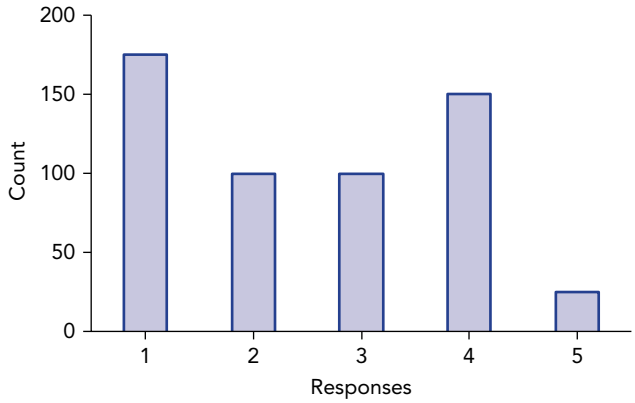
Short-answer questions

1 A survey of 550 people included a question asking for responses to the statement: *Australia's bid to hold the World Cup was a waste of money.*

People were asked to use the following scale. A bar chart of the responses is on the next page.

1 = strongly disagree, 2 = disagree,
3 = neutral, 4 = agree, 5 = strongly agree

ISBN 978-1-107-65599-7



a Can we describe the data as bimodal? Why or why not?

b Describe the data without using the word 'bimodal'.

c How much does using the word 'bimodal' help in describing the data?

2 In section 1-2, the data for Cavendish's 1798 experiment to measure the density of Earth (as a multiple of the density of water) are given and plotted in histograms.

a Would you describe the data as skewed to the left or to the right or asymmetric? Why?

b Would you describe the data as unimodal or bimodal? Why?

c Obtain the mean, median and range of these data and comment on them with reference to the histograms.

3 In question 3 of Exercise 1B, the scores (scaled to be out of 50) are given for 30 people assisting in the checking of a list of words to be used for hearing tests. Look at the histogram you drew for that question.

a Would you describe the data as skewed to the left or to the right or asymmetric? Why?

b Would you describe the data as unimodal or bimodal or unknown? Why?

c Obtain the mean, median and range of these data and comment on them with reference to the histogram.

4 Question 5 of Exercise 1B uses the data of the song lengths for Indie and Alternative rock songs as given in question 4 of Exercise 1A. Look at the histograms you drew for question 5 of Exercise 1B.

a For each type of song, would you describe the data as skewed to the left, skewed to the right or asymmetric? Why?

b For each type of song, would you describe the data as unimodal, bimodal or unknown? Why?

c Obtain the mean, median and range of the data for each type of song. Use these and the histograms to compare these two datasets.

ISBN 978-1-107-65599-7

5 In question 5 at the end of Chapter 1, histograms are given of the weights minus the stated weights (in g) for the three brands of crisps. The three brands have different stated weights: 19 g, 20 g and 21 g.

a For each brand, would you describe the data as skewed to the left, skewed to the right or asymmetric? Why?

b For each brand, would you describe the data as unimodal, bimodal or unknown? Why?

c The means, medians and ranges of the three datasets are given below. Use these and the histograms to compare these three datasets. Are there any surprises?

Variable	Stated weight	N	Mean	Median	Range
Weight-stated weight (g)	19	72	1.833	2.0	9
	20	40	2.250	2.0	4
	21	40	1.775	1.0	6

6 The following is an extract from a 2006 ABS report on the Children's Participation in Cultural and Leisure Activities Survey. This extract is on the issue of time children spend reading for pleasure.

Children who read for pleasure				
Time spent reading in last two weeks (hours)	Age group (years)			Total
	5–8	9–11	12–14	
	Number ('000)			
2 or less	206.9	134.8	112.3	453.9
3–4	150.1	103.8	106.1	360.0
5–9	268.2	207.2	160.6	636.0
10–19	125.9	142.0	153.0	420.9
20 or more	20.4	37.2	55.7	113.3
Total	771.5	624.9	587.7	1984.0

a Can we draw a histogram of these data?

b Why can't we calculate the mean, median and range of these data for each age group?

c What can we find approximations for?

Extended-response questions

7 In Extended-response question 6 of Chapter 1, histograms are given of the number of days after the expiry date on which mould was first noticed on slices of bread of two different brands and two different thicknesses, left in the sun and out of the sun.

a For each histogram, would you describe the data as skewed to the left, skewed to the right, asymmetric or unable to be described? Why?

ISBN 978-1-107-65599-7

b For each histogram, would you describe the data as unimodal, bimodal or unknown? Why?

c The means, medians and ranges of the datasets are given below, except for brand A with thick bread. Why are these data not considered? Use these summary statistics and the histograms to compare these datasets. Are there any surprises?

Brand A, Thin

Variable	Sunlight	N	Mean	Median	Range
Days after	No	55	0.600	1.0	6
	Yes	35	0.400	0.0	8

Brand B, Thick

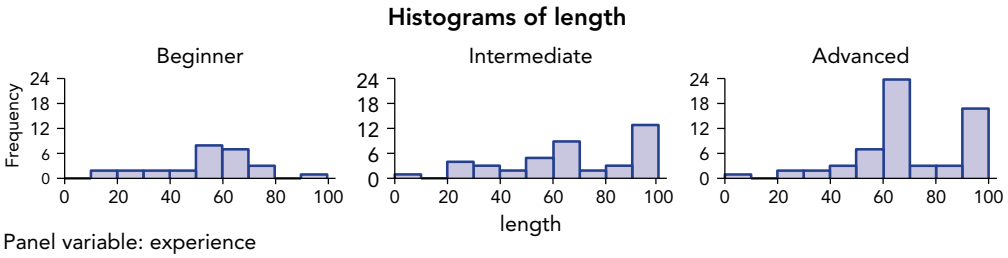
Variable	Sunlight	N	Mean	Median	Range
Days after	No	31	2.677	2.0	6
	Yes	29	2.138	2.0	6

Brand B, Thin

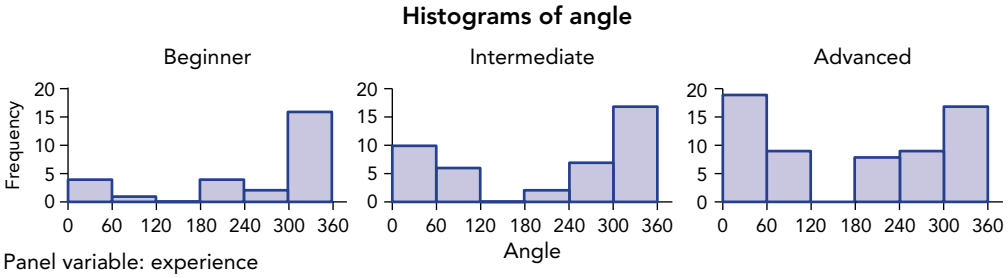
Variable	Sunlight	N	Mean	Median	Range
Days after	No	20	4.500	4.5	6
	Yes	20	3.450	3.0	4

8 In a study on male cricketers, the results for each ball faced by batsmen of different levels of experience (beginner, intermediate and advanced) were recorded. Of those balls that were hit, the ball was not fielded and the length (in m) and the angle (in degrees clockwise from the front) were recorded. A total of 90 balls were faced by beginner and advanced batsmen, and 60 by intermediate batsmen. On the next page are histograms on the same scale, and summary statistics for the lengths and the angles of the balls hit.

ISBN 978-1-107-65599-7



Variable	Experience	N	Mean	Median	Range
Length	Advanced	62	70.06	65.0	91
	Beginner	27	52.46	55.0	87
	Intermediate	42	66.95	65.0	98



Variable	Experience	N	Mean	Median	Range
Angle	Advanced	62	169.2	180.0	320
	Beginner	27	245.2	300.0	320
	Intermediate	42	202.9	260.0	320

a For each histogram, would you describe the data as skewed to the left, skewed to the right, asymmetric or unable to be described?

b For each histogram, would you describe the data as unimodal, bimodal or unknown?

c For the data you described as bimodal, do the data indicate sub-groups or are there other possible reasons for the bimodality?

d For the lengths, use the summary statistics and the histograms to compare the lengths for the levels of experience.

e Compare the data on the angles for the levels of experience. Are the summary statistics of any assistance? Why or why not?

ISBN 978-1-107-65599-7

CHAPTER

3

Probabilities for combinations of two or three events

What you will learn

Introduction

In Australia, many young people, particularly young women, consider changing their diet to one that is vegetarian. There seems to be a wide range of reasons for such a change, including concern for animals that are sometimes badly treated when they are being raised for food, and a desire to lose weight in order to look more attractive. However, there are some concerns that vegetarianism may be linked to a range of eating disorders, which raises many questions.

Do healthy people become vegetarians? Is being a vegetarian healthy for you? Or is it likely that becoming vegetarian represents one step along the road to an eating disorder? Probability can help us examine these and related questions.

An article in *The Sydney Morning Herald* reports on a recent study of the relationship between vegetarianism and eating disorders (www.cambridge.edu.au/statsAC910weblinks). The study found that people with a history of eating disorders are more likely to be vegetarian now, or to have been a

Cambridge GO

AUSTRALIAN CURRICULUM

AC

Statistics and probability

- Chance
- List all outcomes for two-step chance experiments, both with and without replacement using tree diagrams or arrays. Assign probabilities to outcomes and determine probabilities for events **(ACMSP225)**
- Calculate relative frequencies from given or collected data to estimate probabilities of events involving 'and' or 'or' **(ACMSP226)**
- Describe the results of two- and three-step chance experiments, both with and without replacements, assign probabilities to outcomes and determine probabilities of events. Investigate the concept of independence **(ACMSP246)**
- Use the language of 'if ... then', 'given', 'of', 'knowing that' to investigate conditional statements and identify common mistakes in interpreting such language **(ACMSP247)**

ISBN 978-1-107-65599-7

vegetarian in the past. Not only that, around two-thirds of people who had had an eating disorder thought that their vegetarianism was related to it.

But what comes first? Does a decision to change to a vegetarian diet lead to an increased chance of developing an eating disorder? Or does having an eating disorder lead to an increased chance of becoming a vegetarian? Of course, the link – if there is one – shows up in only a proportion of people. There are many young people who are vegetarian and have no problems at all with their eating. People who are vegetarians for religious or cultural reasons are not at increased risk of an eating disorder, but those who have a real focus on what they are eating seem to have an increased risk.

So how can we use ideas of probability to investigate an important health topic such as this? What techniques of probability can we use to assess the risks to ourselves and to people we know? And how can we decide whether there really is a link between having a vegetarian diet and developing an eating disorder? In this chapter, we will continue our study of probability to answer these and other questions.

Terms you will learn

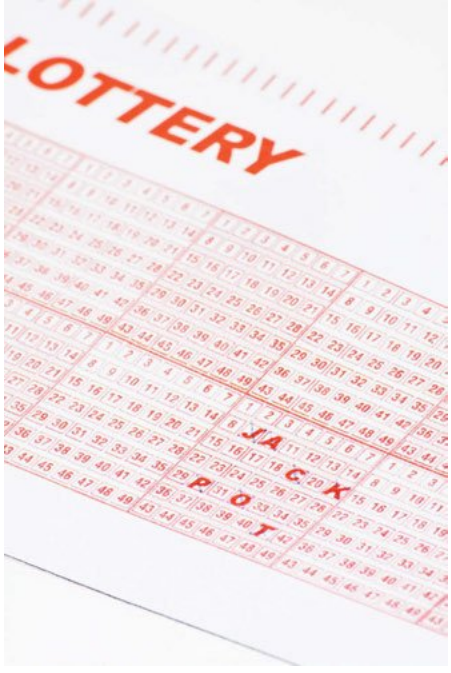
A and B
A or B
complementary
compound event
conditional phrases
conditional probability
disjoint
equally likely outcome
event
exclusive or
experiment
inclusive or
independent
mutually exclusive
probability
sample space
tree diagram
with replacement
without replacement

Pre-test

1 For each pair, state the probability that represents the more likely event:

a 5% or $\frac{1}{25}$ **b** 90.5% or 0.95 **c** $\frac{1}{6}$ or 16%

2 We buy a ticket in a small lottery. Only 50 tickets are sold, and there are three prizes – first, second and third. What is the probability that we will win a prize?



3 Imagine that you buy two tickets in the lottery of the previous question. You are interested in the event 'win at least one prize'. What is the complementary event?

4 An American roulette wheel has 38 numbered sections – 18 red, 18 black and 2 green. When the wheel is spun, what is the probability that one of the green numbers (0 or double-0) comes up?

5 On a Venn diagram show two events, A and B, and shade:

a the event 'A and B' in one colour

b the event 'A or B' using the 'exclusive or' in another colour.

ISBN 978-1-107-65599-7

3-1 Brief review of probability

An experiment is any situation where we don't know what will happen in advance. Although we don't know exactly what will happen, we can list all the possibilities in a **sample space**. An **event** is any of the individual outcomes in the sample space, called a simple event, or any combination of these outcomes, called a **compound event**. **Probability** is a way of measuring chance using a scale from 0 to 1. An event that is impossible has a probability of 0, an event that is certain has a probability of 1, and an event that is as likely to occur as not has a probability of 0.5.

We can estimate the probability of an event using a frequency approach – by repeating the experiment many times and finding the proportion of times that it occurs. If we can't repeat an experiment, we could estimate the probability subjectively using our beliefs about the situation. A third possibility is to find a probability using a modelling approach – for instance, by assuming that the outcomes are **equally likely**.

If the outcomes in a sample space are equally likely, we can share the total probability of 1 between them. For example, in rolling a die there are six possible outcomes; if the die is fair they will each have a probability of $\frac{1}{6}$. If an event consists of several of these outcomes, its probability depends on how many individual outcomes are included. For example, if we want to find the probability of rolling an odd number, we see that there are three possible odd numbers (1, 3 and 5) out of the total six possibilities, so the probability is $\frac{3}{6}$ or 0.5.

Sample space: A list of possible outcomes

Event: Any individual outcome or collection of outcomes

Compound event: Event that consists of several simple events or individual outcomes … *see glossary*

Probability: A way of measuring chance, represented by a number between 0 and 1

Equally likely outcomes: Outcomes of an experiment that are equally likely to occur

In order to use this modelling approach, we need to describe the outcomes carefully and check that they are equally likely. We can do this by repeating the experiment many times and comparing the frequencies for each outcome with the equal frequencies that we expect if the model is correct.

Often we are interested in two (or more) events in the same sample space.

Events A and B are **disjoint** (or **mutually exclusive**) if they have no points in common. For instance, 'rolling a number less than 3' and 'rolling a 6' are disjoint.

Disjoint or **mutually exclusive events:** Outcomes that cannot occur at the same time

ISBN 978-1-107-65599-7

Cambridge University Press

Events A and B are **complementary** if they are disjoint and they cover the sample space completely. For instance, 'rolling an odd number' and 'rolling an even number' are complementary.

The phrase '**A and B**' specifies that both of the events A and B occur. For instance, if A = 'rolling an even number' and B = 'rolling a number greater than 3' then 'A and B' represents the outcomes 4 and 6.

The phrase '**A or B**' is ambiguous as it can mean 'at least one of A and B occurs' (the '**inclusive or**') or it can mean 'exactly one of A and B occur' (the '**exclusive or**'). The first of these is more common, but we have to be careful to specify which we mean. For instance, if A = 'rolling an odd number' and B = 'rolling a prime number (2, 3, 5)', then 'A or B' represents the outcomes 1, 2, 3 and 5 if we are using the 'inclusive or'.

Complementary: Two **disjoint** events that cover the entire sample space

A and B: A situation where both events A and B occur

A or B: A situation where at least one of the events occurs (**inclusive or**) or exactly one of the events occurs (**exclusive or**)

Example 1: Two-up

The Australian game of Two-up was played by miners on the goldfields in the early 1800s, and by Australian soldiers, the Diggers, in both World Wars. Because of this history, it is a legal form of gambling on Anzac Day. In the game, the 'spinner' tosses two ordinary coins.

a What is the sample space for this experiment, and what is the probability of each outcome?

b You can bet on *Heads* (both coins land showing heads) or *Tails* (both coins land showing tails. What is the probability that the event '*Heads* or *Tails*' occurs? Should we use the 'inclusive or' or the 'exclusive or' here?

c The complementary event to '*Heads* or *Tails*' is called *Odds*. What outcomes are included in *Odds* and what is its probability?

d What do you think happens if *Odds* occurs a second time?

Solution

a The sample space can be written as {hh, ht, th, tt}, the two letters showing the result on the two coins. Each of these outcomes is equally likely, so each has probability 0.25.

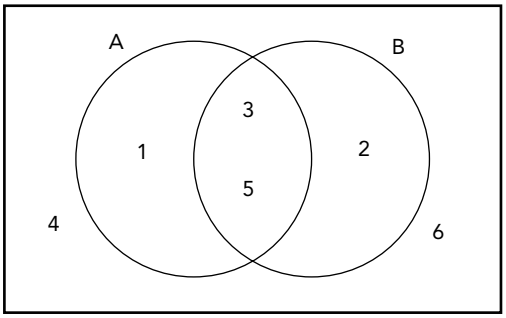
b The probability of '*Heads* or *Tails*' is 0.5, since this event includes the two outcomes hh and tt. There is no ambiguity about the 'inclusive or' or the 'exclusive or' here, since the two outcomes are disjoint. We can get two heads or two tails but we can't get both on the same toss.

c The outcomes ht and th are included in *Odds*, so its probability is 0.5.

d If *Odds* occurs a second time, the coins are tossed again, and so on. But after a run of five *Odds*, the spinner wins all the bets on *Heads* and all the best on *Tails*.

ISBN 978-1-107-65599-7

We can use a Venn diagram to show events visually as circles, either interlocking circles if the events can occur together, or separate circles if the events are disjoint. For instance, the Venn diagram below shows the events A = 'rolling an odd number' and B = 'rolling a prime number'.



We can use a two-way table to show information about two variables. The frequencies can be used to estimate probabilities of individual events or combinations of events. We will give more details in the following section.

Key ideas

- Definition of experiment, sample space and event.
- Probability as a numerical value between 0 and 1.
- Probability estimated using frequency, belief or modelling.
- Equally likely outcomes all have the same probability.
- Definition of disjoint (mutually exclusive) and complementary events.
- 'A and B' means that both A and B occur.
- 'A or B' means that at least one of A and B occur ('inclusive or'), or sometimes that exactly one of A and B occurs ('exclusive or').
- Venn diagrams and two-way tables can be used as visual representations.

Exercise 3A

1 List a sample space for each of these experiments:

a You ask five friends to your birthday party.

b A student is sitting for examinations in three subjects: mathematics, English and history.

c A dozen bottles of wine are opened and checked to see whether they are 'corked' (given a bad taste and smell from a chemical that is sometimes present in the cork).

ISBN 978-1-107-65599-7

2 For each of these probabilities, choose the best description from this list: impossible, unlikely, as likely to happen as not, likely, certain

a 18% **b** $\frac{5}{90}$ **c** 0.52

d $\frac{1}{20}$ **e** 100% **f** 0.95

3 A batch of 80 batteries is being tested as part of a quality control program. The batch actually contains five that are defective.

a If one battery is selected at random and tested, what is the probability that it will be defective?

b The first battery selected was defective. A second one is selected for testing. What is the probability that this one is also defective?

4 Dice with ten sides are used in various games (two of them can be used to select percentages). Such dice have faces that are in the shape of a kite, numbered from 0 to 9 and each is equally likely to be rolled. The solid has a technical name – it is called a 'pentagonal trapezohedron'. We define two events: A = 'roll an odd number' and B = 'roll a number 7 or greater'. Find the following:

a P(A)

b P(B)

c P(A and B)

d $P(B^C)$, the complement of B

e P(A or B) using the 'inclusive or'

f P(A or B) using the 'exclusive or'

5 The Year-12 students at a college were surveyed about their use of cigarettes and alcohol. The results are shown in the table.

	Drinks alcohol	Doesn't drink	Total
Smokes cigarettes	16	2	18
Doesn't smoke	48	37	85
Total	64	39	103

What is the probability that a randomly selected Year-12 student:

a smokes cigarettes and drinks alcohol?

b smokes cigarettes or drinks alcohol?

c doesn't smoke cigarettes or drink alcohol?

ISBN 978-1-107-65599-7

Enrichment

Is eye colour linked to hair colour?

6 The university students in a large lecture course in statistics were classified by their eye colour and their (natural) hair colour. The table summarises the results (the 'other' eye colours include hazel and green).

		Eye colour			
		Brown	Blue	Other	Totals
Hair colour	Black	68	20	20	108
	Brown	119	84	83	286
	Red	26	17	28	71
	Blond	7	94	26	127
	Totals	220	215	157	592

One of these students is selected at random.

a What is the probability that:

i the student has red hair?

ii the student has blond hair and blue eyes?

iii the student has brown hair or brown eyes?

b Are blond-haired students or brown-haired students more likely to have blue eyes? Explain how you can find an answer to this question.

ISBN 978-1-107-65599-7

3-2 Pairs of outcomes

In this section, we will focus on experiments that involve two stages and show how we can represent the outcomes using two-way tables. We will start with a simple example – rolling two dice. Each die can land showing faces 1 to 6, so we can represent the sample space from rolling two dice in a table, with the row showing the result on the first die and the column showing the result on the second die. The sample space consists of the 36 cells showing the pair of numbers obtained on the dice. If the dice are fair or 'unbiased', and if they are rolled properly, maybe using a container to shake them, then all 36 outcomes are equally likely, each with probability $\frac{1}{36}$.

		Second die (green)					
		1	2	3	4	5	6
First die (red)	1	1,1	1,2	1,3	1,4	1,5	1,6
	2	2,1	2,2	2,3	2,4	2,5	2,6
	3	3,1	3,2	3,3	3,4	3,5	3,6
	4	4,1	4,2	4,3	4,4	4,5	4,6
	5	5,1	5,2	5,3	5,4	5,5	5,6
	6	6,1	6,2	6,3	6,4	6,5	6,6

If the two dice are of different colours, the colour can be used to show which was rolled first and which was rolled second. But even if they are the same colour, or they are rolled at the same time, the table will still show the sample space and the outcomes will still have probability $\frac{1}{36}$ each.

Example 2: Pairs of outcomes

Use the table above to find the probability of:

a rolling a double-1 (known as 'snake eyes' in dice games such as Craps)

b rolling any double

c getting a six on the first or the second die

d getting a total of 7.

Solution

a The double-1 is equivalent to getting a 1 on the first die and a 1 on the second die. This has probability $\frac{1}{36}$, the same as any individual outcome.

b Since there are six ways to get a double (1,1; 2,2; 3,3; 4,4; 5,5; 6,6), the probability of getting a double is $\frac{6}{36} = \frac{1}{6}$.

ISBN 978-1-107-65599-7

c The 'inclusive or' seems appropriate here – getting a six on at least one of the dice – on the first die, or on the second die, or on both. Since there are 11 ways of doing this, the probability is $\frac{11}{36}$.

d A total of 7 can be obtained from six different outcomes (6,1; 5,2; 4,3; 3,4; 2,5 and 1,6), so the probability is $\frac{6}{36} = \frac{1}{6}$.

Data that have been collected can also be presented in a two-way table. The rows of the table represent the possibilities for one aspect of the data, and the columns of the table represent possible outcomes for another aspect. The cells inside the table are the frequencies with which each combination occurs in the data. As an example, we can look at the data from the vegetarianism and eating disorders study from the introduction to this chapter.

	Current vegetarian	Former vegetarian	Never vegetarian	Totals
History of eating disorder	22	26	45	93
No history of eating disorder	4	4	59	67
Totals	26	30	104	160

The rows indicate whether a person in this study has a history of eating disorders or not, and the columns classify them in terms of their vegetarian status – current, former or never. If a person from the 160 participants in this study is chosen at random, each person will have the same chance of being chosen, $\frac{1}{160}$. We can find the probability of various events involving eating disorders or vegetarian status or both.

Example 3: Using two-way tables

A respondent is selected at random from the participants in the vegetarianism and eating disorders study. Find the probability that:

a the person has a history of eating disorders

b the person is a current vegetarian and has no history of eating disorders

c the person has never been a vegetarian or has never had a history of eating disorders.

Solution

a Since the sample of participants in the study contains 93 people who had a history of eating disorders, the probability that one of them is selected is $\frac{93}{160} = 0.581$.

b There are 4 current vegetarians with no history of eating disorders, so the probability is $\frac{4}{160} = 0.025$.

c Using the 'inclusive or', there are 4 + 4 + 59 + 45 people in the study who satisfy these conditions, so the chance that one of them is selected is $\frac{112}{160} = 0.7$.

ISBN 978-1-107-65599-7

Key ideas

- The outcomes from two-stage experiments can be represented in a two-way table.
- Collected data with information on two variables can be presented in a two-way table.
- In each case, we can find probabilities of single events or combinations of two events from the information in the table.

Exercise 3B

1 In horse race betting, an Exacta is a bet that requires you to specify the winner and the second place horses in that order. (This is different from a Quinella, in which you specify the winner and second place horses in either order.) In one particular horse race there are five starters: Abracadabra, Basil Boy, Close Call, Defiant and Eucalyptus.

a Draw a table showing the possible Exacta bets that you can make, with rows representing the first horse and columns representing the second.

b How many outcomes are there in the sample space? Are they all equally likely from the point of view of a gambler?

c Do the outcomes all have the same probability of occurring in the race?

2 Here is some information about senior students at a high school:

	Studies art	Doesn't study art	Total
Studies music	13	9	22
Doesn't study music	1	37	38
Total	14	46	60

What is the probability that a randomly selected student:

a studies music but not art?

b studies music and art?

c studies music or art?

ISBN 978-1-107-65599-7

3 A famous series of experiments was carried out at Stanford University in the late 1960s. Four-year-old preschool children were offered a marshmallow, but told that they would get two instead of one if they waited 15 minutes. About 10 years later, a short questionnaire concerning the coping and mental competence of the children was mailed to parents. Other follow-up questionnaires were given at later times.

Variations of this experiment are still being run, nearly 50 years later. The original results are quite complex, but here is a summary that shows their main characteristics:

	Below average competence	Above average competence	Total
Ate marshmallow within 15 minutes	24	36	60
Waited 15 minutes to get the second marshmallow as well	10	25	35
Total	34	61	95

a What is the probability that a preschool child in this study gave in to temptation and ate the first marshmallow within 15 minutes?

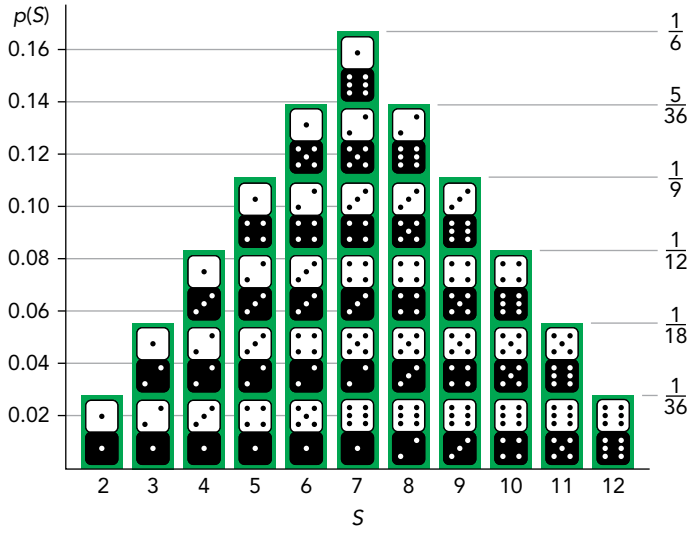
b What percentage of 14 year olds were rated as 'above average' on coping and mental competence? Is this a surprising value? Can you suggest any reasons for it?

c What is *deferred gratification*? What do you think the researchers concluded from these results about deferred gratification?

4 The famous German mathematician Gottfried Leibniz (1646–1716) was co-inventor of calculus (which you may study in Years 11, 12 and beyond) together with Isaac Newton. Leibniz wrote in one of his works that '... with two dice, it is as feasible to throw a total of 12 points as to throw a total of 11'. What do you think?

ISBN 978-1-107-65599-7

5 The diagram below shows another way of representing the outcomes from rolling two fair dice, in this case one white and the other black.



a Explain how the outcomes of the sample space are organised in the diagram.

b What is the probability that the two dice land showing a total of 10? What other total has the same chance of occurring?

Enrichment

What are you most likely to die from?

6 According to data from the Australian Bureau of Statistics, the leading cause of death for Australians is ischaemic heart disease, which includes angina, blocked arteries in the heart and heart attacks. Ischaemic heart disease is responsible for 15.1% of all deaths. Cerebrovascular disease, including strokes and blocked arteries of the brain, is the second most common cause of death, accounting for 7.8% of all deaths. The third most common cause of death is dementia and Alzheimer's disease, which causes 4.4% of all deaths. (The figures are from 2010 ABS data, (www.cambridge.edu.au/statsAC910weblinks). In that year, there were 143 473 deaths recorded.)

a What was the probability that a death in 2010 was caused by cerebrovascular disease?

b What was the probability that a death in 2010 was caused by ischaemic heart disease or cerebrovascular disease? Are you using the 'inclusive or' or the 'exclusive or' to answer this question?

c Do you think that the leading causes of death would be the same for male and female Australians? If we divide Australians into three age groups – young, middle aged, old – do you think that the leading causes of death would be the same for all three groups?

ISBN 978-1-107-65599-7

3-3 Conditional language and applications

The table of data from the vegetarianism and eating disorders study points out that we sometimes want to find probabilities under conditions that are different from the whole sample.

	Current vegetarian	Former vegetarian	Never vegetarian	Totals
History of eating disorder	22	26	45	93
No history of eating disorder	4	4	59	67
Totals	26	30	104	160

We saw in Example 2 that, for a person selected at random from the participants in this study, the probability of having a history of eating disorder is $\frac{93}{160} = 0.581$. But of course if you pick someone randomly from Years 9 and 10 in your school, or from the general population, the chance that they have a history of eating disorder is not 58%. It is likely to be much smaller than this. The people in this study were selected specifically from two groups: an eating disorder clinic (the 93 in the top row), or a 'control' group that had no history of eating disorder (the other 67).

So it would make sense to look at the probabilities for these two groups separately. We could ask:

- What is the probability of being a current or former vegetarian if a person has a history of eating disorder? In this group, the probability is $\frac{48}{93} = 0.516$.
- What is the probability of being a current or former vegetarian if a person has no history of eating disorder? In this group, the probability is $\frac{6}{67} = 0.119$.

The chance of being a current or former vegetarian is more than four times as high for people with an eating disorder than for those who don't have an eating disorder.

Phrases such as 'if ...', 'given that ...' and 'knowing that ...' are called **conditional phrases**, and the probabilities that are based on them are called **conditional probabilities**. For a conditional probability, the reference group is a sub-group of the original data – a sub-group defined by the condition.

Conditional phrases: Phrases such as 'if ...', 'given that ...' and 'knowing that ...'

Conditional probability: Probability based on a smaller reference group – a sub-group of the sample space or set of data

The same idea can be used to find probabilities from information about the sample space. For example, we can consider the sample space from rolling two dice that we looked at in the previous section. We can replace the pairs of outcomes from the dice by their totals, to modify the table like this:

(Totals)		Second die (green)					
		1	2	3	4	5	6
First die (red)	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

ISBN 978-1-107-65599-7

Each of the 36 cells in the table represents an equally likely outcome in the original sample space. We can find probabilities involving totals under various conditions, as the next example shows.

Example 4: Probabilities involving totals

Two unbiased dice are rolled and the total number of points is noted. Using the table on page 65, find the probability that:

a the total is 10

b the total is 10, if it is known that at least one six was rolled

c that the total is 10, given that the total is greater than 7

d that the total is 10, if the two dice showed the same number.

Solution

a The reference group is the whole 36 points in the sample space. The probability is $\frac{3}{36} = \frac{1}{12}$, since the total of 10 occurs in three of the outcomes. This is not a conditional probability.

b Here the reference group is all cases where at least one of the dice showed a six – the 11 outcomes in the bottom row and the rightmost column. In two of these cases we have a total of 10, so the conditional probability is $\frac{2}{11}$.

c The reference group is all outcomes where the total is more than 7, the 15 outcomes in the lower-right triangle. In three of these cases we have a total of 10, so the conditional probability is $\frac{3}{15}$.

d The reference group is the six cases along the diagonal where both dice show the same number. Only one of these has a total of 10, so the probability is $\frac{1}{6}$.

When we see in a question about probability a phrase such as 'if ...', 'given that ...' or 'knowing that ...' we know that we are being asked about a conditional probability. We must be careful to identify the correct reference group for the calculation. And if we are talking about a conditional probability, we must be careful that we make it clear what the reference group for the calculation actually is.

Key ideas

- Conditional probabilities are based on a reference group that is smaller than the whole sample space, or the whole set of data.
- In a two-way table, conditional probabilities are useful for comparing groups
- Phrases such as 'if ...', 'given that ...' and 'knowing that ...' are indications that we are looking for conditional probabilities.

ISBN 978-1-107-65599-7

Exercise 3C

1 Students in a university class took a short survey that asked them, among other things, whether they smoked and what level of exercise they regularly carried out. Here are the results from these two questions:

	Low	Moderate	High	Totals
Non-smoker	82	106	32	220
Smoker	13	15	2	30
Totals	94	121	34	250

a What is the probability that a student selected at random from this group:

i does a high level of exercise regularly, if it is known that they are a smoker?

ii is a smoker, if it is known that they do a high level of exercise regularly?

b Explain carefully the difference between these two probabilities.

2 The students in the previous questionnaire were also asked whether they drank alcohol regularly. Their height and weight were used to calculate their 'body mass index' (BMI) – if this was over 25, they were classed as 'overweight'. Here are the results:

	Normal	Overweight	Totals
Drinks alcohol	82	28	110
Does not drink	123	17	140
Totals	205	45	250

Use this data to estimate the probability that:

a a student is overweight and drinks alcohol regularly

b a student is overweight, given that they drink alcohol regularly

c a student drinks alcohol regularly given that they are overweight.

3 To play Craps in a casino, you roll two dice. You win immediately if you roll a total of 7 or 11, and lose immediately if you roll a total of 2, 3 or 12. Otherwise the total that you have rolled becomes your 'point', and you continue rolling the dice. If your point comes up again, you win, but if a total of 7 comes up you lose; any other total is disregarded.

a What is the probability that the game is decided on the first roll of the dice?

b What is the probability that your point is 5 if the game is not decided on the first roll?

c Assume that your point was 5 and the game was finished on the second roll. What is the probability that you won?

ISBN 978-1-107-65599-7

4 On 15 April 1912, the RMS *Titanic* collided with an iceberg during the ship's maiden voyage from Southampton to New York. The ship sank and only around one-third of the people onboard survived. The table below shows information about the group to which each person belonged (crew or first-, second- or third-class passenger) and whether they survived or not.

	Survived	Didn't survive	Totals
Crew	212	673	885
First class	203	122	325
Second class	118	167	285
Third class	178	528	706
Totals	711	1490	2201

a What was the probability of survival if a person was part of the crew?

b What was the probability of survival if a person was a first-class passenger? What about second- and third-class passengers?

c If a person was found to have survived, what is the probability that they had been a third-class passenger?

d Comment on these results. What other variable might be associated with a greater chance of survival?

5 Let's look again at the results from the Stanford experiment on delayed gratification in young children. Four-year-old preschool children were offered a marshmallow, but told that they would get two instead of one if they waited 15 minutes. About 10 years later, a short questionnaire concerning the coping and mental competence of the children was mailed to parents. The summary in the table shows the main characteristics of the results.

	Below average competence	Above average competence	Total
Ate marshmallow within 15 minutes	24	36	60
Waited 15 minutes to get the second marshmallow as well	10	25	35
Total	34	61	95

a What is the probability of being rated as 'above average competence' for those students who ate the marshmallow at age 4?

b What is the probability of being rated as 'above average competence' for those students who waited at age 4 (and got the second marshmallow as well)?

c What could you conclude from these conditional probabilities?

d Could you suggest an explanation for the results?

ISBN 978-1-107-65599-7

Enrichment

Has the mortality rate changed?

6 In London in 1661, John Graunt published an early 'life table' showing the numbers of people from a group of 100 births surviving to various ages.

Here is the table:

Age	0	6	16	26	36	46	56	66	76	86
Survivors	100	64	40	25	16	10	6	3	1	0

a What was the probability in London at that time of surviving to age 6? Does this seem correct?

b What was the probability that a 6 year old would still be alive at age 26? Explain why this is a conditional probability.

c How many times more likely was a 36 year old than a 16 year old to reach the age of 76?

d As a comparison, find the current probability in Australia that a newborn baby survives to age 6.

ISBN 978-1-107-65599-7

3-4 The concept of independence

In ordinary language, independence means that one thing is not affected by another. If you are an independent thinker, your views are not influenced by the views of your friends. If your brother is living independently, he has his own home and does not have to get permission to go out to a party. If a country is independent, it can determine its own policies and direction in the world, rather than being told what to do by another country.

In our study of probability we have to be more careful about the meaning of the term. We need a formal definition of independence – one that is written down in terms of a logical statement that does not depend on the ordinary meanings of the word. Here is the definition:

Two events A and B are **independent** if and only if P(A and B) = P(A) × P(B).

We can apply this definition to solve problems involving probabilities.

Independent events: Two events A and B are independent if P(A and B) = P(A) × P(B)

Example 5: Space tourism

Tourism in space is not science fiction but current fact! Several wealthy individuals have already been on extended trips into space, and bookings are currently being taken for less expensive 'suborbital' spaceflights (www.virgingalactic.com) and more expensive trips to the moon (www.spaceadventures.com). Students in a university statistics class were introduced to these websites and asked if they would be interested in going into space as tourists.

	Not interested	Interested	Totals
Female	18	27	45
Male	22	33	55
Totals	40	60	100

a One of these students is selected at random. What is the probability that they are interested in becoming a space tourist?

b What is this probability if the student is female? How does this relate to the result from part **a**?

c What is the probability that the selected student is female? What is the probability that they are female and interested in becoming a space tourist? How do these relate to the result in part **a**?

d Does there seem to be a relationship between sex and interest in space tourism?

Solution

a P(interested) = $\frac{60}{100}$ = 0.6.

ISBN 978-1-107-65599-7

b P(interested given female) = $\frac{27}{45}$ = 0.6. Note that this is a conditional probability. So P(interested given female) = P(interested). The conditional probability is the same as the overall or unconditional probability.

c P(female) = $\frac{45}{100}$ = 0.45. P(female and interested) = $\frac{27}{100}$ = 0.27. We can see that P(female and interested) = P(female)×P(interested) since 0.27 = 0.45 × 0.6.

d In this situation, shown by the results in parts **b** and **c**, the two events are independent.

In summary, two events A and B are independent if

- the conditional probability P(A given B) is equal to the (unconditional) probability P(A), or
- the probability of both events P(A and B) is equal to the product of the individual probabilities P(A)×P(B).

We can apply either of these to solve problems involving probabilities.

Example 6: Independent events

A demolition company uses explosive charges to blow up disused buildings. Each charge has a probability of 0.95 of detonating. Two independent charges are laid, each one sufficient to blow up a warehouse.

a Find the probability that:

i both charges detonate

ii neither charge detonates

iii at least one charge detonates, and the warehouse is blown up.

b Illustrate your results using a Venn diagram.

Solution

a i We define two events: A = 'first charge detonates' and B = 'second charge detonates'. The description that 'two independent charges are laid' implies that the events A and B are independent. The probability that both charges detonate is P(A and B) = P(A) × P(B) = 0.95 × 0.95 = 0.9025. Therefore there is just over 90% chance that both charges detonate.

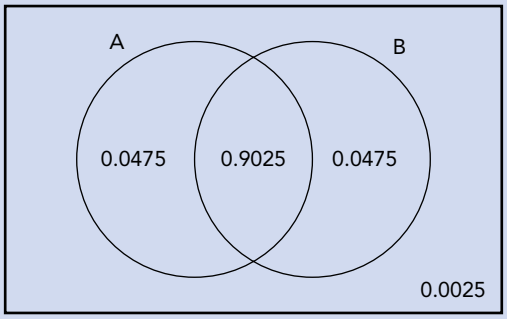
ii The probability that neither charge detonates is P(AC and BC) = P(AC) × P(BC) = 0.05 × 0.05 = 0.0025.

iii 'At least one charge detonates' is the event 'A or B' using the 'inclusive or', and this event is the complement of 'neither charge detonates', so P(A or B) = 1 – P(AC and BC) = 1 – 0.0025 = 0.9975.
There is almost 100% chance that at least one of the charges will detonate – which is, of course, why the company uses two charges.

HINT
Note that if A and B are independent, the complementary events are also independent.

ISBN 978-1-107-65599-7

b We can show the situation using a Venn diagram with a circle for each of the events A and B. We know that each circle contains probability 0.95. In order to fill in the probabilities in the each region of the circles, we start from the inside. From part **a i**, we know that the intersection contains probability 0.9025, and so the probability in A but not in B will be 0.95 – 0.9025 = 0.0475. This is also the probability in B but not in A. From part **a ii**, we know that the probability outside both circles is 0.0025. Finally, we can check that the probabilities sum to 1.



If we have data in a two-way table, we can calculate the probability of individual events A and B and the probability that A and B both occur. We can use these probabilities to check whether the events are independent. Here is an example.

Example 7: Are the events independent?

A study examined the relationship between survival of heart disease patients and pet ownership. Each patient in the study was classified by whether they owned a pet and whether they survived for at least one year.

The table below shows the data collected.

	Died	Survived	Totals
No Pet	11	28	39
Pet	3	50	53
Totals	14	78	92

Is survival for at least a year independent of having a pet? Or is there some relationship between these two events?

Solution

If we define the events A = 'survived' and B = 'having a pet' then we can use the data to estimate the probability of each event: $P(A) = \frac{78}{92} = 0.848$, $P(B) = \frac{53}{92} = 0.576$.
We can also estimate the probability of surviving and having a pet: $P(A \text{ and } B) = \frac{50}{92} = 0.543$.

ISBN 978-1-107-65599-7

We check the definition of independence: P(A and B) = 0.543, P(A) × P(B) = 0.488. The two events don't seem to be independent. Since P(A and B) > P(A) × P(B), we would conclude that the two events occur together more often than they would if they were independent.

Having a pet and surviving seem to go together more than we would expect if the events were independent. The conditional probabilities tell the same story:

- P(survived if you have a pet) = $\frac{50}{53}$ = 0.943
- P(survived if you don't have a pet) = $\frac{28}{39}$ = 0.718

Key ideas

- Ideas about independence are influenced by the way the word is used in English.
- In probability, A and B are independent if and only if P(A and B) = P(A) × P(B).
- Independent events are shown on a Venn diagram with overlapping circles with the probabilities satisfying the equation for independence.
- Data in a two-way table can be used to check for the independence of two events.

Exercise 3D

1 In an intuitive way, we could say that two events A and B are independent if the probability that A occurs is not affected by whether B occurs or doesn't occur, and vice versa. Sometimes we can see intuitively that events are independent. Explain why these events would be independent:

a An experiment consists of tossing a coin and then rolling a die.
A = 'the coin lands heads' and
B = 'the die shows a 6'.

b An experiment consists of interviewing two people selected at random and asking them which party they plan to vote for at the next election.
C = 'first person plans to vote Labor' and D = 'second person plans to vote Liberal'.

2 A bag contains five balls: three are red and two are black. We draw out two balls from the bag at random and **with replacement**. If A is the event that we draw out a black ball on the first draw, and B is the event that we draw out a black ball on the second draw, are the events A and B independent or dependent? If we draw the two balls from the bag **without replacement**, are the events A and B independent or dependent?

3 An oil company drills holes in two separate locations. At each location there is a probability of 0.15 of striking oil, and success at one location is independent of success at any other location. Find the probability that the company finds oil:

a at both locations

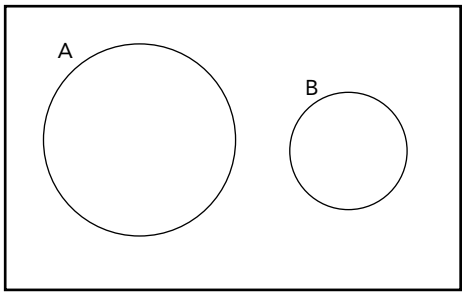
b in at least one of these locations.

With replacement: Item selected at the first stage is replaced before the next selection

Without replacement: Item selected at the first stage is not replaced before the next selection

ISBN 978-1-107-65599-7

4 A Venn diagram shows two events A and B as non-overlapping circles. Are the two events independent?



5 Consider the study on pet ownership and survival after a heart attack, but with the results changed from what actually happened:

	Died	Survived	Totals
No Pet	10	30	40
Pet	15	45	60
Totals	25	75	100

One of the participants in this study is selected at random.

a Find the probability that a person had a pet, and also the probability that the person survived.

b Find the probability that the person had a pet and survived.

c What is the probability of survival if a person had a pet? What is the probability of survival if the person did not have a pet?

d What can you conclude from these answers about the relationship between independence and conditional probability?

Enrichment

Can we assume independence in a system?

6 The ideas of independence and probability are widely used to assess the reliability of electronic or mechanical systems. A famous instance was the failure of the *Challenger* Space Shuttle in January 1986. An investigation after the disaster showed that the problem was caused by a failure in one of the O-ring seals in the Shuttle's right booster rocket. For any particular O-ring, the probability of failure was 0.023, an acceptably low value. However, for the launch to be successful, all the O-rings had to work. You can find more information about this at www.cambridge.edu.au/statsAC910weblinks.

a Explain why P(at least one O-ring fails) = 1 – P(none of the six O-rings fail).

b Assuming that the O-rings work independently, how would you calculate P(all six O-rings hold)?

c What is the overall probability of problems with the launch?

d The O-rings were known to be more likely to fail in low temperatures. What does this say about the assumption of independence?

ISBN 978-1-107-65599-7

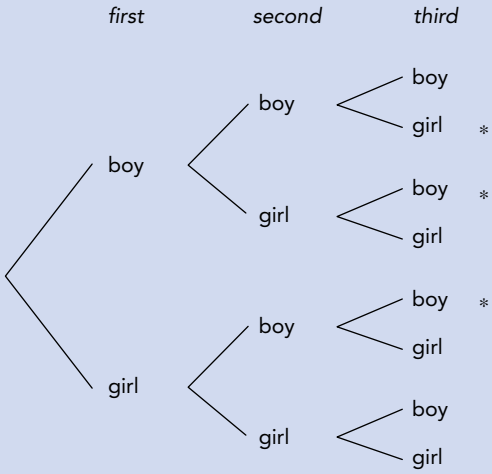
3-5 Special case of two- or three-stage chance experiments

Sometimes we can consider an experiment in several stages. One stage of the experiment is done first, then another stage, then maybe further stages. In this case, the sample space can be shown in the form of a **tree diagram**. Here is a simple example.

Tree diagram: Used to display the results of a two- or three-stage experiment (or even more stages, but then the tree gets complex); probabilities can be written on each connecting line and multiplied together to get the probability for each branch

Example 8: Using tree diagrams

Consider the experiment of having a family of three children in terms of whether each child is a boy or a girl. Show the sample space on a tree diagram. Explain how the tree diagram is constructed and show that the probability of getting two boys is $\frac{3}{8}$. What assumption do we have to make here?



Solution

The tree diagram shows the sample space. The first child is a boy or a girl. Whichever occurs, the second child is also a boy or a girl. And for every combination for the first two children, the third child is a boy or a girl. The lines indicate the possible outcomes making up a branch – for example, the top branch represents the outcome of getting three boys.

ISBN 978-1-107-65599-7

We could list the sample space as {bbb, bbg, bgb, bgg, gbb, gbg, ggb, ggg}, using b = boy and g = girl. However, the tree diagram gives a visual representation showing how the three stages of the experiment were carried out.

The three outcomes with stars next to them represent the event 'getting two boys', as long as we interpret this as 'getting exactly two boys'. If we assume that boys and girls are equally likely at each stage of the experiment, the eight outcomes each have probability $\frac{1}{8}$ and so P('getting exactly two boys') = $\frac{3}{8}$.

If we interpret 'getting two boys' to mean 'getting at least two boys', we must also include bbb as part of the event. In this case, P('getting at least two boys') = $\frac{4}{8} = \frac{1}{2}$.

If the outcomes in the sample space are not equally likely, we find probabilities in a different way. At each stage in the tree diagram, we write the probabilities on the lines. The overall probability of a branch is found by multiplying the probabilities written on the lines that make up that branch. Here is an example showing how this is done. An urn is a type of jar, often used in probability modelling of physical situations.

Example 9: A model of heat

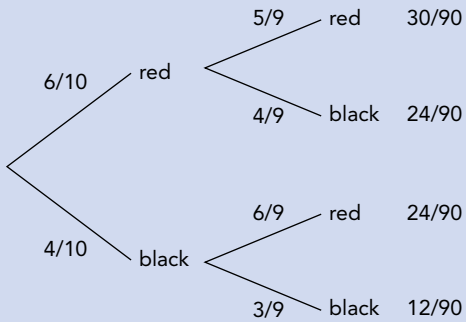
Radiation and reflection of heat can be modelled using an urn containing red and black balls. The red balls are interpreted as 'hot' and the black balls are interpreted as 'cold'. An urn contains 6 red balls and 4 black balls.

a Two balls are selected at random without replacement. Show the sample space on a tree diagram. Find the probability that at least one red ball was selected.

b Repeat if the balls are selected with replacement.

Solution

a The tree diagram is shown. Since there are different numbers of red and black balls, we need to write the different probabilities on each line at each stage.



At the first stage, the probabilities are $\frac{6}{10}$ for red and $\frac{4}{10}$ for black. If red is selected first, that leaves 5 red and 4 black balls in the urn. So the probabilities for red and black at the second stage are $\frac{5}{9}$ and $\frac{4}{9}$ respectively. Multiplying the probabilities along each branch gives $\frac{30}{90} = \frac{1}{3}$ for the chance of selecting two red balls.

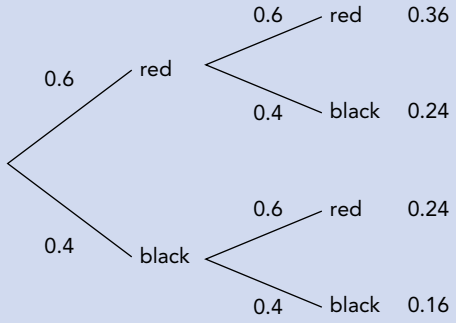
The probability that at least one red ball was selected can be found by adding the probabilities of the first three branches: $\frac{30}{90} + \frac{24}{90} + \frac{24}{90} = \frac{78}{90} = \frac{13}{15}$. Alternatively, we can find the probability of selecting two black balls, $\frac{12}{90} = \frac{2}{15}$, and subtract this from 1 to get the probability of the complementary event, $1 - \frac{2}{15} = \frac{13}{15}$.

If you are working with fractions, it is usually convenient not to simplify each so that the probability for each branch has the same denominator – in this case 90. Of course, we simplify fractions for our final answers.

ISBN 978-1-107-65599-7

b If the balls are selected 'with replacement', the first ball is replaced before the second is selected. Therefore, the probabilities are the same at each stage of the selection. The results are shown in the tree diagram on the previous page. Since there are 10 balls at each stage, it is convenient to work with decimals.

The probability that at least one red ball was selected can be found as before by adding the probabilities of the first three branches: 0.36 + 0.24 + 0.24 = 0.64. Alternatively, we note that we need the probability of the complementary event to 'both balls selected were black', the bottom branch: 1 – 0.16 = 0.64



Note carefully the difference between:

- 'selection without replacement' – after the first ball has been selected, it is not put back before the second ball is selected; the probabilities are different at each stage.
- 'selection with replacement' – after the first ball has been selected, it is replaced before the second ball is selected; the probabilities are the same at each stage.

Key ideas

- The sample space for an experiment that has several stages can be shown using a tree diagram.
- In some cases, the branches are (or can be assumed to be) equally likely.
- In other cases, we write the probability on each line and find the probability of a branch by multiplying probabilities along the branch.
- In selection without replacement, the probabilities are different at each stage.
- In selection with replacement, the probabilities are the same at each stage.

Exercise 3E

1 A bag contains six balls: three red, two white and one yellow. Two balls are selected one after the other, at random and without replacement.

a Show the sample space on a tree diagram (note that there is only one yellow ball).

b Write the correct probabilities on each line and calculate the probabilities for each branch. Check that they sum to 1.

c Find the probability that the yellow ball was selected.

d Find the probability that both balls selected are the same colour.

e Find the probability that both balls selected are of different colours.

2 In Example 8, having a family of three children, can we consider the selection to be carried out 'with replacement' or 'without replacement'? Explain your answer carefully.

ISBN 978-1-107-65599-7

3 John Graunt was the first person to find that, of births, boys are more likely than girls. His data collected during the 1600s in London showed that there were approximately 14 boys born for each 13 girls. In present-day Australia, the probability that a newborn child is a boy is around 0.52. Modify the tree diagram in Example 7 using the knowledge that P(boy) = 0.52. Find the probability that exactly two of the three children are boys.

4 Legionnaires' disease is caused by a bacterium that can sometimes be found in airconditioning systems. If you are over 50, the probability of dying from this disease is about 10%. Three people are brought into a hospital ward suffering from Legionnaires' disease.

a Show the sample space on a tree diagram (use S = survive and D = die).

b Write the correct probabilities on each line and calculate the probability for each branch of the tree.

c Why should the probabilities of the branches sum to 1? Check that they do in this case.

5 Use your tree diagram from the previous question to find the probabilities of these events:

a None of the patients die.

b Exactly one of the patients dies.

c At least one of the patients dies.

ISBN 978-1-107-65599-7

Enrichment

Can probability be useful in quality control?

6 Probability is an important tool in industrial quality control. If a machine is producing electronic components, it is important to know what percentage of these components is faulty. As it isn't possible to test all the components individually, we only test a sample of them. Here is an example of how calculation of probabilities is used in the quality control process.

Information from previous testing is that 2% of the components produced by the machine are faulty. We select three components randomly from the machine's output during the last hour and test each of them.

a Show the possibilities on a three-stage tree diagram (use F = faulty and W = working).

b Write the probabilities on each line and use them to calculate the probability of each branch.

c What is the probability that all three components tested are faulty? What is the probability that two of them are faulty? What is the probability that only one is faulty? What is the probability that they are all working?

d Imagine that you do the testing and find that two of the three selected components are faulty. What do you think you might conclude from this?

e What if you found that exactly one component was faulty?

ISBN 978-1-107-65599-7

Chapter summary

Brief review of probability

- Definition of experiment, sample space and event
- Probability as a numerical value between 0 and 1
- Probability estimated using frequency, belief or modelling
- Equally likely outcomes all have the same probability
- Definition of disjoint (mutually exclusive) and complementary events
- 'A and B' means that both A and B occur
- 'A or B' means that at least one of A and B occur ('inclusive or'), or sometimes that exactly one of A and B occurs ('exclusive or')
- Venn diagrams and two-way tables can be used as visual representations.

Pairs of outcomes

- The outcomes from two-stage experiments can be represented in a two-way table
- Collected data on two variables can be presented in a two-way table
- Probabilities of events (single or combined) from the information in the table.

Conditional language and applications

- Conditional probabilities are based on a restricted reference group
- In a two-way table, conditional probabilities are useful for comparing groups
- Phrases such as 'if ...', 'given that ...' and 'knowing that ...' imply conditional probabilities.

The concept of independence

- Ideas about independence are influenced by the way the word is used in English
- In probability, A and B are independent if and only if $P(A \text{ and } B) = P(A) \times P(B)$
- Independent events are shown on a Venn diagram with overlapping circles with the probabilities satisfying the equation for independence
- Data in a two-way table can be used to check for the independence of two events.

Two- and three- stage experiments

- The sample space for an experiment that has several stages can be shown using a tree diagram
- In some cases, the branches are (or can be assumed to be) equally likely
- In other cases, we write the probability on each line and find the probability of a branch by multiplying probabilities along the branch
- In selection without replacement the probabilities are different at each stage
- In selection with replacement the probabilities are the same at each stage.

Multiple-choice questions

1 A ten-sided die has faces numbered 1 to 10. Each face is equally likely to be rolled. The probability of rolling a prime number is:

A $\frac{1}{2}$ **B** $\frac{2}{5}$

C $\frac{3}{10}$ **D** $\frac{7}{10}$

ISBN 978-1-107-65599-7

2 Three students are asked whether they are in favour of wearing a school uniform. The sample space is represented as {yyy, yyn, yny, nyy, nny, nyn, ynn, nnn}, where y = yes and n = no. The event A = 'two students said yes, the other said no'. Which of the following events is disjoint from A?

A All three had the same view
B The first student said yes
C A majority was in favour
D At least one student said no

3 Two fair dice are rolled. The probability that a total of 3 is obtained is

A $\frac{1}{6}$
B $\frac{1}{18}$
C $\frac{1}{12}$
D $\frac{1}{36}$

4 You have bought five tickets in a fire brigade lottery. There are three prizes in the lottery: first prize is a car, second prize is a trip to Surfers Paradise and third prize is a dishwasher. The complement to the event 'you win at least one of the prizes' is

A 'You win at most one of the prizes'
B 'You don't win the car'
C 'You don't win any of the three prizes'
D 'You win two or three of the prizes'

The following information relates to questions 5 and 6.

The students at a senior high school are classified by whether they are taking music or not, and whether they are taking drama or not. A student is selected at random to give a welcome speech at the beginning of the year.

	Drama	No drama	Totals
Music	18	4	22
No music	12	63	75
Totals	30	67	97

5 What is the probability that the student will be taking music or drama?

A $\frac{63}{97}$
B $\frac{18}{97}$
C $\frac{52}{97}$
D $\frac{34}{97}$

ISBN 978-1-107-65599-7

6 What is the probability that the student will be taking drama given that they are **not** taking music?

A $\frac{4}{67}$ **B** $\frac{18}{22}$

C $\frac{12}{75}$ **D** $\frac{63}{97}$

7 Two independent events, A and B, are such that P(A) = 0.5 and P(B) = 0.6. The probability that neither A nor B occurs is:

A 0.2 **B** 0.3

C 0.5 **D** 0.1

8 Two disjoint events, A and B, are such that P(A) = 0.4 and P(B) = 0.5. The probability that neither A nor B occurs is:

A 0.2 **B** 0.3

C 0.5 **D** 0.1

9 Three coins are tossed. Which of these probabilities is **not** a conditional probability?

A P(three heads if we know that at least one head occurred)

B P(first toss was a head, and the other two tosses were tails)

C P(a majority of heads knowing that at least one head turned up)

D P(all three tosses were the same given that the first was a head)

10 A jar contains 5 olives – 3 black and 2 green. If two olives are taken from the jar without looking, what is the probability that one of each colour is selected?

A 0.48 **B** 0.3

C 0.6 **D** 0.24

Short-answer questions

1 Pat, Chris and Kim are having lunch in the cafeteria. They decide to toss a coin each to see who will pay for the coffees. If two of them get heads and the other gets a tail (or vice versa) then the odd one out buys the coffees. If they all get the same result (that is, 3 heads or 3 tails come up) then they each buy their own coffees. What is the probability that they each buy their own coffees?

2 In 1788, the convicts of the First Fleet were transported to Australia in six ships. The following page shows a summary of the sexes of the convicts on each ship.

ISBN 978-1-107-65599-7

	Female	Male	Total
Alexander	0	212	212
Charlotte	19	88	107
Friendship	21	75	96
Lady Penrhyn	102	0	102
Prince of Wales	50	1	51
Scarborough	1	209	210
Totals	193	585	778

You have been asked to write a short report on one of the convicts using the historical evidence available. To be fair to all the people in the class, 'your' convict will be selected randomly.

a What is the probability that your convict was female and arrived on the *Friendship*?

b What is the probability that your convict was female or someone who arrived on the *Friendship*?

c What is the probability that your convict was female given that she was transported on the *Friendship*?

d What is the probability that your convict was transported on the *Friendship* if you know that she was female?

3 Lee, Jan and Alex have been chosen to represent their school in a debating team. They decide to choose who will be first speaker by putting their names in a hat and drawing out a name at random. They repeat the process with the remaining two names to choose who will be second speaker. Of course, the remaining person will speak third.

a Show the sample space for the two stages of selection on a tree diagram. Is this an example of sampling with replacement or sampling without replacement? How will you decide on the probability for each branch?

b What is the probability that Lee will be either the first or the second speaker?

c What is the probability that Alex will be the third speaker?

4 Colour-blindness is a 'sex-linked' characteristic: the genes that cause the condition are carried on the X-chromosome and so males are much more likely than females to suffer from it. In Australia, about 8% of males are colour-blind. Three Australian men are chosen at random.

a Use a tree diagram to show the sample space and the probabilities.

b What is the chance that none of the men is colour-blind?

c What is the chance that a majority of them is colour-blind?

d For a female to be colour-blind, she has to have the colour-blindness gene on both of her X-chromosomes. The two X-chromosomes are independent (one comes from her father, the other from her mother). What percentage of women will suffer from the condition?

5 Four electronic components operate independently in an alarm system, and each component has a probability of 0.025 of failing.

a 'In series' means that the system fails when any one of the components fails. If the components are connected in series, find the probability that the system works properly.

ISBN 978-1-107-65599-7

b 'In parallel' means that the system fails only if all individual components fail. If the components are connected in parallel, find the probability that the system works properly.

c Comment on the practical application of these results.

Extended-response questions

1 In the Vedic period in India, in the sixth century BCE, Sushruta wrote a famous text on medicine, which included a discussion of diet and its relationship to health. He examined the combinations that could be made with six different tastes (called *rasa*) – bitter, sour, salty, astringent, sweet and hot – taking them a certain number at a time.

a You are planning to cook a potato and cauliflower curry and rather than follow a modern recipe you want to go back to Sushruta's discussion. You decide to flavour the dish with two rasa selected from the six. In how many different ways could you flavour your curry? How could you show the possibilities visually?

b How many different ways are there to flavour your curry using three rasa?

c The number of ways of flavouring with four rasa is the same as the number of ways of flavouring with two rasa. Explain why this is so. (Hint: consider the rasa that you **don't** use in the dish.)

2 A *placebo* is a dummy medical treatment (e.g. a sugar pill) that may have a positive effect psychologically or even physically in treating a medical problem. One medical study examined the benefits of treating patients who had suffered a heart attack with a 'beta blocker' called propranolol. Patients were randomly given either propranolol or a placebo, and neither the patients nor the doctor treating them knew which pill they were getting. This way of carrying out such a study is called 'double blind'.

The study concluded that propranolol reduced the chances of dying from a follow-up heart attack.

a Why were around half of the patients 'treated' using the placebo instead of the propranolol?

b Why do you think the study was carried out 'double blind'?

c Of the 1094 patients assigned to take the placebo, 1037 said that they took their pills regularly while the rest said that they took their pills 'sporadically' (that is, they often forgot to take them). In the regular group, 31 people died during the following year, while in the sporadic group, 4 people died. Show this information in a two-way table.

d What was the probability of dying during the following year given that you took your (placebo) pills regularly? What was the probability of dying if you took your pills sporadically?

e Could you suggest any explanation for these results?

3 (You might like to try this question in a group of two or three.) What topic or idea in this chapter did you find most difficult to understand? Can you say why you found it difficult? Write a question about this topic – a short-response question, or a discussion question – that might help other people in your class (or a future class) understand the topic. Include an answer or suggested points for an answer.

ISBN 978-1-107-65599-7

CHAPTER

4

Boxplots

What you will learn

4-1 Quartiles, interquartile range and five-number summary

4-2 Boxplots and interpretation

4-3 Using boxplots to compare datasets

Do students run on stairs?

A large study recorded how long commuters took to go up or down the stairs between the platform and the pedestrian walkway at a busy city bus station. Over 5 weekdays, commuters were chosen at random and timed. The day was divided into Peak 1 (7.30 am–10.30 am), Off-Peak (10.30 am–4 pm) and Peak 2 (4 pm–7 pm). As well as direction (up or down the stairs), gender, and time to go up or down (in seconds), the type of commuter was recorded. In the overall dataset, there were 363 students – school or tertiary students.

So to investigate how quickly students use the stairs at this bus station, we might be interested in considering the time of day, comparisons between male and female students as well as whether the students were going up or down the stairs. Unless we split the data, that's $3 \times 2 \times 2 = 12$ groups to compare. And that's just the students!

The variable we are interested in is time in seconds. This is a quantitative variable and a continuous variable, so data on it might be measured very accurately and take many different values. We have seen how to use dotplots, stem-and-leaf plots and histograms to compare quantitative datasets, but we have also seen that choice of intervals can affect the plot appearance. It is not easy comparing 12 groups at once using any of these plots. Is there another type of plot we can use?

AUSTRALIAN CURRICULUM

Statistics and probability

- Data representation and interpretation
- Determine quartiles and interquartile range **(ACMSP248)**
- Construct and interpret box plots and use them to compare datasets **(ACMSP249)**
- Compare shapes of box plots to corresponding histograms and dot plots **(ACMSP250)**

ISBN 978-1-107-65599-7

Pre-test

1 Below is a sample of 28 observations on the time in seconds it takes people to complete a puzzle. They are arranged from smallest to largest.

110 115 115 120 125 125 125 125 130 130 130 130 130 130 130 135
135 135 135 135 135 140 140 140 140 140 145 145

a Draw a histogram of these data with bin length 5 seconds, starting at 110 seconds.

i Do you think the data can be described as symmetric or skewed to the left, skewed to the right or asymmetric?

ii Does the histogram appear to be bimodal or unimodal?

b Draw a histogram of these data with 6 bins, starting at 110 seconds.

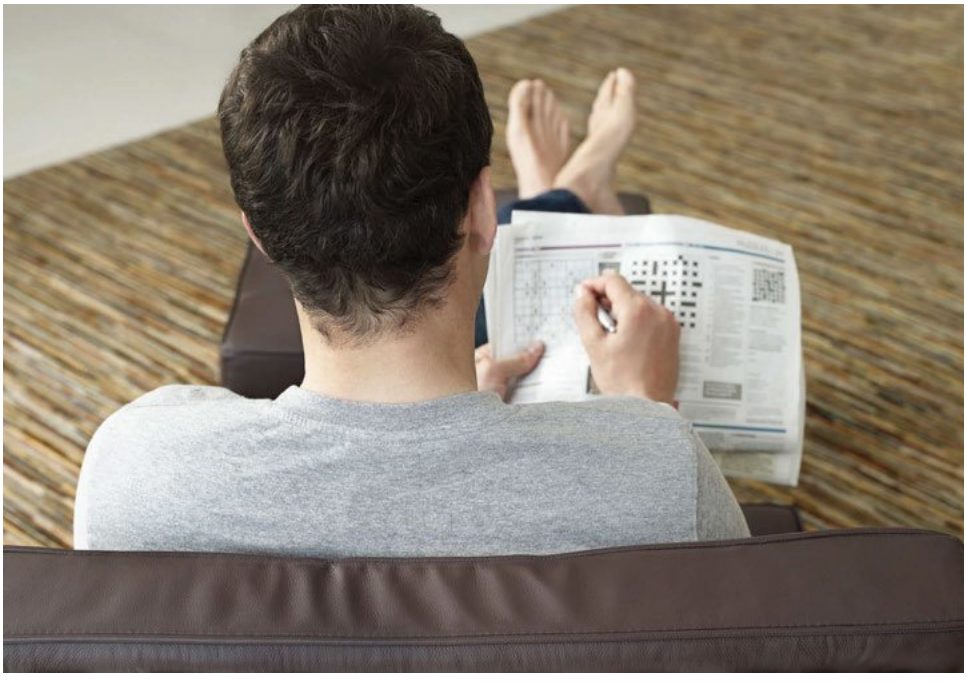
i Do you think the data can be described as symmetric or skewed to the left, skewed to the right or asymmetric?

ii Does the histogram appear to be bimodal or unimodal?

iii Are your answers different from part **a** above?

c Calculate the mean and range of the data and show that the median is 130 seconds.

d Is there anything surprising about the values of the mean and median of these data?



ISBN 978-1-107-65599-7

2 Below is another set of 28 times in seconds to complete the puzzle. They are arranged from smallest to largest.

110 115 116 120 125 125 126 128 130 130 132 132 133 134 134 135
135 136 136 137 138 140 141 143 143 144 145 145

a Show that a histogram of these data with bin length 5 seconds, starting at 110 seconds is exactly the same as in question **1 a**.

b Draw a histogram of these data with 6 bins, starting at 110 seconds.

i Do you think the data can be described as symmetric or skewed to the left, skewed to the right or asymmetric?

ii Does the histogram appear to be bimodal or unimodal?

c Calculate the mean and range of the data and show that the median is 134 seconds.

d Comment on the data using the summary statistics and the plot(s).

e The dataset in question 1 was obtained from this dataset. How?

3 Below are data of the amount of protein in grams per 100 grams of cereals from two manufacturers.

Manufacturer A

15.0	15.0	11.6	6.7	6.7	6.0	7.8	7.8	7.8	7.1	7.1
5.4	5.4	5.4	6.1	6.4	6.4	8.1	8.1	7.1	9.4	10.6
9.1	21.9	23.1	21.9	6.7	6.7	6.7	6.7	19.7	19.7	19.9
8.9	8.9	9.5	9.5	9.5						

Manufacturer B

11.80	9.60	9.60	9.60	11.90	9.00	14.90	8.40	8.40	9.80	9.80
11.20	11.20	9.00	9.40	9.40	10.10	9.60	9.19			

a Draw histograms of these data, starting at 4 and with bin length of 4 g per 100 g.

b Calculate the means, medians and ranges of these data.

c Use the summary statistics and plots to compare these two datasets.

Terms you will learn

box-and-whiskers plot
boxplot
extrapolation
first quartile
five-number summary
hinges
interquartile range
lower quartile
quartile
theoretical or population median
third quartile
upper quartile
whiskers

ISBN 978-1-107-65599-7

4-1 Quartiles, interquartile range and five-number summary

We have seen how useful the median is as a measure of location or centre of quantitative data. Half the data are less than it, and half are greater. If observations are changed, added or removed from the dataset, the data median may change, but it won't change if the values of the largest (or smallest) observations are changed but are still the largest (or smallest).

When the data are a random sample of a general situation or population, the data median estimates a special quantity of the general situation or population. A random observation from the general situation or population is equally-likely to be greater or less than this special quantity. This special quantity is called the **theoretical** or **population median**.

The data mean or average is also an important measure of location, or centre, of the data. The data mean and the data median are examples of summary statistics.

Other summary statistics are the data minimum, maximum and the data range. The data range gives a measure of how spread out the data are, but of course, this is a measure of the overall spread of the data. Two datasets may have completely different ranges, but only because a very small number – possibly as small as one – of observations are very different.

Two other very useful summary statistics for quantitative data are the **quartiles**. Together with the median, the quartiles divide the data into four groups with equal numbers of observations in each. The median has half the observations less than it and the **lower quartile** is the median of this lower group. The **upper quartile** is the median of the observations greater than the median. These are useful in themselves, but their difference = upper quartile – lower quartile is also a useful measure of spread as it gives the spread of the middle half of the data; it is called the **interquartile range**.

Theoretical or population median: A special quantity for a general situation or population … *see glossary*

Quartiles: These divide a dataset (including the data median) into four groups with equal numbers of observations in each

Lower quartile: This has a quarter of the observations less than it … *see glossary*

Upper quartile: This has a quarter of the observations greater than it … *see glossary*

Interquartile range: The difference (upper quartile) – (lower quartile)

LET'S START Judging distances

In section 1-1 are the data, with dotplots and a back-to-back stem-and-leaf plot, of the guesses of a distance of 5 m by 25 males and 19 females who did not wear glasses or contact lenses. We have seen how useful a stem-and-leaf plot is for finding the data median. Let's use the back-to-back stem-and-leaf plot and the data to find the median and quartiles for the data for the males.

From the plot, the 13th observation for the males is the second 5.2 m. Looking at the data, this is 5.21 m if we want the original accuracy. This divides the data into two groups of 12. So the median of the lower group is halfway between the 6th and 7th observations; this is 4.75 m. Again, if we want the original accuracy, the 6th and 7th observations are 4.61 and 4.95, so halfway is 4.78 m. This is the lower quartile. In the upper group, we want halfway between the 6th and 7th observations from the top; these are the 19th and 20th observations from the bottom but it's easier to work from the top. In the plot these are

ISBN 978-1-107-65599-7

HINT
If we have only the stem-and-leaf plot, then the accuracy of the median and quartiles depends on what's in the plot. But, as you can see, going to the accuracy of the original data is not making much difference.

both 5.6 m. If we want the original accuracy, they are 5.61 m and 5.66 m, so the upper quartile is 5.635 m.

Just using the stem-and-leaf plot, the median is 5.2 m, the lower quartile is 4.75 m and the upper quartile is 5.6 m. The middle half of the males' guesses are between 4.75 m and 5.6 m. The interquartile range is 0.85 m; this gives the range of the middle half of the data and is a useful measure of spread. The minimum male guess is 3.9 m and the maximum is 6.4 m.

Note that the median is about halfway between the lower and upper quartile. This is because these 25 observations are reasonably symmetric. We see more about this in section 4-2.

The minimum, lower quartile, median, upper quartile and the maximum of a dataset are called the **five-number summary**. Together they divide the ordered data into quarters with equal number of observations in each quarter.

HINT
Even with only 25 observations, finding the median and quartiles from the original data is difficult unless the data are ordered from smallest to largest. And even if we lose a decimal place (or more), obtaining the median and quartiles from a stem-and-leaf plot is usually accurate enough for practical purposes.

Key ideas

- For quantitative data, the lower quartile is the median of the data below the overall median, and the upper quartile is the median of the data above the overall median.
- The minimum, lower quartile, median, upper quartile and maximum are called the five-number summary. They divide the ordered data into quarters of equal frequency.
- The interquartile range = (upper quartile) – (lower quartile).

Five-number summary: The set of numbers consisting of the median, the upper and lower quartiles, and the minimum and maximum

ISBN 978-1-107-65599-7

Example 1: The Kiama blowhole

Below is a stem-and-leaf plot of the data on the 64 times in seconds between eruptions of the Kiama blowhole, as in Example 2 of Chapter 1.

Question: What are the quartiles and five-number summary for these data?

There are 64 observations so the median is halfway between the 32nd and the 33rd, which are both 28 seconds; the median is 28 seconds. There are 32 observations less than the median, so the lower quartile is halfway between the 16th and the 17th, which are 14 and 15 seconds; the lower quartile is 14.5 seconds. The upper quartile is halfway between the 16th and the 17th observations from the top, which are both 60 seconds; the upper quartile is 60 seconds. The five-number summary is: minimum = 7 seconds, lower quartile = 14.5 seconds, median = 28 seconds, upper quartile = 60 seconds, maximum = 169 seconds.

Leaf unit = 1.0

0	78888899
1	00001124567778888
2	1556778899
3	45667
4	027
5	1456
6	001189
7	37
8	23379
9	15
10	
11	
12	
13	
14	6
15	
16	9

HINT
Note that the lowest quarter of the data are squashed between 7 s and 14.5 s, the next quarter spread a bit more between 14.5 s and 28 s, the next quarter spread out between 28 s and 60 s, and the top quarter of the data spread out between 60 s and 169 s! The spread of the data increases greatly from smallest to largest observations. This is a classic case of skewed to the right.

Exercise 4A

1 In section 1-1, the back-to-back stem-and-leaf plot also gives the guesses of a distance of 5 m by 19 females who did not wear glasses or contact lenses.

a Use this plot to obtain the median and quartiles of the 19 observations, and give the interquartile range.

b Compare the five-number summary with that of the 25 males.

2 In question 4 of Exercise 1A, the lengths of Indie and Alternative Rock songs in seconds are given, obtained from a top 100 chart in one year.

a Use the back-to-back stem-and-leaf plots you drew for that question to obtain the medians and quartiles for both types of songs.

b The values in the stem-and-leaf plots are rounded down to the 10 s below. Refer to the original data to get better accuracy for the medians and quartiles. Does it make much difference?

c Use the five-number summaries to comment on the lengths of the two types of songs.

ISBN 978-1-107-65599-7

3 Question 5 of Exercise 1A gives data from an experiment to compare people's perception of short periods of time when reading to when not reading. The period of time considered in the data given was 10 s. Consider the differences given by (guess when reading) – (guess when not reading).

a Using stem-and-leaf plots (or otherwise) obtain the medians and quartiles of the data to the nearest second below (that is, ignore the decimal place) for both the males and females.

b Use the five-number summaries to comment on the effect of reading in guessing 10 s for males and females.

4 Another way of thinking of the median of a dataset with n observations is to take the $\frac{(n+1)}{2}$ observation, using **extrapolation** if necessary. For example, if we have 10 observations, the median is the $\frac{11}{2}$ = 5.5th observation, so we take halfway between the 5th and the 6th. If we have 11 observations, the median is the $\frac{12}{2}$ = 6th observation.

a Continuing like this, the lower or **first quartile** can be thought of as the $\frac{(n+1)}{4}$th observation from the bottom, using extrapolation if necessary. The upper or **third quartile** is then the $\frac{3(n+1)}{4}$th observation from the bottom. Consider sample sizes of 8, 9, 10 and 11 to demonstrate that the third quartile is the $\frac{(n+1)}{4}$th from the top observation.

b Show that the first data quartile obtained by this method for the Kiama blowhole data in Example 1 is 14.25 s, while the third quartile is still 60 s.

Extrapolation: Refers to ways of approximating values between given or known values

First quartile: Another name for lower quartile

Third quartile: Another name for upper quartile

Enrichment

What estimate of Earth's density should we give from the Cavendish data?

5 In section 1-2, the 1798 Cavendish data on the density of Earth (as a multiple of the density of water) are given. Short-answer question 2c of Chapter 2, asks you to obtain the mean, median and range of these data.

a Obtain the quartiles of these data as the medians of the lower and upper halves of the data.

b Obtain the first and third quartiles of these data using the method of question 4 above.

c This dataset is a fairly small one with only 29 observations. Is there much difference between these two methods for these data? Make up a small dataset – for example, of 20 observations – where there is quite a difference between the lower quartile obtained as the median of the bottom half and the first quartile obtained as the $\frac{(n+1)}{4}$th observation from the bottom.

d There are different estimates of the density of Earth we could give based on these data. We could give the data mean or the data median of the original data, or we could omit the smallest observation (4.07) and give the data mean or median.

i What are the data mean and median without the smallest observation?

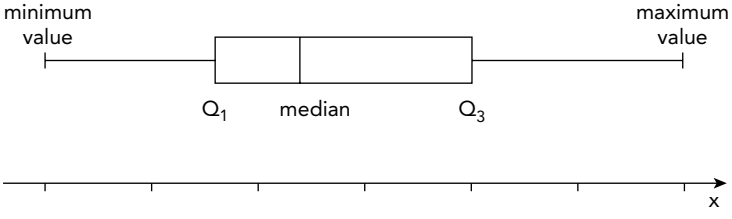
ii What are the quartiles (as medians of the bottom and top halves of the data) without the smallest observation?

ISBN 978-1-107-65599-7

4-2 Boxplots and interpretation

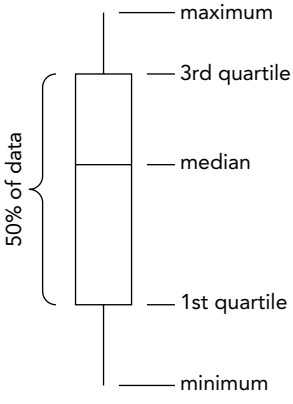
Boxplot of five-number summary

A plot called the **boxplot** is a useful way of presenting the five-number summary as shown below. In this sketch, Q_1 and Q_3 are the lower and upper quartiles respectively.



Boxplot: A graph for quantitative data (usually from a continuous variable) which divides the range of values of the data into intervals with a quarter of the data in each … *see glossary*

However, boxplots are almost always presented vertically these days because this is how computer packages produce them. Below is a vertical boxplot.



HINT
The reason the upper quartile is sometimes called the third quartile is because the median is the second quartile; it marks the end of the second quarter of data.

Hinges: The edges of the box – the lower and upper quartiles

Box-and-whiskers plot: The original and full name for the boxplot

The edges of the box, marking the lower and upper quartiles, are sometimes called the **hinges**. The full name of the boxplot is **box-and-whiskers plot**, but we almost always just call them boxplots now. They were introduced by John Tukey (1915–2000), an outstanding statistician who invented many modern statistical methods and always emphasised the importance of looking at and exploring data.

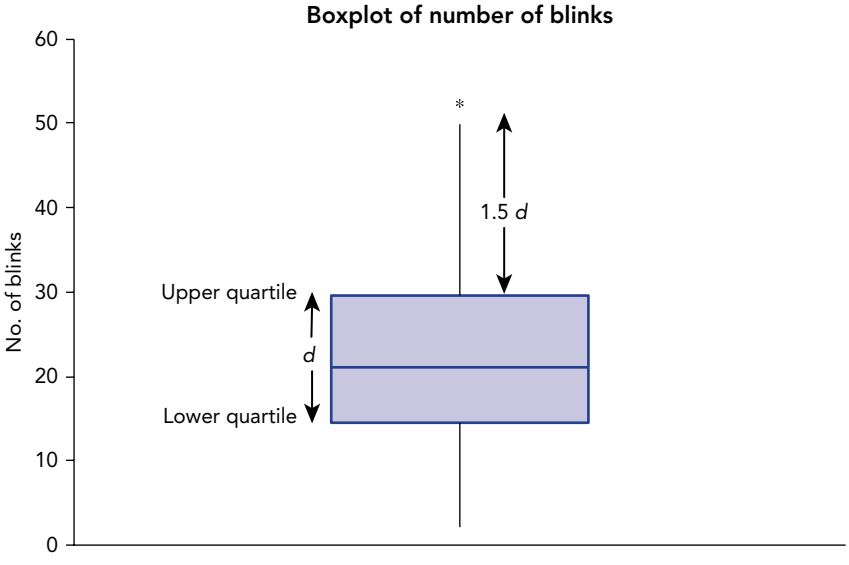
CAUTION
In the horizontal boxplot above, note the small bars at the end of the whiskers. The ones in this plot are small enough not to be a distraction, but there is no need for them, and larger ones can distort the picture.

Better boxplots

The above diagram of the five-number summary is the simplest form of boxplot, but a problem is that we don't know how far the minimum and the maximum are from the rest of the data. Boxplots more often used in statistics draw the **whiskers** from the box to the data points that are within a certain distance from the edges of the box, and marks the data points that are outside this distance by a star or an asterisk (*). The distance most commonly used is 1.5 times the interquartile range. We will use this here.

Whiskers: The lines extending from the edges of the box

The boxplot below is for a dataset from an experiment that investigated the number of times people blinked in a minute. In the diagram, d is the value of the interquartile range. The whiskers are drawn out to the last observation that is within $1.5d$ from the edge of the box. Observations further away than $1.5d$ from the box are marked by *.



In this dataset, there is only one observation marked by a star, so the simpler five-number summary does not hide much information for these data, but in general this better version of the boxplot provides valuable information. This is especially so when we are using boxplots to compare a number of datasets or groups.

Boxplot are often very good representations of skewness as well as location as the median, spread as the interquartile distance, and overall range. What can't we see in a boxplot? We can't see how many observations there are and we can't see if there are groups in the data – for example, bimodality.

HINT
The length of the box is the interquartile distance. The width doesn't have any meaning, and just depends on how many boxplots there are.

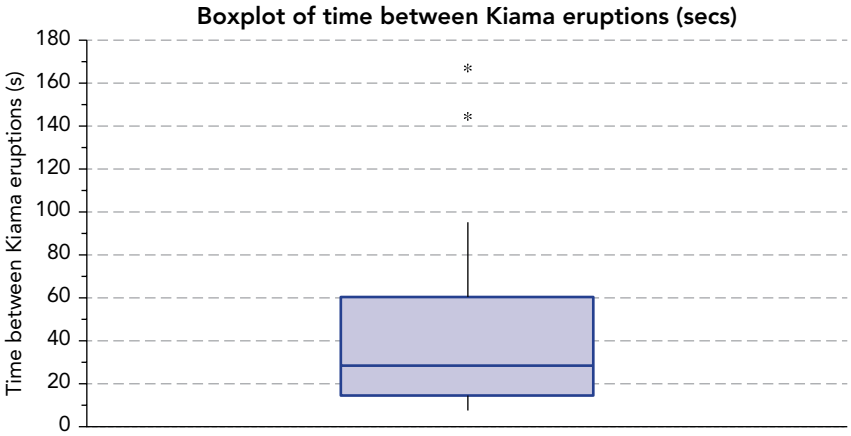
CAUTION
We should not use boxplots for small datasets. Guidelines are sometimes given, but we can see that 20 or more observations per boxplot is reasonable, and that a boxplot for fewer than 12 observations could be misleading.

LET'S START How skewed data looks in a boxplot

In the boxplot above of the number of blinks per minute in an experiment, notice how the whisker to the larger numbers of blinks is longer than the whisker to the smaller numbers of blinks. There is also another large observation outside the upper whisker. Also in the box, the distance from the median to the upper quartile is slightly longer than the distance between the median and lower quartile. There are 100 observations in this dataset. So the boxplot shows that the upper 50 observations are a little more spread out than the lower 50. The dataset is skewed to the right – not strongly but we can see it in the boxplot.

HINT
Even when the boxplot is vertical as above, we still call the upper whisker the right-hand whisker, and the lower whisker the left-hand whisker.

What happens for a very skewed dataset? Let's draw the boxplot for the Kiama blowhole data. From Example 1, we have the interquartile range = 60 – 14.5 = 45.5. So 1.5 times this is 68.25 seconds. The upper whisker goes to the last observation less than 60 + 68.25 = 128.25 seconds. The lower whisker goes to the last observation greater than 14.5 – 68.25 which is negative! So the lower whisker goes to the minimum which is 7 seconds. We need to look at the original data in section 1-2 to see that there are two observations larger than 128.25 seconds: 146 seconds and 169 seconds. And the last observation less than 128.25 seconds is 95 seconds. The boxplot is below. Gridlines have been added so you can see the values easily.



HINT
Note that the gridlines are faint so as not to distract from the picture of the data.

In Example 1, we could tell just from the five-number summary that the data became increasingly spread out as we moved through the quarters. The boxplot shows this very clearly. The first 25% of the observations are completely squashed up in the tiny lower whisker. Then the next 25% of observations are about three times less squashed in the lower section of the box. Then the third quarter of the observations are spread out much more in the upper section of the box. And the top 25% of the data are spread out over the upper whisker plus there are two observations much further away. The data are very skewed to the right.

HINT
Unlike dotplots, stem-and-leaf plots and histograms, a boxplot for a dataset does not change – its appearance does not depend on choice of intervals or starting point.

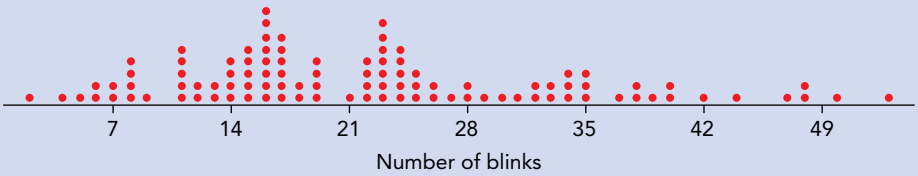
ISBN 978-1-107-65599-7

Key ideas

- The five-number summary can be represented by a simple boxplot, with the edges (or hinges) of the box marking the quartiles and the whiskers going out to the minimum and maximum.
- Better boxplot versions have the whiskers going out to the last observation within a chosen distance of the box. This distance is usually 1.5 times the interquartile range. Observations outside this distance are marked by a star or an asterisk (*).
- Boxplots show how the four quarters of data are clustered or spread, and if there are small or large observations very spread out.
- Boxplots quickly show location (median), spread and shape, and are constant (unchanging) for a dataset, but they do not show the number of observations nor sub-groups of data.

Example 2: How does a boxplot compare with other plots?

Below is a dotplot of the dataset on the numbers of blinks per minute that is plotted in the boxplot on page 98.



Question: How does the dotplot compare with the boxplot?

Both plots show how the data are skewed to the right. The boxplot tells us that the middle 50% of the data are between approximately 14 and just under 30. We can see in the dotplot that this is the central part of the data, although not easily that it's the middle 50%. The boxplot tells us that the data are more spread out above 30 than below approximately 15, and we can see that in the dotplot also. However, the dotplot indicates that there may be two groups of data and we can't see that in the boxplot.

Example 3: Skewed or asymmetric?

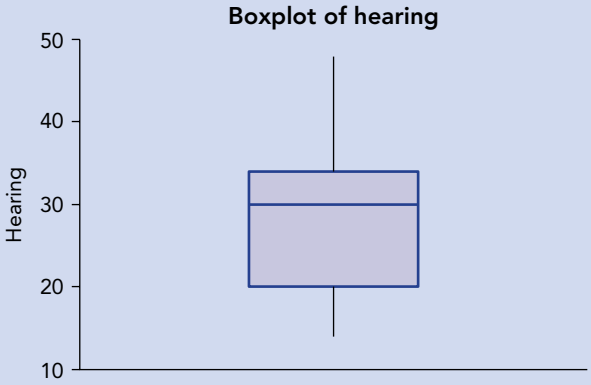
In a hearing test experiment, four lists each of 25 words were tested on a group of 24 people to investigate lists of words that can be used. Below is a boxplot of all the scores scaled to be out of 50.

Question: Are the data symmetric, asymmetric or skewed?

In the boxplot, the upper whisker is longer than the lower whisker, indicating some skewness to the right. But in the box, the lower part of the box is larger than the upper part, so that the data in the second quarter are more spread out than the data in the third quarter. The data are not symmetric, but based on the boxplot, we cannot really

ISBN 978-1-107-65599-7

say they are skewed in a particular direction. We can say they are asymmetric; we could say that the right-hand tail is longer than the left.



Example 4: How far away is too far?

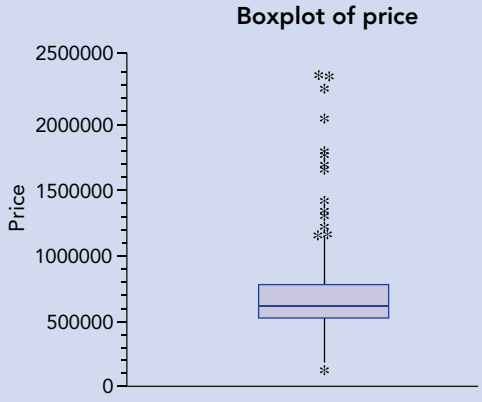
You have seen the word *outlier*. This is a name sometimes given to extreme values. Sometimes an extreme observation needs to be investigated in case it's a mistake or it is an observation that doesn't belong with the rest of the data. We also know that the data mean is more affected by outliers than the data median.

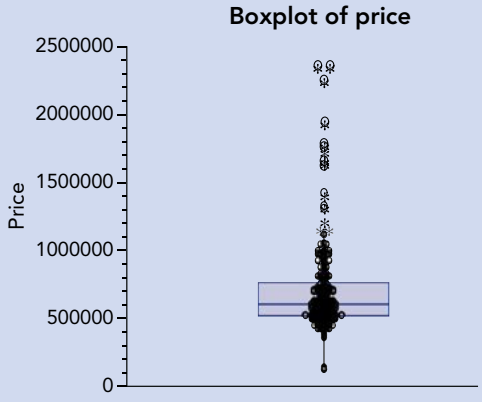
Question: Are they outliers?

On the next page are two boxplots of the same dataset of selling prices of houses in an Australian city in 2007. The second boxplot includes a dotplot of the individual observations. The data are very skewed to the right, with a few very large values, but we can see that this is a natural 'tailing off' of the data – typical of real-estate prices. Of course, to investigate these prices, we would want to know other information such as suburb, size of land and so on. But the whole dataset represents the overall region, and the largest observations are part of the data.

HINT
Most reports of real-estate prices quote median prices – you can see why! Most real-estate data are skewed to the right.

ISBN 978-1-107-65599-7





CAUTION
Think of what horizontal bars on the ends of the whiskers in these plots would look like – they distract the eye from the flow of the data.

Exercise 4B

1 In question 3 of Exercise 1B, the scaled scores out of 50 are given for 30 people assisting in the checking of a list of words to be used for a hearing test, as follows:

28 24 32 30 34 30 36 32 48 32 32 38 32 40 28 48 34
28 40 18 20 26 36 40 20 16 38 20 34 30

a Draw a boxplot of these scores.

b Use the boxplot to describe the data, including the shape of the data.

c Compare the boxplot with the histogram obtained in question 3 of Exercise 1B.

2 In question 4 of Exercise 1A, the lengths of Indie and Alternative rock songs in seconds are given, obtained from a top 100 chart in one year:

Lengths of Indie songs

219 200 204 199 203 275 226 186 278 237 208 200 232
250 190 288 233 227 233 226 197 332 239 234 192 182
226 255 130 257 219 248 221 216 254 258

Lengths of Alternative rock songs

499 191 201 200 485 181 406 258 326 298 197 213 181 152 188 220
252 213 362 275 234 250 194

a Use the five-number summaries obtained in question 2a of Exercise 4A (and the data in the stem-and-leaf plots in Exercise 1A) as well as this original data to draw boxplots on the same scale for these datasets.

b Use the boxplots to describe the data.

c Compare the boxplots with the stem-and-leaf plots drawn in question 4 of Exercise 1A.

3 Refer to question 3 of Exercise 1A and Let's start of section 1-3 on the lengths of New Zealand South Island rivers.

- **a** Use the stem-and-leaf plots obtained in Exercise 1A and the data there to draw boxplots of the lengths of the rivers on the same scale.
- **b** Use the boxplots to describe the data.
- **c** Compare the boxplots with the dotplots and the stem-and-leaf plots drawn in question 3 of Exercise 1A.

Enrichment

Reaction times

4 Below are the reaction times in seconds, correct to two decimal places, for the dataset of Example 4 in Chapters 1 and 2, in groups of males (M) and females (F) and whether they had drunk coffee in the last hour (no, yes). The data are ordered from smallest to largest in each group to help you.

Reaction_F-No

0.17 0.17 0.17 0.18 0.18 0.19 0.19 0.19 0.20 0.20 0.20

Reaction_F-Yes

0.14 0.17 0.18 0.19 0.19 0.19 0.19 0.19 0.19 0.19 0.20 0.20 0.20 0.20
0.22 0.23 0.24

Reaction_M-No

0.17 0.17 0.18 0.18 0.18 0.19 0.19 0.19 0.19 0.20 0.20 0.20 0.21 0.21
0.21 0.21 0.22 0.22 0.22 0.23 0.23

Reaction_M-Yes

0.16 0.16 0.17 0.17 0.17 0.17 0.17 0.17 0.18 0.18 0.18 0.19 0.19 0.19
0.19 0.19 0.20 0.20 0.20 0.20 0.21 0.22 0.22 0.23 0.23 0.24 0.25

- **a** Draw boxplots of these four datasets on the same scale.
- **b** Where is the line for the median in the boxplot for females who had drunk coffee?
- **c** Comment on the boxplots for the females. Give at least one reason why the boxplots look as they do.
- **d** Use the boxplots to comment on the data for the males.
- **e** Comment on the boxplots compared with the dotplots and histograms in Example 4 of Chapter 1.

ISBN 978-1-107-65599-7

5 In a survey of patrons of cafés, random samples of people over a number of days were observed and the time they spent over a cup of coffee was recorded in minutes. The times recorded are given below, grouped by gender and whether the coffee was large or small/medium (called small). Each data group is ordered from smallest to largest.

Time spent_Female-Large

5.0 5.0 5.0 6.0 7.5 10.0 10.0 12.5 12.5 12.5 15.0 15.0 15.0 15.0 15.0
17.5 17.5 20.0 20.0 25.0 30.0 37.5

Time spent_Female-Small

5.0 7.5 10.0 10.0 10.0 10.0 11.0 12.5 15.0 15.0 15.0 15.0 15.0 15.0 15.0
15.0 15.0 15.0 15.0 17.5 17.5 20.0 20.0 20.0 20.0 20.0 20.0 20.0 20.0
20.0 21.0 22.5 22.5 22.5 25.0 25.0 25.0 30.0

Time spent_Male-Large

1.0 2.5 2.5 5.0 5.0 5.0 7.5 7.5 7.5 10.0 10.0 10.0 10.0 10.0 10.0 10.0
10.0 10.0 11.0 12.5 12.5 12.5 12.5 12.5 15.0 15.0 15.0 15.0 15.0 15.0
15.0 15.0 15.0 17.5 20.0 20.0 20.0 24.0

Time spent_Male-Small

1.0 1.0 2.5 2.5 2.5 5.0 7.5 7.5 9.0 10.0 10.0 10.0 10.0 10.0 10.0 10.0
10.0 10.0 10.0 10.0 10.0 10.0 10.0 12.5 15.0 15.0 15.0 15.0 15.0 15.0
17.5 20.0 20.0

a Draw boxplots for these four groups on the same scale.

b Use the boxplots to comment on the data.

c To what accuracy have the data been recorded? Do you think there may have been problems in recording the data?

ISBN 978-1-107-65599-7

4-3 Using boxplots to compare datasets

Section 4-2 focuses on constructing and interpreting boxplots, and what they can and can't do. But a number of examples and exercises in section 4-2 already demonstrate how easy it is to compare boxplots of groups on the same scale.

As you can see in some of the examples and exercises in section 4-2, boxplots of small data groups and/or of data with many equal values may not present as much information as other plots. For example, dotplots clearly show if there are many repeats of the same value such as when the recording accuracy has not been good. For small datasets, it is difficult for boxplots to show much and they can be misleading.

Their advantages are that there is only one boxplot for a dataset, and they are excellent for an overall comparison across a number – even many – groups. Their advantages are greatest for medium to large datasets. For very large datasets, for example in the thousands, either histograms or boxplots tend to give good representations of data.

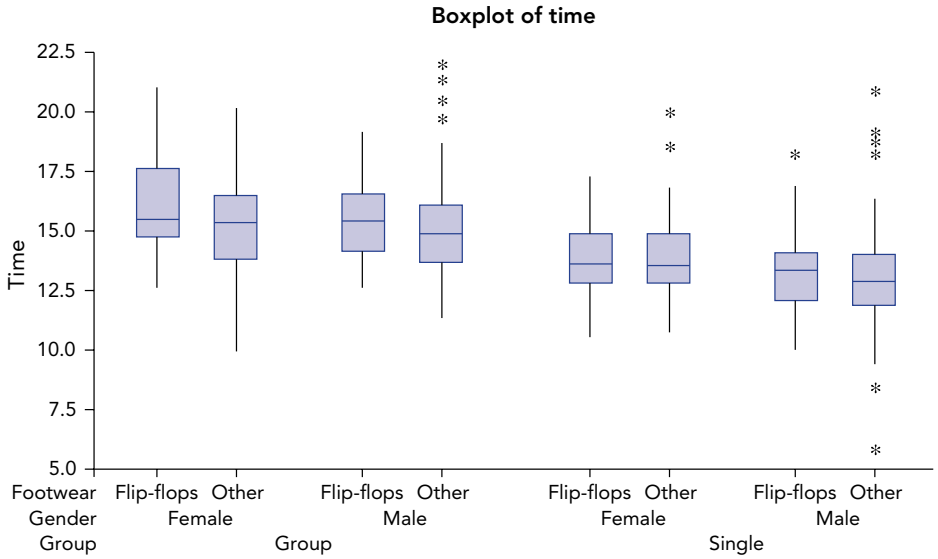
CAUTION
Remember that boxplots, like histograms and stem-and-leaf plots, are for quantitative data.

LET'S START How fast do pedestrians walk?

The case study described at the beginning of Chapter 2 and in question 6 of Exercise 2B, recorded times in seconds for pedestrians walking a 15-metre section of city footpath that was used by many people as a thoroughfare. Subjects were chosen 'randomly' and any who exited or stopped in the section were not included. If a group walked by, the time for the group was recorded and a pedestrian classified as 'group' was chosen at random from the group and was the only person recorded for that group. As well as gender, the type of footwear (flip-flops or other) was recorded.

HINT
There is no best plot of any dataset. Which plot gives a good representation of the data depends on the data, the context and the purpose, as well the sample size, features and shape of the data.

Below is a boxplot of the times for the datasets formed by male or female, single or group, and flip-flops or other footwear.



ISBN 978-1-107-65599-7

Question: What does the boxplot tell us?

There is a lot of overlap in times (and therefore speeds) across all datasets, so that variation within datasets tends to swamp differences between them. However, those walking in groups generally tend to be slower than those walking alone. We can see that because the medians are higher for those walking in groups than those walking alone, and the interquartile ranges are not very different. There are some much slower males walking in groups wearing 'Other' footwear, some much faster and slower males walking alone wearing 'Other' footwear, and a few slower females and males walking alone in 'Other' footwear (females) and flip-flops (male). The data are skewed to the right for females wearing flip-flops (for singles and groups), for males wearing 'Other' footwear in groups and for females wearing 'Other' footwear walking singly. The other groups are either close to symmetric or are asymmetric – not having skewness in one direction.

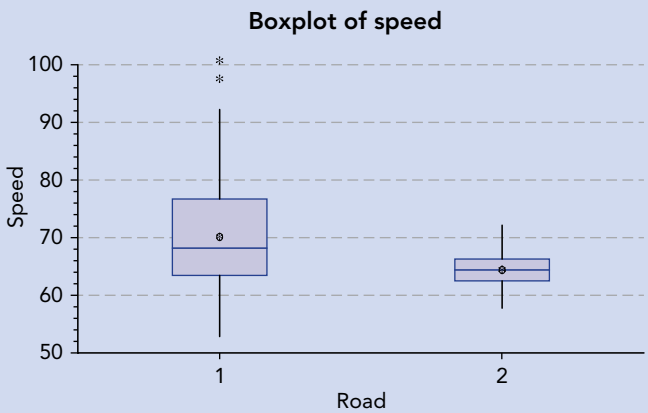
Looking at the medians, the males wearing flip-flops may be slightly slower in general than males wearing 'Other' footwear, and males may tend to be slightly faster than females in general. But there is a lot of variation across all types of pedestrians.

Key ideas

- Provided datasets are not too small, boxplots on the same scale are a good presentation for comparing a number of quantitative datasets. These are often subsets of a larger dataset split into groups by categorical variables.
- Boxplots can compare location in general and represented by medians, spread represented by interquartile range and overall range, skewness and extreme or unusual observations. They cannot tell us how many observations we have or if there is bimodality which could be sub-groups.

Example 5: What's the speed limit?

The boxplot shown is of the speeds of vehicles in km/h over 300 m on straight stretches of one of two roads, with no side roads or traffic lights. The road coded 1 is part of a highway but with a speed limit of 60 km/h because of its situation. Road coded 2 also has a speed limit of 60 km/h. The data means have also been marked on these boxplots. This is always an option, but can be distracting if comparing many boxplots, especially if they are very different in shape. Gridlines help us in comparing the boxplots with values of speeds.



ISBN 978-1-107-65599-7

Question: How does vehicle speed compare on the two roads?

For both roads, the median and average speeds are over 60 km/h. The median speed on road 1 is over 65 km/h and the average speed is almost 70 km/h! The speeds on road 1 are skewed to the right, with speeds reaching 90 km/h and two extreme speeds of approximately 100 km/h. The median and average speeds on road 2 are approximately 63 km/h and the speeds on road 2 are symmetric, ranging from approximately 56 to 71 km/h. The speeds on road 1 range from just over 50 up to 100 km/h, and the central 50% of speeds vary from about 62 to about 76 km/h.

So people tend to speed more on road 1, but the main contrast between the two roads is the variation, with the speeds on road 1 being highly variable compared with the speeds on road 2.

Exercise 4C

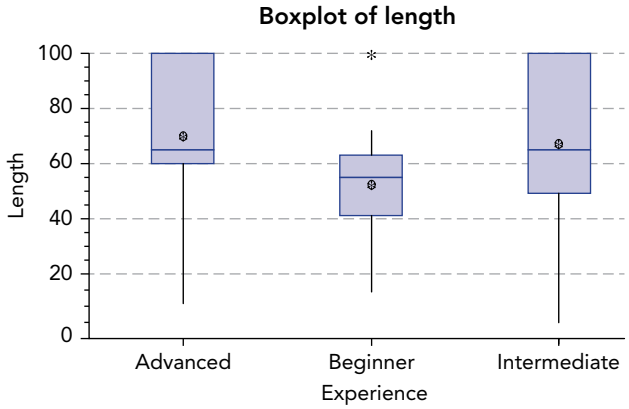
1 Extended-response question 8 in Chapter 2 reports a study on male cricketers. The results for each ball faced by batsmen of different levels of experience (beginner, intermediate and advanced) were recorded. Of those balls that were hit, the ball was not fielded and the length (in m) and the angle (in degrees clockwise from the front) were recorded. A total of 90 balls were faced by beginner and advanced batsmen and 60 by intermediate batsmen. Below is a boxplot of the lengths hit by batsmen of different experience, including symbols for the average lengths.

a Using only the boxplots, comment on the three sets of lengths.

b Comment on how the average length compares with the median length in each case.

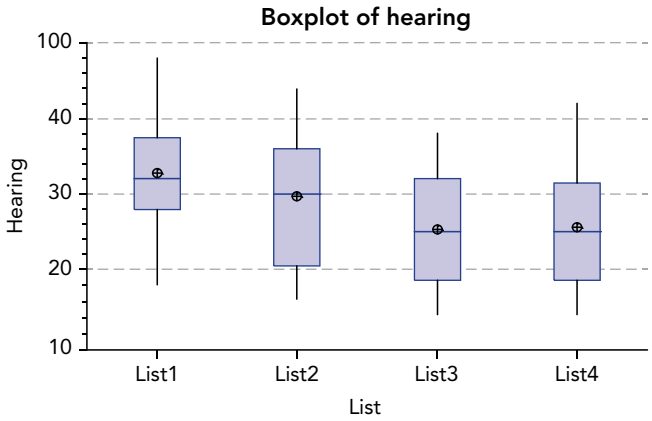
c Refer to the histograms in Extended-response question 8 in Chapter 2 to explain why there are no upper whiskers in two of the boxplots. This also helps to explain part **b** above.

d Why is there no point in drawing boxplots of the angles hit by the batsmen?



ISBN 978-1-107-65599-7

2 Example 3 above refers to an experiment in which four lists each of 25 words were tested on a group of 24 people to investigate lists of words that can be used in hearing tests. Below are boxplots of the scaled scores out of 50 for each list, including symbols for the average scores.

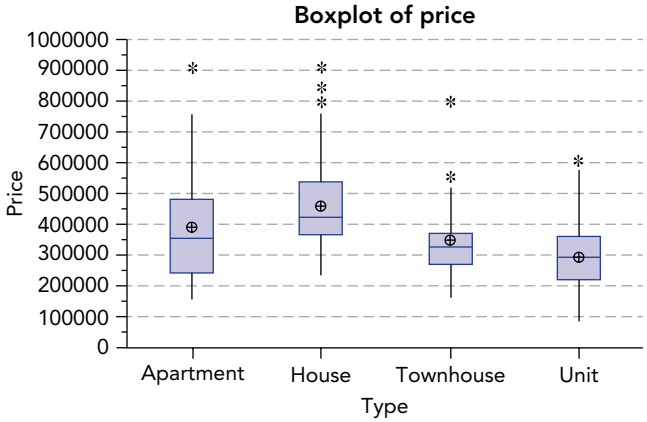


a Use the boxplots to compare the lists.

b The same group of 24 people was used to test each list. What would be of interest to investigate?

3 Data were collected from a real-estate website on prices of various types of properties in an Australian city. These data are presented in the boxplots below, which include symbols for the average prices.

ISBN 978-1-107-65599-7

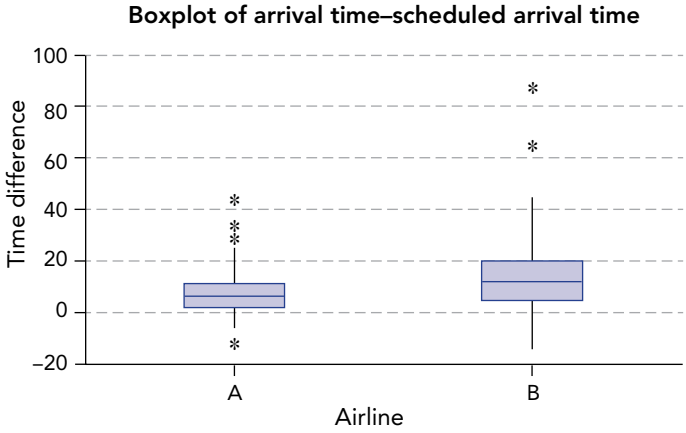


a Use the boxplots to comment on the prices of the different types of properties.

b Why are the averages greater than the medians for three of the four types of properties?

c For units, the average price is close to the median price. Is this surprising? Is there a clue in the boxplot as to why this might be?

4 The differences (actual arrival time – scheduled arrival time) in minutes were recorded for two airlines at the same airport for three days (Monday–Wednesday) for two weeks. The data are presented in the boxplots below.



Use the boxplots to comment on the data.

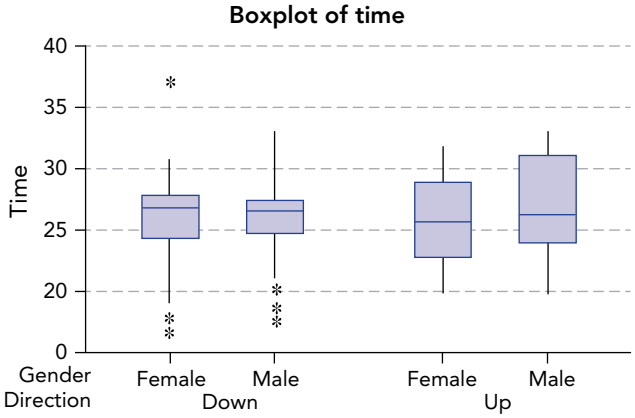
Enrichment

Do students run on stairs?

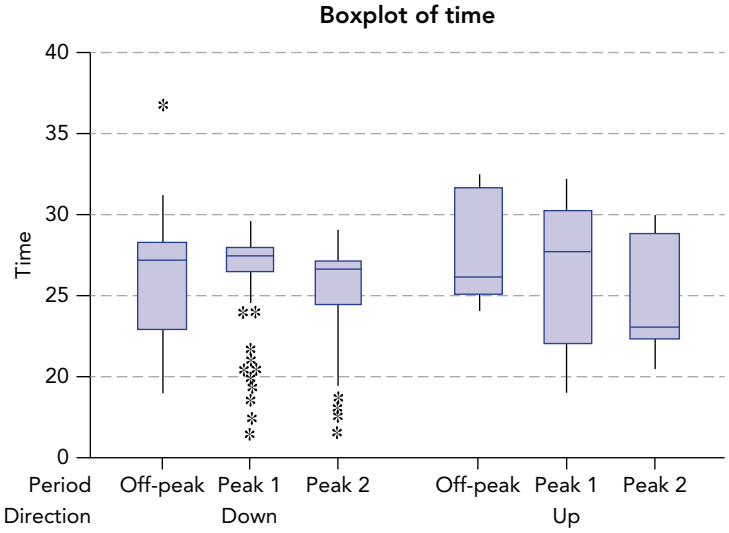
5 The beginning of this chapter referred to an investigation on how long, in seconds, commuters took to go up or down the stairs between the platform and the pedestrian walkway at a busy city bus station. Over 5 weekdays, commuters were chosen at random and timed. The day was divided into Peak 1 (7.30 am–10.30 am), Off-Peak (10.30 am–4 pm), Peak 2 (4 pm–7 pm). In the overall dataset, there were 363 students – school or tertiary students.

ISBN 978-1-107-65599-7

a The boxplots below present the times for the students according to their gender and direction. Use these boxplots to comment on the comparisons between male and female students going up or down the stairs.



b The boxplots below present the times for the students according to direction and period of the day. Use these boxplots to comment on the comparisons between the three periods of the day and whether students are going up or down the stairs.



c Suggest a possible arrangement of data for the times going up the stairs that could explain the boxplots for going up in the different periods of the day. (Hint: one possibility is something similar to part of the boxplots of the cricket data in question 1 above.)

ISBN 978-1-107-65599-7

Chapter summary

Quartiles, interquartile range and five-number summary

- For quantitative data, the lower (upper) quartile is the median of the data below (above) the overall median
- The minimum, lower quartile, median, upper quartile and maximum are called the five-number summary. They divide the ordered data into quarters of equal frequency
- Interquartile range = (upper quartile) – (lower quartile)

Boxplots and interpretation

- The edges (or hinges) of the box in a boxplot mark the first and third quartiles. The median is marked by a line across the box
- Boxplots have whiskers, which go out to the last observation within a chosen distance of the box. This distance is usually 1.5 times the interquartile range. Observations outside this distance are marked by a star or an asterisk (*)
- Boxplots show how the four quarters of data are clustered or spread
- Boxplots are unchanging for a dataset, but they do not show the number of observations nor sub-groups of data.

Using boxplots to compare data

- Provided datasets are not too small, boxplots on the same scale are a good presentation method for comparing a number of quantitative datasets
- Boxplots can compare location, spread, skewness and unusual observations.

Multiple-choice questions

1 The quartiles and the median divide which of the following into four equal groups?

A The observations in the order collected
B The observations ordered from smallest to largest
C The range of the data
D The values of the data

2 The lower quartile of data is defined as the median of the observations less than the median because

A This has a quarter of the observations less than it
B This has a quarter of the observations between it and the data median
C This has three quarters of the observations greater than it
D All of these

3 The values of the data quartiles and the median may change if

A The smallest observation is decreased
B The largest observation is increased
C The data values are rounded
D None of these

4 Data are collected on people's favourite colour. Which of the following can be used to present these data?

A Stem-and-leaf plot
B Histogram
C Boxplot
D None of these

5 Which of the following has sufficient information for drawing a boxplot?

A The mean, median and range of the data
B A histogram of the data
C A stem-and-leaf plot of the data
D None of these

ISBN 978-1-107-65599-7

6 Boxplots are **not** good for

A Comparing a number of datasets
B Looking at location and spread of data
C Looking at skewness
D Looking for sub-groups in data

Short-answer questions

1 Example 1 in Chapter 1 gives data on the density, in g/cm^3 to the nearest 0.01, of the heads of a type of coral in the Great Barrier Reef less than 20 km and more than 20 km from the coastline.

a Draw boxplots of the two datasets on the same scale.

b Comment on the features of the boxplots.

c Draw a boxplot of all of the data. What is hidden?

d Looking at the dotplots and stem-and-leaf plots in Example 1 of Chapter 1, which plots do you think best represent these data? Why? What is a problem with any of the plots?

2 Refer to questions 1 and 2 of the Pre-test of this chapter.

a Draw boxplots of the two datasets of questions 1 and 2 on the same scale.

b What are the differences between the boxplots?

c Which features of the histograms show clearly in the boxplots?

3 Refer to question 3 of the Pre-test of this chapter.

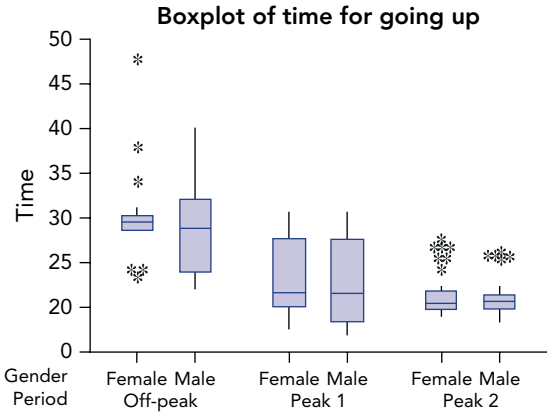
a Draw boxplots on the same scale of the amounts of protein per 100 g of cereal for the two manufacturers.

b What are the differences between the boxplots?

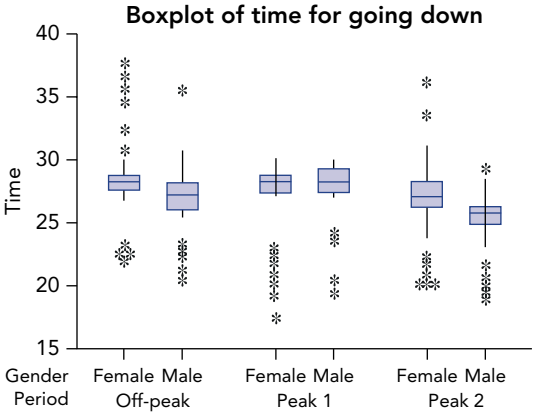
c Which difference shows in the boxplots more than in the histograms you drew for these data?

ISBN 978-1-107-65599-7

4 Question 5 of Exercise 4C looks at boxplots of the time students took to go up or down stairs at a busy city bus station. The dataset also included people classified as city workers by their style of dress. Below are boxplots for the 323 male and female city workers recorded going up the stairs at different times of the day – Peak 1 (7.30 am–10.30 am), Off-Peak (10.30 am–4 pm), Peak 2 (4 pm–7 pm). Use the boxplots to comment on the speeds of city workers going up the stairs.



5 Refer to question 4 above. Below are boxplots for the 482 male and female city workers recorded going down the stairs at the different times of the day.



a Use the boxplots to comment on the speeds of city workers going down the stairs.

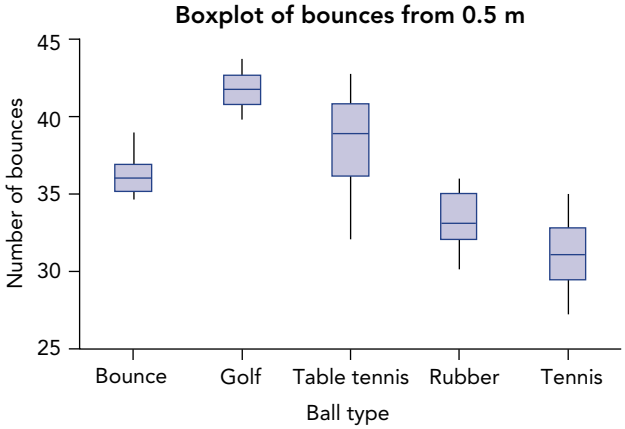
b Give one difference between going up and down the stairs for city workers.

ISBN 978-1-107-65599-7

Extended-response question: How bouncy is a golf ball?

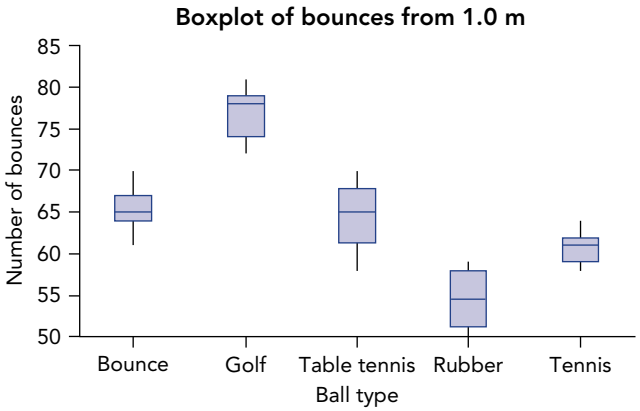
6 An experiment to compare the bounciness of different types of balls considered bounce, golf, table tennis, rubber and tennis balls. Twenty balls of each type were dropped from three different heights and the numbers of bounces before the balls came to a stop were counted.

a Below are boxplots of the numbers of bounces when the balls were dropped from 0.5 m. Use the boxplots to compare the bounciness of the balls.

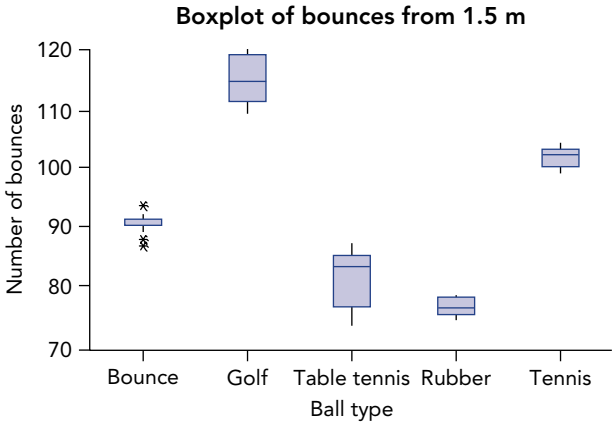


ISBN 978-1-107-65599-7

b Below are boxplots of the numbers of bounces when the balls were dropped from 1 m. Use the boxplots to compare the bounciness of the balls.



c Below are boxplots of the numbers of bounces when the balls were dropped from 1.5 m. Use the boxplots to compare the bounciness of the balls.



d What can you see happening as the heights increase?

ISBN 978-1-107-65599-7

CHAPTER

5

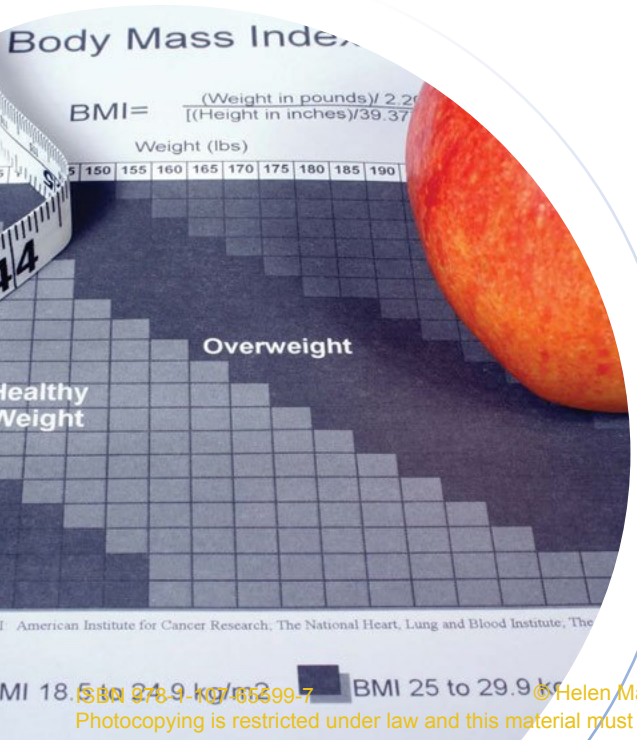
Scatterplots

What you will learn

5-1 Scatterplots of two quantitative variables

5-2 Scatterplots involving more information

5-3 Plots against time



What is BMI?

Body Mass Index (BMI) was developed by Adolphe Quetelet between 1830 and 1850 as a simple way of allowing for height in considering people's weights with the aim of being able to find ideal weights. It is usually defined as weight divided by height squared. The units are usually of weight in kg and height in m.

Quetelet's index is still used. There has been, and will continue to be, considerable debate and research on how it should be used and what ranges should be considered as ideal for different nationalities and ages. But it still provides a useful measure. For example, for children it compares data for girls and boys of the same age separately in the same way that heights and other developmental measures are compared. So parents will typically be told where their child is in relation to other children of the same age. For example, a girl might be in the top 5% of heights for her age.

From a statistical point of view, what we want for a BMI is that it is weight adjusted for height – that is, that it not depend on height. Obviously there will be a lot of variation across types of people, amounts of muscle, bone structures and levels of fitness. Generally speaking, tall people tend to have a leaner build, which is why the BMI is weight/(height squared) and not weight/(height cubed). Quetelet and researchers since have suggested that a better index is somewhere between these two. However, because there are so many other matters to consider in discussing weight and health of individuals, the BMI is just one index or measure to consider.

AUSTRALIAN CURRICULUM

Statistics and probability

- Data representation and interpretation
- Use scatter plots to investigate and comment on relationships between two numerical variables **(ACMSP251)**
- Investigate and describe bivariate numerical data where the independent variable is time **(ACMSP252)**

ISBN 978-1-107-65699-7

The BMI and associated measures of weight, health and fitness, are all about investigating relationships between quantitative variables. Graphing and exploring such relationships are introduced in this chapter.

PRE-TEST

1 The following table shows weekly February rainfall in millimetres in two Australian towns.

Town A				Town B			
13	18	5	10	58	29	66	42
0	0	1	2	14	20	29	31
96	17	38	0	83	79	73	66
7	1	130	11	52	38	36	34

a How many years are the data for?

b What do we need to know to be able to compare the February rainfalls in these two towns?

c What would be a better way to present the data in a table?

ISBN 978-1-107-65599-7

2 An experiment was carried out on the lifetimes (in weeks) of wire specimens to test their thermal endurance. The tests were carried out in two laboratories. Wire specimens were heated to a test temperature of 200°C. The data were presented in the table below.

Lab 1	14	16	17	18	20	22	23	25	27	28
Lab 2	27	28	29	29	29	30	31	31	33	34

a What impression is given by this table of data?

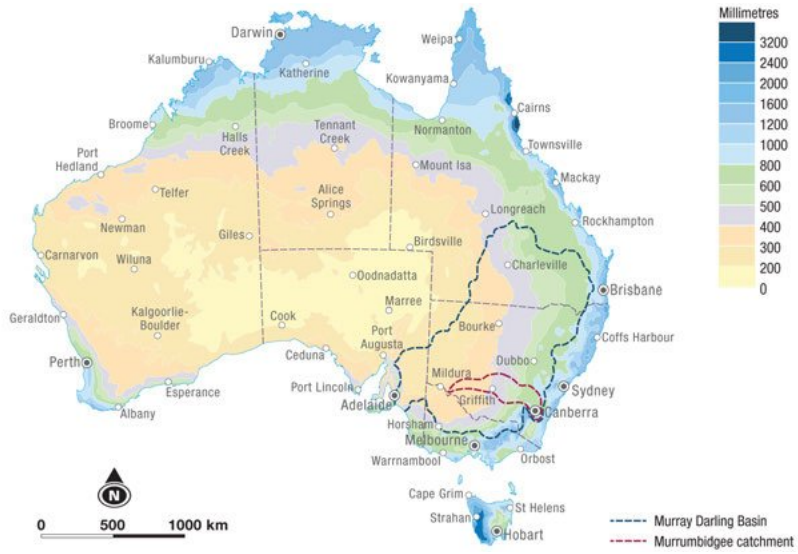
b What information do we need about the experiment so that we can interpret the table?

c What is strange about the data as presented in the table? Does it provide any clues in interpreting the table?

3 To monitor petrol prices 'at the pump' over time in a region, how should care be taken in the collection of the prices to ensure suitable data?

4 The following are part of a dataset to monitor weekly rainfall in a region. What are two problems with these observations?

20 mm 54.5 mm 23 mm 15.2 mm 1.5 inches 2 inches



5 It is claimed that crime is increasing in a region. The annual numbers of various types of crimes are recorded. What is essential information to be able to investigate the claim?

6 In a survey, people are asked to name their favourite sport and also to say if there is sufficient news coverage of it.

a The sports are coded using codes such as 1 = cricket, 2 = soccer, 3 = tennis and so on. Do these codes have any numerical meaning?

b The question on news coverage is phrased as 'There is sufficient news coverage of this sport.' People are asked to respond by choosing a number from 1 to 5, where 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree and 5 = strongly agree. For these numbers, does 5 – 4 = 3 – 2?

c A report publishes the data on parts **a** and **b** separately. What's wrong with this?

Terms you will learn

pairs of observations
scatterplot
time series plot
trends, patterns and variation over time
x-axis
y-axis

ISBN 978-1-107-65599-7

5-1 Scatterplots of two quantitative variables

Scatterplot: A plot used to explore relationships between two quantitative variables … *see glossary*

Pairs of observations: Data on two variables that are observed for each subject; each pair of observations belongs to one subject

x-axis: The horizontal axis of a scatterplot

y-axis: The vertical axis of a scatterplot

Throughout Chapters 1, 2 and 4, we have been using various plots to explore and compare datasets of continuous data (and of count data with many different values) across categories of one or more categorical variables. This can also be viewed as exploring relationships between a continuous variable and one or more categorical variables. For example, how do male reaction times compare with those of females? Do drivers speed up when approaching amber lights compared with the speed when approaching green lights? Is this behaviour different for different age groups?

We often need to explore relationships between quantitative variables in many different areas of interest. Are sales related to amount of advertising? How are people's reflexes related to their age? How closely related are head circumference and width of shoulders? Can we predict an adult's height from their length at birth?

A **scatterplot** is used to explore relationships between two quantitative variables. The data are **pairs of observations** – a value for each variable. Each pair belongs to one subject. We set up a horizontal axis, called the **x-axis**, for one variable and a vertical axis, called the **y-axis**, for the other variable. So they are often called the *x*-variable and the *y*-variable. The range of values along the *x*-axis is the range of values of that variable in our data. Similarly, the range of values on the *y*-axis is the range of values of that variable in our data. Then each subject has a point on the graph, with its *x*-value being the observation of the *x*-variable for that subject and the *y*-value being the observation of the *y*-variable for that subject. The *x*-variable can be any quantitative variable (continuous or count) but the *y*-variable should be a continuous variable (or a count variable taking many different values).

To explore possible relationships between two variables, it tends not to matter which one is assigned to the *y*-axis and which to the *x*-axis. However, if we want to see what happens to one variable as the other changes, then the first variable should be assigned to the *y*-axis and the second to the *x*-axis.

LET'S START How do head circumference and age relate?

The measurements of head circumference in centimetres (measured around the eyebrows) and age in years were collected on a random sample of subjects.

Here are the first eight of the pairs of observations, with age first in each pair:

(17, 56.0) (21, 56.0) (48, 63.0) (15, 57.0)
(17, 55.3) (22, 54.0) (53, 57.0) (65, 55.0)

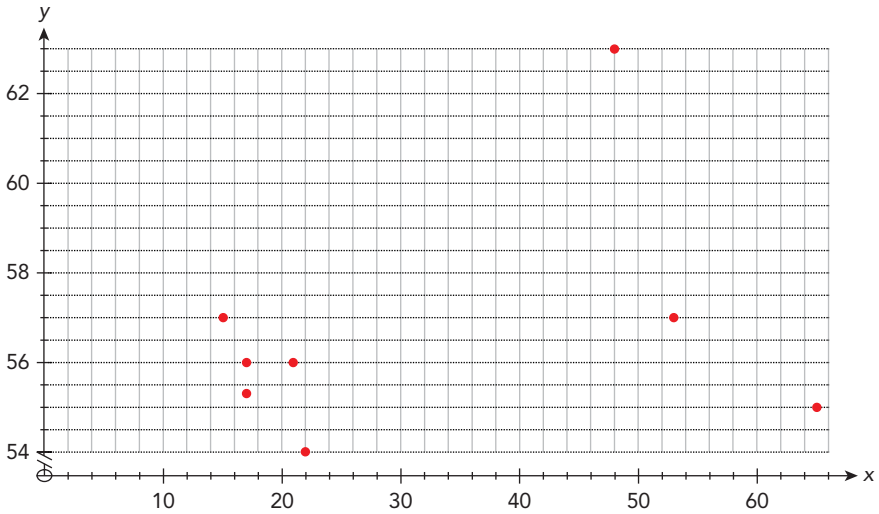
The ages in the data range from 2 to 65, and the head circumferences from 46 cm to 63 cm. Which variable will we put on the horizontal – the *x*-axis? We are usually interested in what happens as age increases, so we'll put age on the *x*-axis and head circumference on the *y*-axis.

Drawing the plot

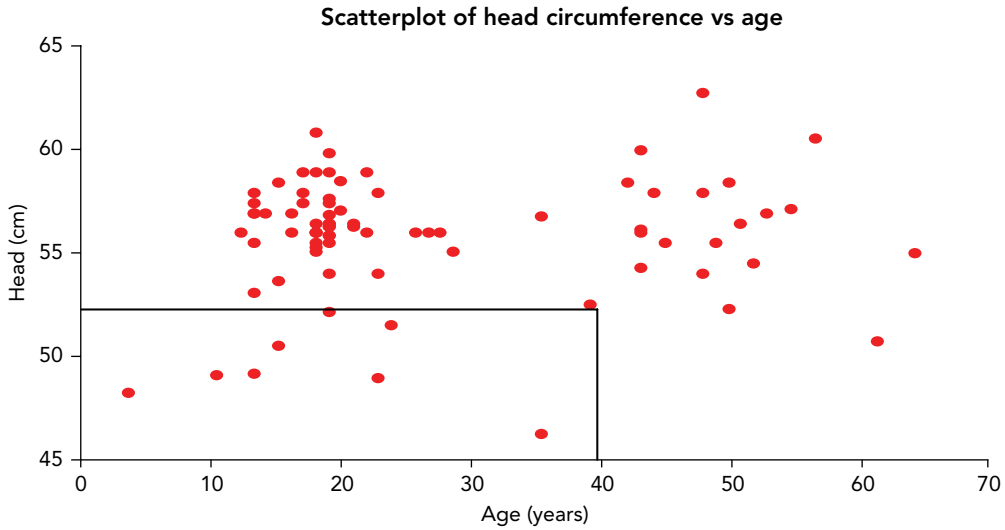
HINT
The data pairs are usually written with the *x*-axis variable first and the *y*-axis variable second.

We can start the *x*-axis at 0 years, and go up to 65 or 70 years. The *y*-axis can't start at 0 because all the observations will be squashed up in a small part in the upper right-hand of the graph; we wouldn't be able see anything of how head circumference and age are related or not! Therefore it is sensible to start the *y*-axis at 40 cm or 45 cm and go up to 65 cm or 70 cm.

The first point entered on the graph is vertically above 17 on the *x*-axis and is level with 56 cm on the *y*-axis. Continuing like this gives the first 8 points on the scatterplot as shown below.



There are 79 subjects in this dataset. Continuing in this way gives the following scatterplot.



CAUTION
It is very important to name the variables on each axis.

ISBN 978-1-107-65599-7

Each dot represents one observation with a pair of values. The age of the pair is on the horizontal axis and the head circumference of the pair is on the vertical axis. For example, the point highlighted in the above plot corresponds to a person of age 39 years with a head circumference of 52.5 cm.

CAUTION
We do not join up points in a scatterplot because we need to consider and allow for variation in looking at relationships. Imagine what the plot above would look like if we tried to join up points!

Is there any relationship?

Looking at the plot, does there seem to be any relationship between head circumference and age? Obviously there are a few children in this dataset with ages less than 12 and smaller heads than most of the older subjects, but not by much. There are quite a few older people with head circumferences as small, and one aged about 35 years with a smaller head circumference than the child aged 2 years, which seems highly unlikely. This data point would have to be checked in case it was a mistake, but there is another subject aged 22 years with a head circumference of not much more. Perhaps the reliability of these measures of head circumference needs to be checked! In general, what the plot shows is that there is very little relationship between head circumference and age in these data, but there is a lot of variability. That is, for people of the same approximate age, there is a lot of variability in their head circumferences.

Key ideas

- To investigate whether and how two quantitative variables are related to each other, we need pairs of observations – observations on each variable for each subject.
- Data on two quantitative variables can be presented and explored in a scatterplot that plots points for the data pairs.
- Which variable goes on the vertical axis and which on the horizontal depends on the context.
- A scatterplot shows how much variation there is in the data as well as suggesting if the two variables are related and in what way.

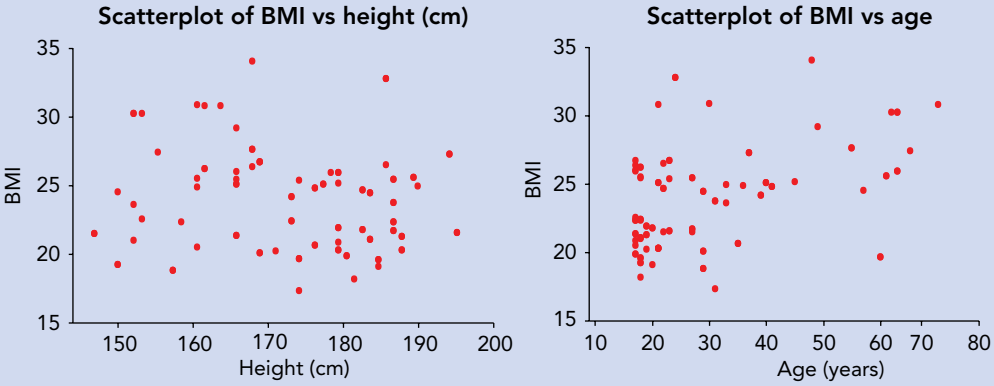
Example 1: What does a scatterplot of BMI look like?

The aim of the BMI is to have an index of body mass that has no or little dependence on height. So if we plot BMI against height, the plot should not indicate that BMI changes in a systematic way with height. Obviously BMI will vary with gender, type of build, amount of muscle and so on. So we expect to see variation in BMI across any random group of people. There are many questions about BMI.

Questions: Does BMI not depend on height? Does BMI in adults tend to increase with age?

Some body statistics were collected for a random sample of 54 young Australians aged between 17 and 25, and another of 45 Australians aged more than 25 years. On the next page are scatterplots of BMI (in kg per (height in m)^2) against height (in cm), and of BMI against age in years.

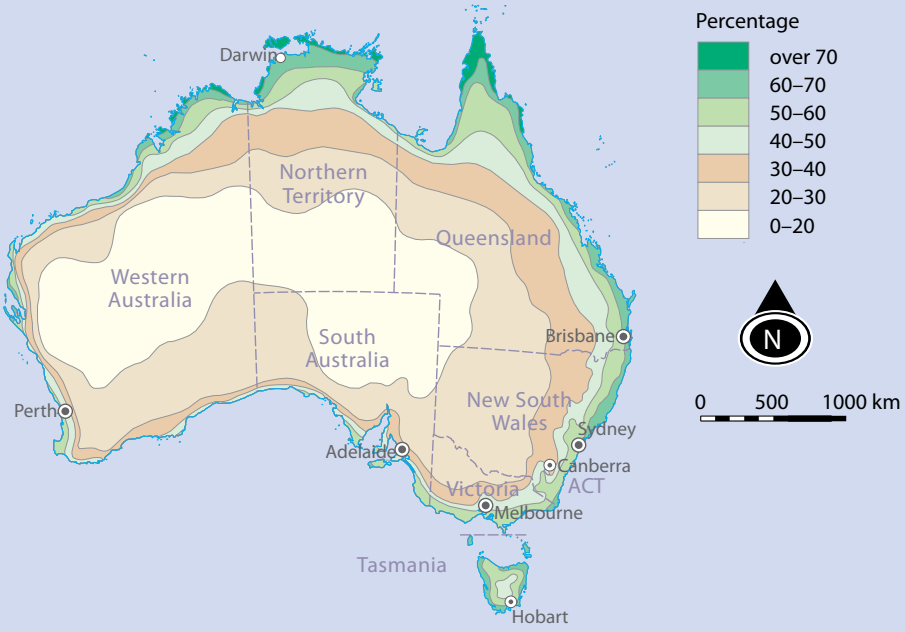
ISBN 978-1-107-65599-7



We see that there is a lot of variation in BMI, but as required, there does not appear to be any relationship to height. In the plot against age, again there is a lot of variation in BMI for those younger than 30. Then there is perhaps some suggestion in the plot that BMI tends to increase with age, but there are not very many people older than 30 in this plot so it is difficult to tell.

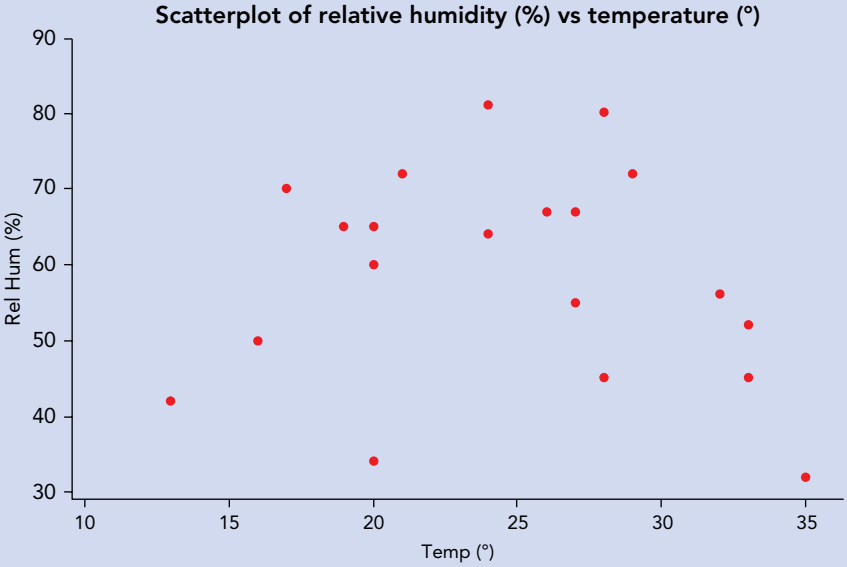
Example 2: What is the weather like?

It is not just the temperature that makes the weather feel pleasant or not, but also the humidity, which is expressed in a percentage as relative humidity. A resident in a particular Australian town took a random sample of 20 days and looked up the maximum temperature (in Celsius) and the maximum humidity (relative humidity in percentage). On the next page is a scatterplot of humidity against temperature.



ISBN 978-1-107-65599-7

Question: How does the humidity vary with the temperature in these data?



As expected, we see that there is quite a lot of variation in the humidity for similar temperatures. This would depend on the type of weather and the season. But we also see that there is a general tendency for the humidity to increase with temperature but then decrease for higher temperatures.

Exercise 5A

1 Question 2 of Exercise 4C refers to an experiment in which four lists each of 25 words were tested on a group of 24 people to investigate lists of words that can be used in hearing tests. Below are the scaled scores out of 50 for the first and second lists for the 24 people.

Person	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Hearing test 1	28	24	32	30	34	30	36	32	48	32	32	38	32	40	28	48	34	28	40	18	20	26	36	40
Hearing test 2	20	16	38	20	34	30	30	28	42	36	32	36	28	38	36	28	34	16	34	22	20	30	20	44

ISBN 978-1-107-65599-7

a Draw a scatterplot of the scores for the first test against the scores for the second test.

b Does it matter in this case which variable is on the vertical axis?

c Does the plot suggest that the scores are related? In what way?

d Does the plot suggest that the 24 people are consistent in their scores over both tests or do they vary from one test to the other?

e Amongst what values of scores is there most variation between the two tests?

2 The data below are of the amounts (in mg) of an unconverted substance from nine similar chemical reactions that were run for different amounts of time.

Amount unconverted (mg)	23.5	16.9	17.5	14.0	9.8	9.1	8.3	7.9	7.2
Reaction time (min)	1	2	3	4	5	6	7	8	9

a Draw a scatterplot of the amounts of unconverted substance against the reaction time.

b Why should the amount of unconverted substance be on the vertical axis in this plot?

c What is the plot suggesting happens as the time of reaction increases?

d Are there any data points that look as though the experiment or recording sheet needs to be checked?

e What do you think might happen if the reaction time was increased to 10 minutes?

3 Refer to question 5 of Exercise 1A, which gives data on an experiment conducted to investigate if people's perception of time is affected by focusing on an activity such as reading.

a Draw a scatterplot for the females of the guesses of time when reading against time when not reading.

b Does the plot suggest that the guesses of females when reading are close to their guesses when not reading?

c Draw a scatterplot for the males of the guesses of time when reading against time when not reading.

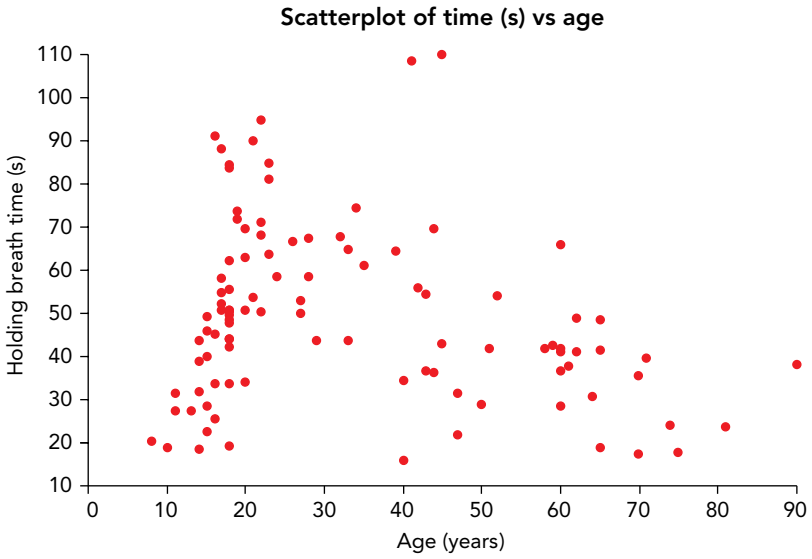
d Does the plot suggest that the guesses of males when reading are close to their guesses when not reading?

e What are the guesses of the female and male who are most out with their guesses when not reading?

f What do you think is the main contrast between the two plots?

4 As part of an experiment a random sample of people covering a wide range of ages were timed to see how long they could hold their breath. On the next page is a scatterplot of their holding-breath times in seconds against their age in years.

ISBN 978-1-107-65599-7



a What is the plot suggesting about how ability to hold breath changes with age?

b What approximate age range has the most variation?

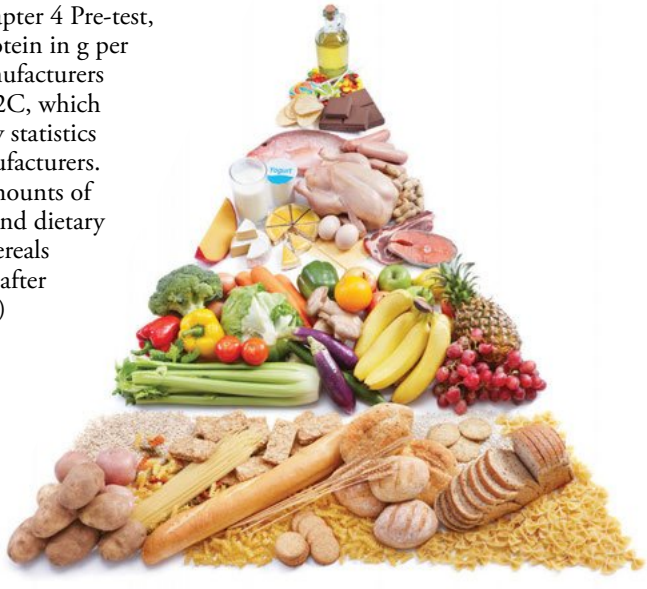
c There are two unusual observations. In what way are they unusual?

d What other information would you like to know about the subjects in this study that might be of help in investigating how long people can hold their breath?

Enrichment

Are the amounts of protein, carbohydrate and dietary fibre in cereal related?

5 Refer to question 3 of the Chapter 4 Pre-test, which gives the amount of protein in g per 100 g of cereals from two manufacturers and to question 5 of Exercise 2C, which gives histograms and summary statistics for 66 cereals from three manufacturers. The following data give the amounts of protein (g), carbohydrate (g) and dietary fibre (g) per 100 g cereal for cereals from manufacturer C. (The T after Carbohydrate stands for total.)



ISBN 978-1-107-65599-7

Manufacturer C

Protein (g)	11.6	8.0	7.6	12.0	11.5	10.3	10.4	7.6	8.3
Carbo. T (g)	73.0	79.0	75.0	67.0	55.9	67.6	66.0	80.0	72.3
D. Fibre (g)	6.7	5.5	6.5	11.0	18.3	9.3	10.4	3.1	9.3

a Draw scatterplots for the data for Protein vs. Carbohydrate, for Protein vs. D. Fibre and for D. Fibre vs. Carbohydrate for the cereal from manufacturer C.

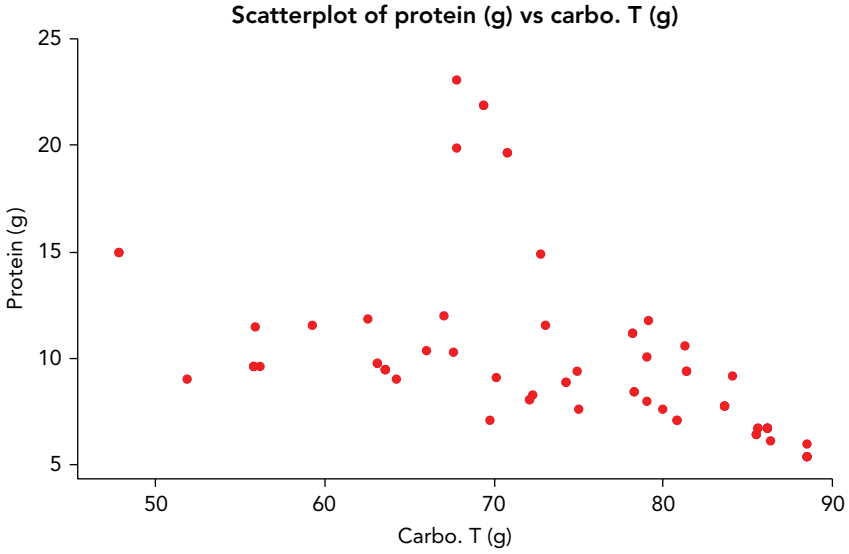
b Comment on what each of the plots indicates.

c Which variables seem to be the most strongly related?

d Which plot seems to have most variation? This indicates that the relationship is not very strong.

e Did it matter in these plots which variable(s) were on the vertical axis?

f Below is a scatterplot of Protein vs. Carbohydrate for all 66 cereals of question 5 of Exercise 2C. Comment on this plot.



ISBN 978-1-107-65599-7

5-2 Scatterplots involving more information

As you have seen in the examples and exercises throughout this book, real datasets usually have more than two variables. Issues worth investigating usually involve a number of questions and a number of possible variables of interest. Often we have to choose which variables we are going to limit the investigation to.

How can we include other variables on a scatterplot? We can include a categorical variable on a scatterplot between two quantitative variables by using different symbols or colours for the different categories of the categorical variable. If we have too many categories the plot can become too confusing to read. However, often a categorical variable of interest has just two or three categories.

Including a categorical variable through different symbols is the most commonly occurring addition to a scatterplot. This can be very useful. For example, in the head circumference versus age plot, the BMI versus height or age plot, and the holding breath plot, the most obvious categorical variable to include is gender. The guess of 20 seconds reading versus non-reading plots in question 3 of Exercise 5A are separate for males and females, so these could be combined in one plot with different symbols for males and females. The result of this is to have two or more scatterplots on the same plot. Sometimes this comes about by combining plots. For example, question 1 of Exercise 5A plots the hearing scores for test 1 versus test 2. But there are four tests altogether. So the scores for tests 2, 3 and 4 could all be plotted against the scores for test 1 on the same plot.

We can include two categorical variables where the different symbols are for the combination of categories, but how good a picture this is depends on the data. The plots can sometimes look quite messy!

Can we do more? Can we include another quantitative variable? Not without very clever computer graphics, and possibly with dynamic plots. Possibly the best known worldwide are the excellent Gapminder resources (www.cambridge.edu.au/statsAC910weblinks). These provide an amazing range of innovative and dynamic plots of data from official international and national sources, particularly focused on public health issues. The plots cleverly combine three quantitative and one categorical variable. A fourth variable of time is able to be included dynamically as the viewer can choose to follow the development of the plots over time. We will see just a few examples below.

LET'S START Can we judge distances one-eyed?

The judging distance data in section 1-1 are part of a larger dataset to investigate how well people can judge a distance of 5 metres. The subjects were classified by gender and it was also noted if they wore glasses or contact lenses. The experiment involved subjects judging when one of the testers, holding a tape measure upside down, had walked to 5 metres away. Subjects first judged this using both eyes, and then (turning around so that they weren't looking at the same surroundings) covering one eye. The subjects could choose the eye they preferred to use.

ISBN 978-1-107-65599-7

To plot how the one-eyed guess compares with the two-eyed guess, we could have different symbols for the four categories from combining glasses/contact lenses or not (Y, N) and male or female (M, F). But below we just look at including the Y and N for glasses/contact lenses or not.

We see that generally there is a relationship between the two guesses; the distance guessed one-eyed tends to increase as the distance guessed two-eyed. But there is a lot of variation – and some way out guesses! There's not much difference in the plot between those wearing glasses/contact lenses and those who don't, although possibly there is more variation in the guesses of those who wear glasses/contact lenses. Probably the most interesting feature is that there is a tendency to guess a smaller distance one-eyed than two-eyed. To see this, put a ruler edge (or similar) on the page so that it joins up the two points (4.0, 4.0) and (6.5, 6.5). There are more points below this line than above it – that is, more points with one-eyed guesses less than two-eyed guesses.

Key ideas

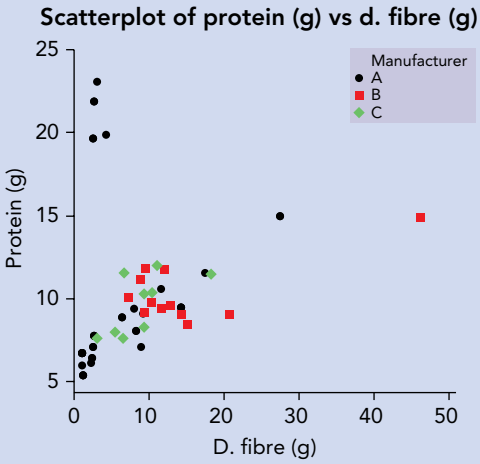
- We can include a categorical variable in a scatterplot between two quantitative variables by using different symbols for the categories of the categorical variable.
- We can include two categorical variables by combining categories but too many different symbols make a scatterplot difficult to interpret.
- Including more variables, possibly quantitative, requires clever graphics and special software.

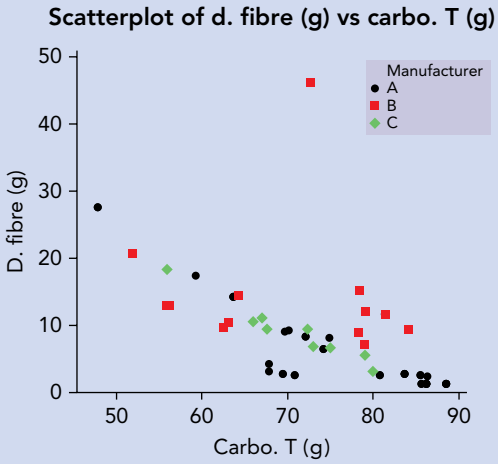
ISBN 978-1-107-65599-7

Example 3: What is happening in our breakfast cereal?

In the plot in question 5 of Exercise 5A, it looks like there may be sub-groups. Are they perhaps for the same manufacturer? There are also two other possible scatterplots – protein and dietary fibre, and carbohydrate and dietary fibre. Can we plot protein and carbohydrate against dietary fibre on the same plot? We can because the same variable is on the horizontal axis, but it will not be a good plot because the other two variables are too different and take different ranges of values.

Below are the two scatterplots versus dietary fibre, using different symbols to represent the different manufacturers.





As it doesn't matter which variable is on the *x*-axis in this example, the above plots have kept protein on the *y*-axis and carbohydrate on the *x*-axis for easy comparison with question 5 of Exercise 5A.

We can see that the group of four with high protein that stood out in question 5 do come from the same manufacturer and have low fibre. There is also one with much higher fibre than all the others. These cereals are specifically designed to focus on a nutritional aspect. It's interesting to see where they are in relation to others. Apart from those, generally increased fibre goes with increased protein, and increased carbohydrate goes with decreased fibre. Manufacturer A seems to have certain groups of cereals, while manufacturer B has quite a bit of variation. Manufacturer C seems to have more regular variation.

ISBN 978-1-107-65599-7

Example 4: Comparing CO_2 emissions, income, population, world region and year

Cambridge GO

Below is a screen capture of one of the Gapminder plots. (www.cambridge.edu.au/statsAC910weblinks). It is plotting CO_2 emissions (in tonnes per person) versus income per person, for each country. The bubble size represents population size and the bubble colour represents a region of the world. The slider along the bottom allows choice of year and the plots can be looked at dynamically to watch how the world situation changes over the years. As the cursor is moved across a bubble, the name of the country appears. We see that as income per person increases, the variation in CO_2 emissions increases a lot. Also the emissions accelerate – the increase is faster as they increase.

The Gapminder plots are all of this type, on a wide range of national and international statistical data.

ISBN 978-1-107-65599-7

Exercise 5B

Questions 1 and 2 refer to the following data. Below is a random sample of 30 observations obtained from the ABS Census At School website (www.cambridge.edu.au/statsAC910weblinks) using Random Sampler. The sample is taken from all the Year-10 students in all Australian states/territories who completed the 2012 CensusAtSchool questionnaire. The questions included here are height without shoes (in cm), armspan (in cm)(measurement from fingertip to fingertip with arms outstretched) number of hours of sleep usually obtained on a school night (to nearest hour), and time it usually takes to get to school (in min).

Gender	F	M	M	M	F	F	F	F	F	M	M	M	M	M	M
Height (cm)	162	172	180	182	164	163	171	156	161	189	190	181	166	184	157
Armspan (cm)	162	172	186	181	159	157	165	156	162	197	150	178	175	190	10
Sleep (h)	6	8	8	9	8	8	8	7	9	8	8	8	9	10	16
Time to school (min)	5	10	20	30	20	5	10	15	75	3	15	12	35	10	45

Gender	F	F	F	F	F	M	F	M	M	F	F	F	M	F	F
Height (cm)	158	168	170	157	166	181	173	173	166	156	159	160	162	163	162
Armspan (cm)	160	152	170	100	163	177	178	177	167	156	162	160	162	167	161
Sleep (h)	8	9	8	10	11	*	8	9	8	6	8	8	9	7	8
Time to school (min)	15	5	10	30	30	10	40	11	10	40	15	10	7	30	2

1 Consider the armspans and heights.

- **a** Are there any observations that are clearly wrong? Omit any observations that are clearly wrong.
- **b** Draw a scatterplot of the remaining observations, using different symbols for males and females.
- **c** Which variable did you choose to put on the vertical axis? Why?
- **d** Do any other points show up on the plot as being so strange that they are probably wrong?
- **e** Comment on the plot.

2 Consider the number of hours sleep and the time to school.

- **a** Are there any observations that are clearly wrong? If so, were they also clearly wrong observations in question 1?
- **b** Omit any observations that are clearly wrong. Draw a scatterplot of the remaining observations, using different symbols for males and females.
- **c** Which variable did you choose to put on the vertical axis? Why?
- **d** Do any other points show up on the plot as being so strange that they are probably wrong?
- **e** Comment on the plot. Why is this plot difficult to comment on?

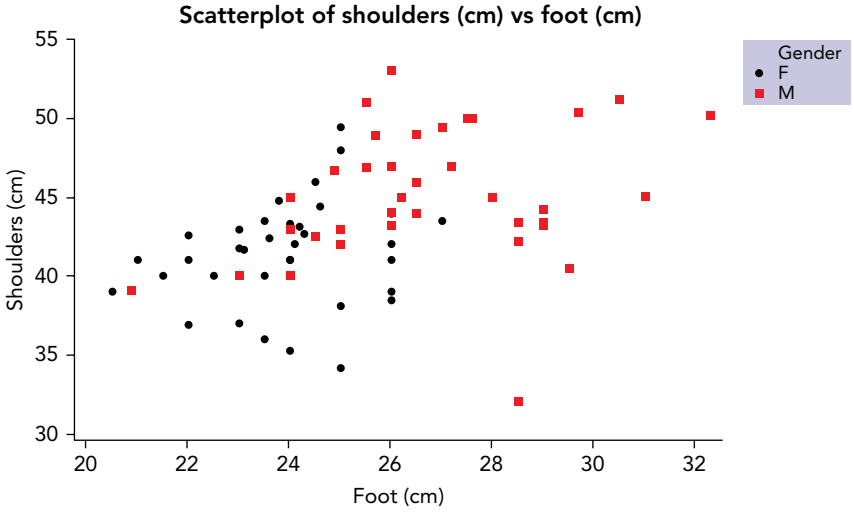
ISBN 978-1-107-65599-7

3 Below is a scatterplot of width of shoulders versus length of the right foot for a random sample of people, with males and females marked separately.

a Describe how the variation in shoulder width changes as the foot length increases.

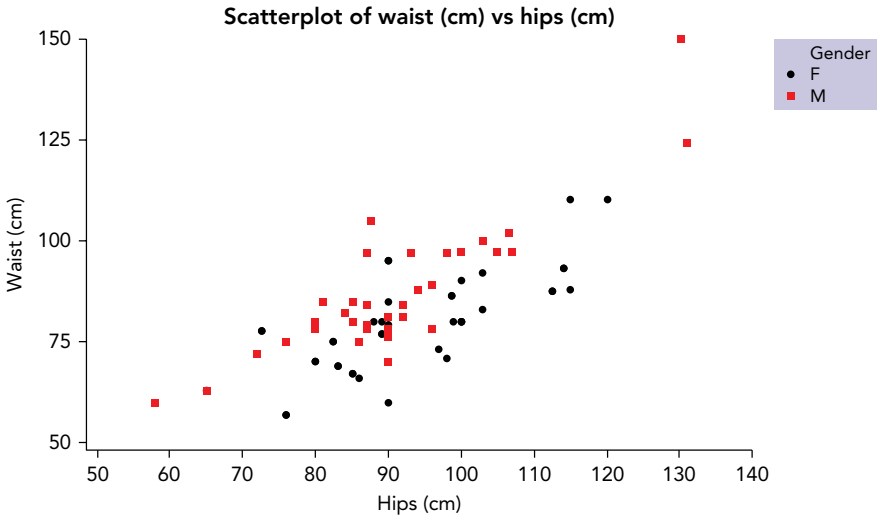
b In what ways are the male and female data different?

c Are there any observations that might need checking? Why or why not?



4 In assessing some health conditions, the ratio of waist measurement to hip measurement is used as well as or instead of BMI. On the next page is a scatterplot of waist measurement (in cm) to hip measurement (in cm) for the random sample of people considered in the BMI dataset above.

ISBN 978-1-107-65599-7

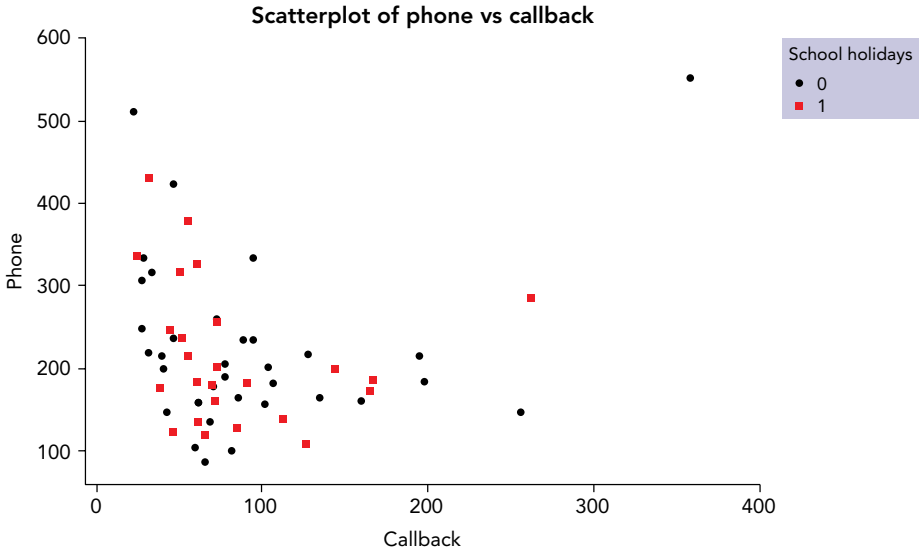


a Do you think the ratio of waist over hip measurement would be reasonably constant as hip measurement changed? Why or why not?

b Does the plot suggest that the relationship between waist and hip is similar between males and females or are there some differences?

c Does the variation look similar or different for males and females? In which ways?

5 Question 4 of Exercise 1C and question 4 of Exercise 2C consider the numbers of complaints about noise per month made in different ways to Sydney airport over a number of years. These complaints are made per month, so the number of complaints of different types may be linked. For example, if the number of phone complaints has increased, have the numbers of other complaints also increased, or has the way of making complaints just shifted? The scatterplot below is of the number of phone complaints per month versus the number of callback complaints per month. Different symbols are used for months with no school holidays (school holidays = 0) and months with school holidays (school holidays = 1).



ISBN 978-1-107-65599-7

a Is the plot suggesting any link between the number of monthly phone complaints and the number of monthly callback complaints? If so, what is the plot suggesting?

b Is the plot suggesting that months with school holidays are different in any way as far as numbers of complaints?

Enrichment

How does life expectancy relate to income? Has it changed over time?

6 The two plots below are Gapminder plots of life expectancy at birth versus income per person in 1950 and then in 2007 (www.cambridge.edu.au/statsAC910weblinks). The population of the country is represented by the size of the bubble, and the region of the world is represented by the colour of the bubble.

a What do you think is meant by Life expectancy (on the *y*-axis) and Income per person (on the *x*-axis) of these plots?

b How has the shape of the graph of the relationship between life expectancy and income per person changed between 1950 and 2007?

c What are some of the major changes between 1950 and 2007 indicated by these plots?

ISBN 978-1-107-65599-7

5-3 Plots against time

Some datasets are collected over time and we want to know how they vary over time. Are they tending to increase or decrease or is there a pattern? Do they vary a lot over time or not much or is the variation greater sometimes than others? In some areas like finance, economics, business, weather and public health, such plots over time are very important and are constantly watched. The Gapminder plotting system (www.cambridge.edu.au/statsAC910weblinks) was developed to provide good plots for public health. In the Gapminder plots, people can look at what happens over time by running plots over time. Here we look at plots against time.

January						
M	T	W	Th	F	Sa	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

February						
M	T	W	Th	F	Sa	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28

March						
M	T	W	Th	F	Sa	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

April						
M	T	W	Th	F	Sa	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30		

May						
M	T	W	Th	F	Sa	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24/31	25	26	27	28	29	30

June						
M	T	W	Th	F	Sa	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

These plots are scatterplots against time, but instead of looking for signs of relationships between the data on the *y*-axis and time on the *x*-axis, we are looking for **trends, patterns and variation over time**. Such plots over time are often called **time series plots** because we are looking at the data on the *y*-axis as a series over time. The data on the *y*-axis can be individual observations or averages or medians or percentages. Time series plots often plot more than one series on the same plot.

Trends, patterns and variation over time: Indications that the data may be increasing or decreasing over time, or that there might be a pattern that occurs again, or that variation might be changing over time

Time series plot: Scatterplot against time in which we are looking for trends, patterns and variation over time ... *see glossary*

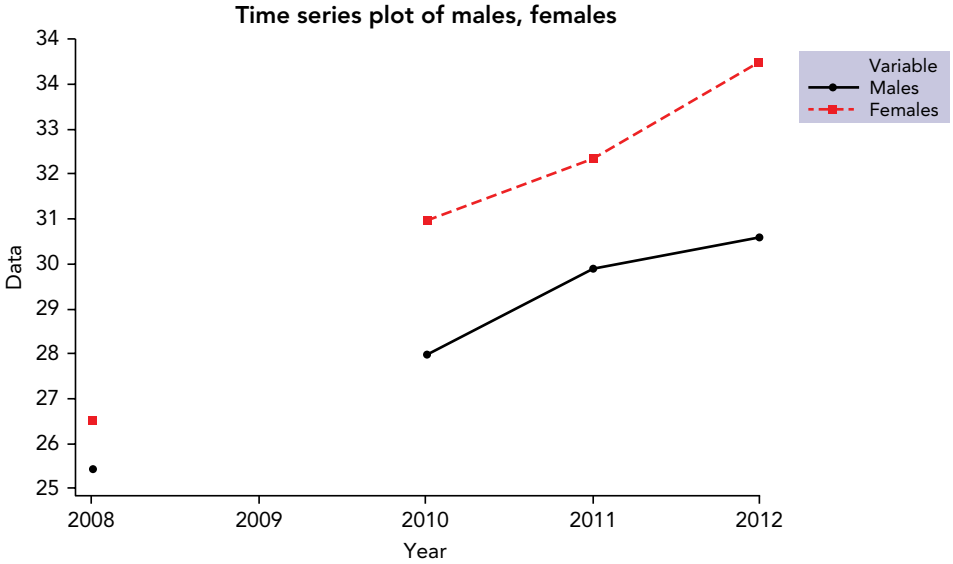
ISBN 978-1-107-65599-7

LET'S START How many school students speak more than one language?

The ABS Census At School website (www.cambridge.edu.au/statsAC910weblinks) publishes total results for some questions over time. For example, the percentage of all male and female students who completed the Census At School survey, who speak more than one language is given in the table below. (The survey asked in how many languages could students hold an everyday conversation.)

Year	2008	2010	2011	2012
% of females	26.5	30.8	32.1	34.2
% of males	25.4	27.9	29.7	30.4

CAUTION
Remember that these are percentages of those students who completed the survey. If you look at the ABS website you will see that the numbers participating in Australia across these years has varied quite a lot.



ISBN 978-1-107-65599-7

These are very simple to plot against year. The 'time series' plot is a simple scatterplot with the points joined unless there is data missing at a time point. Notice data for the year 2009 are missing.

Is this the best plot for these data? Because the *y*-axis starts at 25%, it looks like the percentages are growing very large. But they are only increasing by a bit each year. One possibility is to draw column charts of the percentages over the years for males and females. Sometimes when we want to focus on the size of increases (and decreases) we plot the change over time.

HINT
You will see in the examples below that in time series plots we usually do connect the points to make it easier to see trends, patterns and variation. The above is a very simple graph and the lines could be omitted.

But these data are very simple, and provided we're careful in reading the *y*-axis, we can see that the trend over the years is a steady increase of about 1% per year for boys and about 1.5% per year for girls.

Key ideas

- For some data collected over time, scatterplots of the data against time can show trends, patterns and variation over time. Such plots are often called time series plots because the data are a series over time.
- The points on a time series plot are usually joined to help see trends and patterns over time. If data are missing at a time point, there will be a break in the line.
- More than one data series can be plotted against time on the same plot.

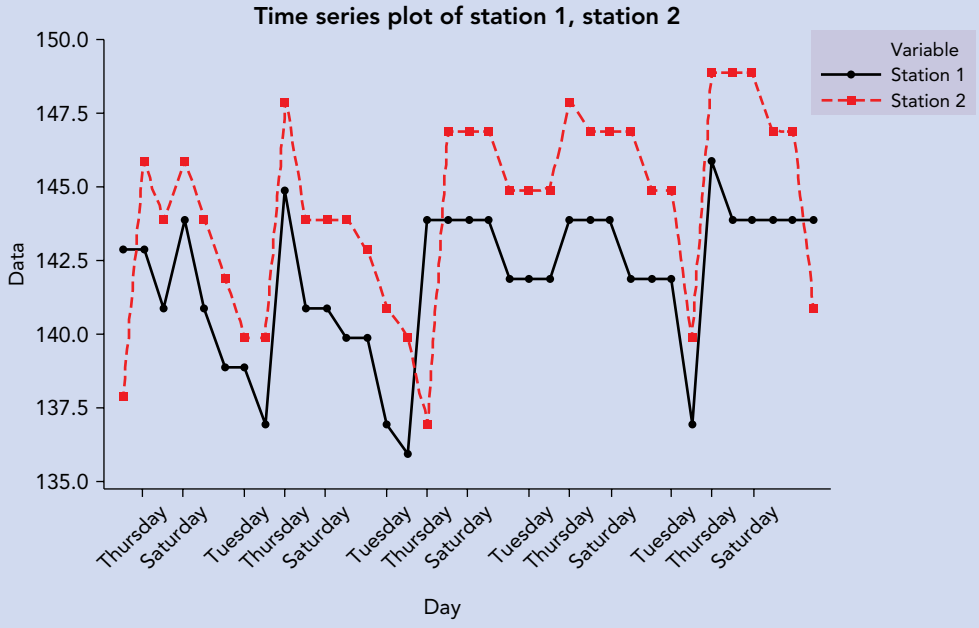
Example 5: Variation of petrol prices over time at two service stations

Question 6 of Exercise 2A and Example 5 of Chapter 2 give some information on the price of unleaded 91 E10 fuel from a number of randomly chosen service stations each day for (almost) 5 weeks. The service stations monitored were of two brands and at different distances from the city centre. On the next page is a time series plot for two service stations at approximately the same distance from the city centre, each selling a different brand of petrol.

Questions: Is there a pattern in the prices over days and weeks? Is the pattern similar for the two brands?

The weeks started on a Wednesday. We see that in weeks 2, 3 and 5, the prices dipped about Tuesday and Wednesday and rose on Thursday or Friday for both stations but in week 4 they did not. What was different about week 4 should be investigated if possible. Generally, station 1 is cheaper than station 2 although the price at station 2 sometimes dips the day after station 1 dips and the day on which station 1 rises. The pattern is reasonably similar for both stations.

ISBN 978-1-107-65599-7



Exercise 5C

1 Another question with total results for each survey posted on the ABS CensusAtSchool website (www.cambridge.edu.au/statsAC910weblinks), is on students' breakfasts. The percentages of students who did not eat breakfast on the morning they completed the survey were: in 2006 – 12.8%; in 2008 – 10.8%; in 2010 – 13.5%; in 2011 – 13.8%; and in 2012 – 13.9%.

a Draw a time series plot of these results.

b Does there appear to be a trend or not?

2 The monthly rainfall in millimetres was carefully recorded in the backyard of a home on the eastern side of Australia. Care was taken to use the same measuring device in the same way over many years. The data from 1978 to 1996 is plotted in two plots below but with the same scale on the *y*-axis. The first plot is from January 1978 to July 1986, and the second from January 1987 to May 1996. There are observations missing from July–December 1986.

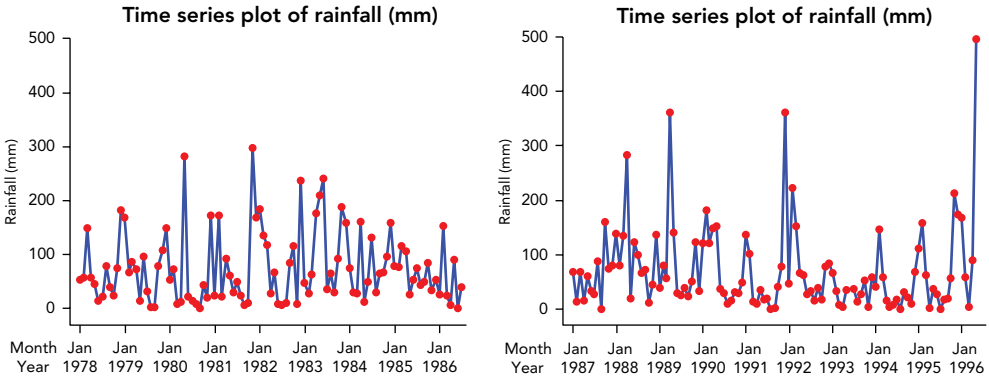
a When does this location tend to get most rain?

b What are two differences between the two plots? That is, differences in monthly rainfall patterns between the period 1978–86 and 1987–96?

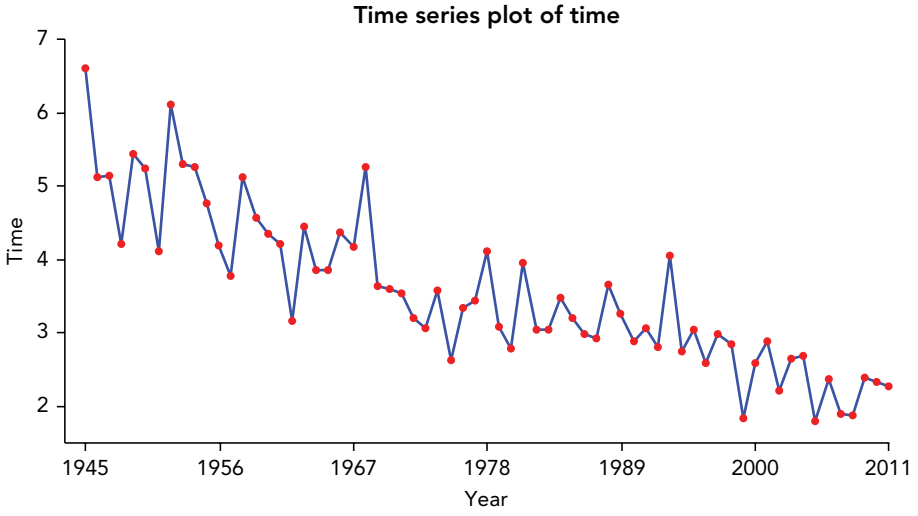
c Which were the driest summers for this location over these years?

ISBN 978-1-107-65599-7

d Describe briefly what happened in the years 1992–96. The last month of the data was May 1996.



3 The winning times for the Sydney–Hobart yacht race are available on the official website for the race. Below is a time series plot of the winning times in days (to two decimal places). What does the plot indicate about the trend and variation in winning times?



Enrichment

How have complaints to Sydney airport varied over time?

4 Question 4 of Exercise 1C and section 2-1 give some information on the monthly numbers of complaints about noise made in different ways to Sydney airport in the period 2001–06. Question 5 of Exercise 5B gives a scatterplot of the monthly numbers of complaints made by phone versus those requesting callback. Other ways of making complaints are by email, by the internet and by letter. Below are two time series plots: the first is of the monthly complaints by phone and by callback, and the second is of the monthly complaints by email and the internet. The same scale on the y-axis has been used on the two plots.

a How have the numbers of complaints of each type changed over time?

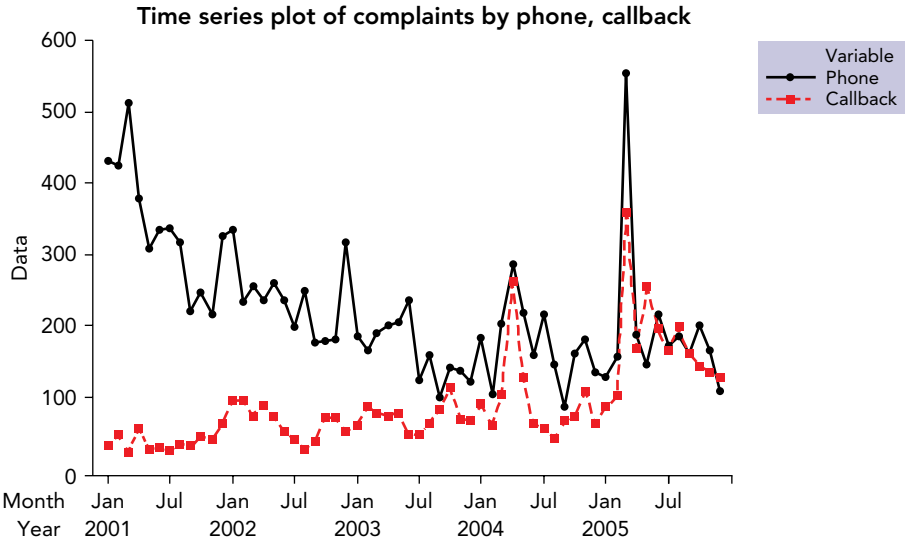
ISBN 978-1-107-65599-7

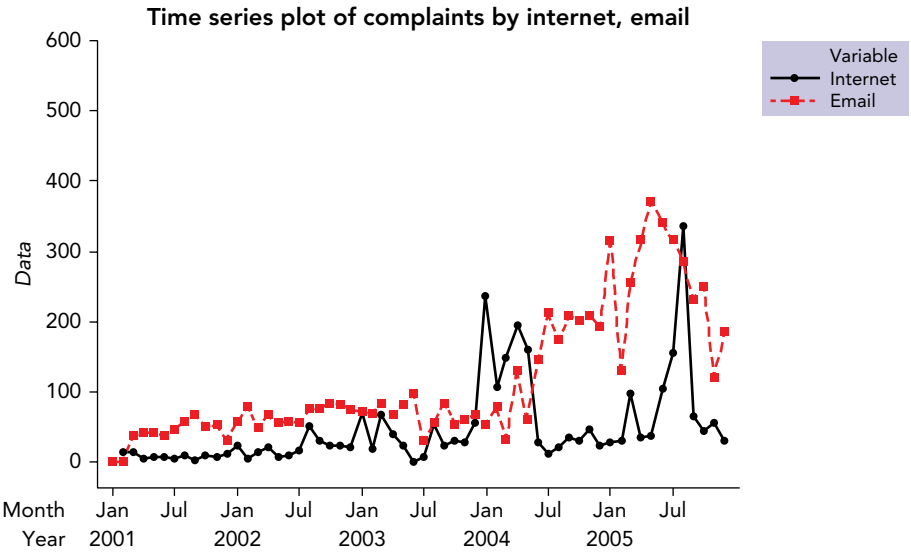
b Have some ways of making complaints become more or less popular?

c Are the numbers of monthly complaints variable or fairly steady?

d Are there any months or periods that stand out for large or small numbers of complaints?

e Are there links between the different ways of making complaints? That is, if there are more by one method, are there fewer by another?





ISBN 978-1-107-65599-7

Chapter summary

Scatterplots of two quantitative variables

- To investigate whether and how two quantitative variables are related to each other, we need pairs of observations for each subject
- Scatterplots plot points for the data pairs
- Which variable goes on the vertical axis and which on the horizontal depends on the context
- A scatterplot shows how much variation there is in the data as well as suggesting if the two variables are related and in what way

Scatterplots involving more information

- Different symbols can be used in a scatterplot to mark different groups
- Including more variables, possibly quantitative, requires clever graphics and special software

Plots against time

- For some data collected over time, scatterplots against time can show trends, patterns and variation over time
- Such plots are often called time series plots because the data are a series over time
- The points on a time series plot are usually joined to help see trends and patterns over time. If data are missing at a time point, there will be a break
- More than one data series can be plotted against time on the same plot

Multiple-choice questions

1 A scatterplot can be used for

A Any numerical data
B Two lists of numerical data with equal numbers
C Two lists of quantitative data
D Pairs of quantitative data

2 A scatterplot helps to explore and represent

A Relationships between quantitative variables
B Whether a quantitative variable is affected by another
C Variation in quantitative variables
D All of these

3 Which variable goes on the x-axis and which on the y-axis in a scatterplot depends on

A The units of the data
B The accuracy of the data
C The context of the data
D None of these

4 In a scatterplot

A The x-axis must start at 0
B The y-axis must start at 0
C Both axes must start at 0
D None of these

5 In a scatterplot, we don't join up the points because

A It takes too much time
B It gets in the way of seeing relationships and variation
C We need to join points to the axes
D None of these

6 A time series plot is a scatterplot in which

A Time is on the x-axis
B The interest is in patterns and variation over time
C The points are usually joined
D All of these

ISBN 978-1-107-65599-7

Short-answer questions

1 We have data on reaction times with the right hand and the left hand of a group of right-handed people. Also recorded are their gender (M, F) and their age group (<30 years, >30 years).

a Which data would you suggest putting on the x-axis? Why?

b How could you include the information about whether a subject is male or female, and whether they are younger or older than 30 years?

2 The world records for some women's track running distances are given in the table below.

Distance (m)	100	200	400	800	1000	1500	2000	3000	5000	10000
Time (s)	10.49	21.34	47.60	113.28	148.98	230.46	325.36	486.11	851.15	1771.78

a Draw a scatterplot of the record time against the distance.

b The plot looks like a straight line. Why can't it be a straight line?

3 Question 2 of Exercise 4C refers to an experiment in which four lists each of 25 words were tested on a group of 24 people to investigate lists of words that can be used in hearing tests. The scaled scores out of 50 for all four lists are given below.

Person	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Hearing test 1	28	24	32	30	34	30	36	32	48	32	32	38	32	40	28	48	34	28	40	18	20	26	36	40
Hearing test 2	20	16	38	20	34	30	30	28	42	36	32	36	28	38	36	28	34	16	34	22	20	30	20	44
Hearing test 3	24	32	20	14	32	22	20	26	26	38	30	16	36	32	38	14	26	14	38	20	14	18	22	34
Hearing test 4	26	24	22	18	24	30	22	28	30	16	18	34	32	34	32	18	20	20	40	26	14	14	30	42

ISBN 978-1-107-65599-7

a Draw a single scatterplot of the scores for hearing tests 3 and 4 against the scores for hearing test 1 – that is, the scores for tests 3 and 4 on the same graph.

b Do the scores seem to be linked – that is, do the 24 people tend to score similarly on the tests?

c How do the plots of scores for tests 3 and 4 against test 1 compare with the plot of the scores of test 2 against test 1?

d Review the scatterplot you drew in answer to question 1a of Exercise 5A, which refers to the same hearing test experiment. There are four points with very good scores on test 1. How did they score in tests 3 and 4? Can you see their scores in test 2?

4 Another question with total results for each survey posted on the ABS CensusAtSchool website (www.cambridge.edu.au/statsAC910weblinks), is on the number of hours sleep students usually get on school nights. The averages over all students who completed the survey in different years for Year 8 and Year 10 are as follows:

Cambridge GO

Year 8: in 2008, 8.7 h; in 2010, 8.6 h; in 2011, 8.5 h; and in 2012, 8.3 h.

Year 10: in 2008, 8.2 h; in 2010, 8.0 h; in 2011, 8.0 h; and in 2012, 7.7 h.

a Draw a time series plot of these results on the same plot.

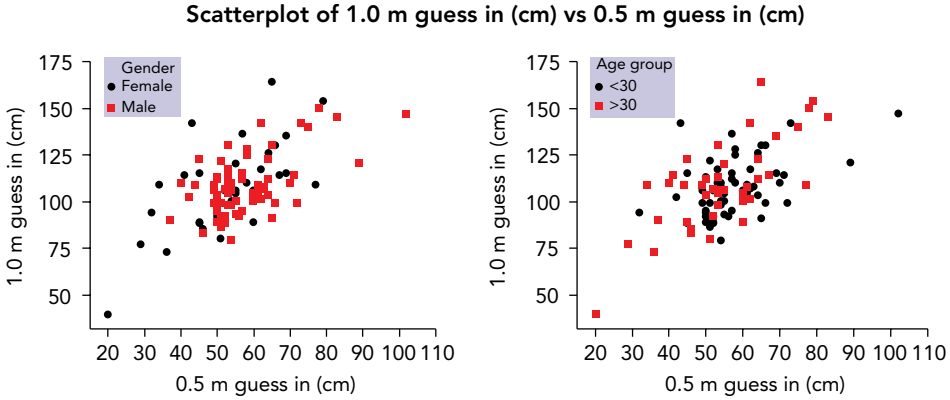
b Does there appear to be a trend or not?

c The medians were constant for each year at 9 h for Year 8 and 8 h for Year 10. What do you think the shape of the data for each school level might be?

d The time series plot of averages gives us no idea of the variation for each school level in each year. If the original data (or a subset) were available to you, what plots would you suggest doing to get an idea of how variable the number of hours sleep on school nights is, and how much it is varying over school levels and years?

ISBN 978-1-107-65599-7

5 Many people have some way of estimating lengths such as 0.5 m and 1 m. For example some people know the approximate length of their pace, although it's not as much as 1 m for most adults. An experiment was conducted to investigate how well people can estimate lengths of 0.5 m and 1 m. Each subject indicated their estimate of the two distances on a piece of string. An estimate of 0.5 m was given first and then an estimate of 1 m without using the 0.5 m estimate as a guide. The subjects were classified as being under 30 years old or more than 30 years old. Below are two scatterplots of people's guesses of 1 m against their guess of 0.5 m. The first scatterplot marks whether the subject was male or female and the second marks their age group.



a Why were the four groups formed by combination of gender and age group not used on just one scatterplot?

b Do people's guesses tend to be linked?

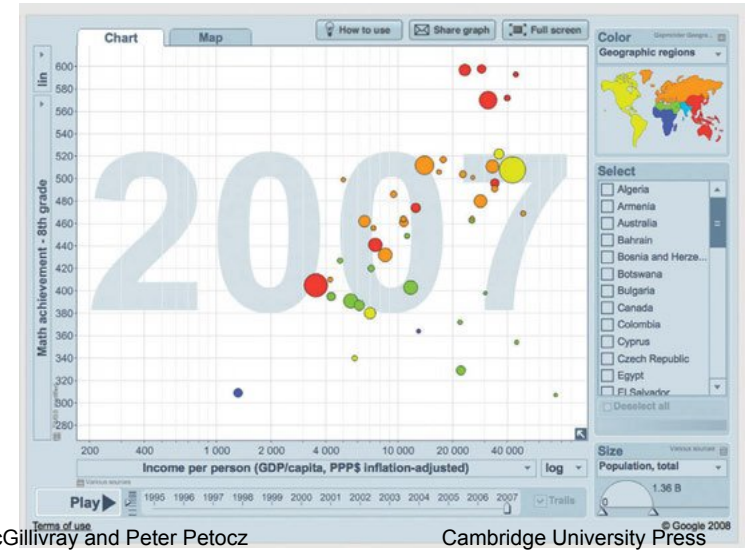
c What's the variation in people's guesses like?

d Do the plots suggest any differences between males and females? If so, what?

e Do the plots suggest any differences between under 30s and over 30s? If so, what?

6 The plot shown is a screen capture from Gapminder of one of the plots over time of Year 8 maths results and income per person. The size of the bubbles represents the relevant population and the colour of the bubble represents region of the world. The data are based on an international maths test for children in Year 8, from the TIMSS (Trends in International Mathematics and Science Study).

When the cursor passes over a bubble on the screen, the country's name appears. Some countries have been identified on this screen capture.



a Does the plot suggest there is any relationship between Year-8 maths results and average income in TIMSS? If so, how strong is this relationship?

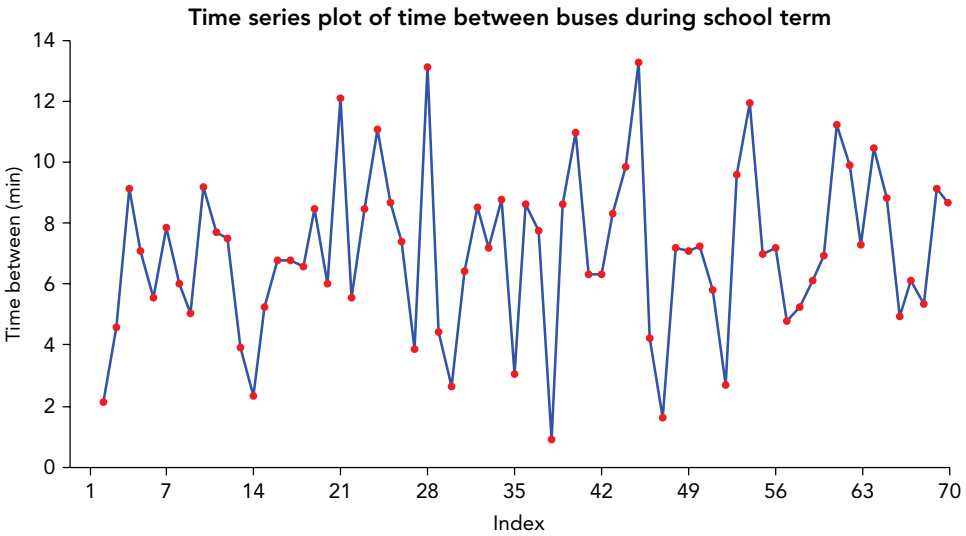
b What else does the plot tell us?

c What type of test is the TIMSS maths test? Does that also tell us something to help interpret the plot?

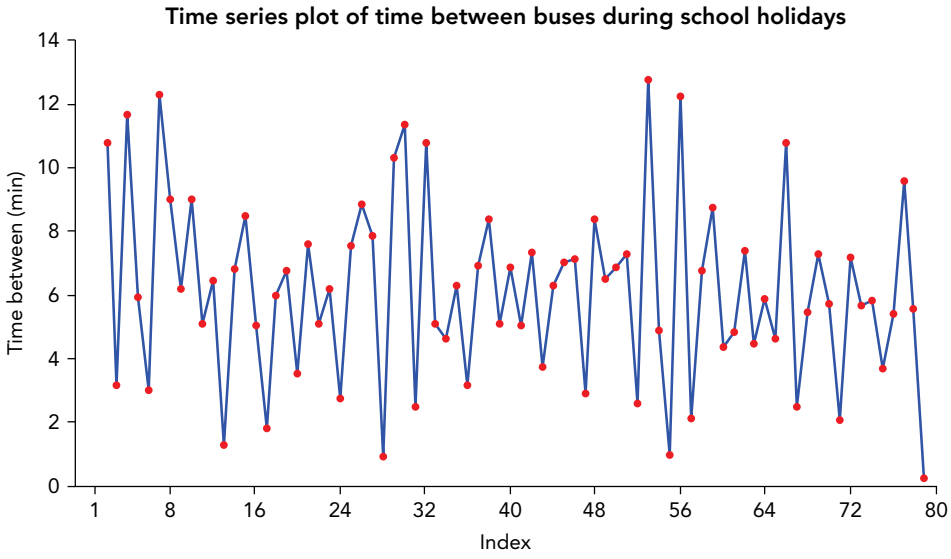
ISBN 978-1-107-65599-7

Extended-response question: City circle buses

7 Cities often have free city circle buses that travel continuously on a route around the central business district of the city (CBD). Schedules usually say approximately how much time there is between buses and give approximate times to certain stops. Usually more than one bus travels the route, and sometimes there may be many, depending on the length of the route and the amount of patronage. Below are time series plots for the time between buses at one stop on a city circle route, recorded for a day during school term, and a day in school holidays.



ISBN 978-1-107-65599-7



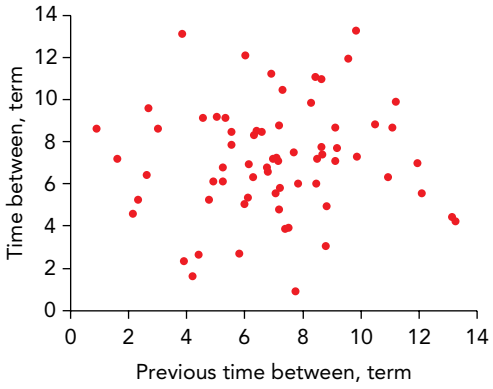
a What is on the *x*-axis in each plot?

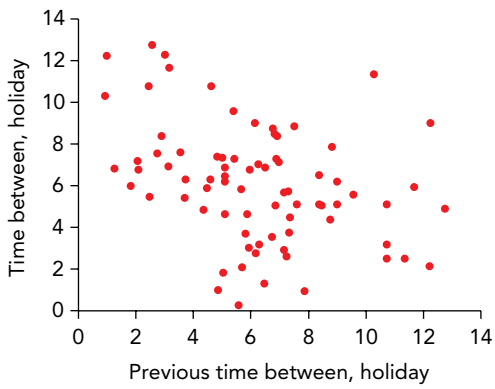
b Comment on the variability in each plot.

c What difference is there between the term time plot and the holiday plot?

d There are a lot of sudden changes – ups and downs – in these plots. Can you see what sort of pattern is happening? Is it happening more in one plot than the other?

e Below are scatterplots of the times between buses plotted against the previous time between buses for term and holiday. What can you see in these plots? Do they provide information that is helpful in answering part **d**? In what way?





ISBN 978-1-107-65599-7

Glossary

Chapter 1

Back-to-back stem-and-leaf plot: Two stem-and-leaf plots placed side by side with a common stem. The leaves on the left of the stem are for one of the plots but with the order of digits in the leaves reversed and aligned on the right. The leaves on the right of the stem are those of the other plot.

Bin: Interval of a histogram

Cumulative frequency plot: Joins up the tops of the steps in a cumulative histogram

Cumulative histogram: A graph of quantitative data that gives the number of observations less than or equal to the values on the horizontal axis. Every time another observed value is reached along the horizontal axis, the plot steps up by the number of observations that have that value.

Distributed: How the data are spread over the range of values

Histogram: A (simple) histogram is a graph of frequencies of quantitative data grouped into equal intervals which cover the range of the data. Rectangles above each interval show the number (frequency) of observations in that interval. Rectangles must share sides (unless an interval contains no observations) as the horizontal axis represents intervals of values. In contrast, the bars of a bar graph must not touch each other as the horizontal axis represents categories. More sophisticated histograms are scaled so that the total area of the rectangles is 1.

Placebo: A dummy treatment; a treatment in which no treatment is given but may pretend to be one; in health, placebos have no active ingredients

Same scale: Refers to plots having the same range of values on the *x*-axis and the same distances between these values; can sometimes also include same scale on the *y*-axis but usually that is specified

Chapter 2

Asymmetric: Data that are not **symmetric**; some data are asymmetric but cannot be readily described as skew to the left or to the right

Bimodal: Ordinal data (or ungrouped count data) with two categories (or count values) with higher frequencies than their neighbours. Quantitative data which have two clearly indicated clumps of data may be said to look bimodal. This may indicate that there are different **sub-groups** within the data.

Multimodal: When quantitative data have more than two clearly indicated clumps of data, this may indicate that there are different **sub-groups** within the data.

Skewed to the left: Refers to quantitative data in which the variation of the data for values greater than the 'middle' is less than for values less than the 'middle'

Skewed to the right: Refers to quantitative data in which the variation of the data for values greater than the 'middle' is more than for values less than the 'middle'

Sub-groups: Groups of observations within data which are different and may refer to different categories of a categorical variable

Symmetric: Refers to quantitative data in which the variation of the data for values greater than the 'middle' is the same as for values less than the 'middle'. This is almost always approximate because we have to allow for sampling variation.

Tails of data: Refers to spreading out of data for the smaller or larger data values. A data tail has low frequencies spread over a range of small (left tail) or large (right tail) data values.

Unimodal: Ordinal data (or ungrouped count data) with only one category (or count value) with greater frequency than its neighbours. Quantitative data which have one region in which the data are most clumped may be said to look unimodal.

Chapter 3

A and B: A situation where both events A and B occur

A or B: A situation where at least one of the events occurs (**inclusive or**) or exactly one of the events occurs (**exclusive or**)

Complementary: Two **disjoint events** that cover the entire sample space

Compound event: Event that consists of several simple events or individual outcomes; if the individual outcomes are equally likely then the probability of a compound event can be found as the proportion of outcomes it contains

Conditional phrases: Phrases such as 'if ...', 'given that ...' and 'knowing that ...' are indications that we are looking for conditional probabilities.

Conditional probability: Probability based on a smaller reference group – a sub-group of the original sample space or set of data

ISBN 978-1-107-65599-7

Disjoint or **mutually exclusive events:** Outcomes that cannot occur at the same time

Equally likely outcomes: Outcomes of an experiment that are equally likely to occur

Event: Any individual outcome or collection of outcomes

'Exclusive or': Exactly one of events A and B occur

Experiment: Any situation where we don't know the outcome in advance

'Inclusive or': At least one of events A and B occurs

Independent events: Two events A and B are independent if P(A and B) = P(A) × P(B)

Mutually exclusive or **disjoint events:** Outcomes that cannot occur at the same time

Probability: A way of measuring chance, represented by a number between 0 and 1

Sample space: A list of possible outcomes

Tree diagram: Used to display the results of a two- or three-stage experiment (or even more stages, but then the tree gets complex); probabilities can be written on each connecting line and multiplied together to get the probability for each branch

With replacement: Item selected at the first stage is replaced before the next selection

Without replacement: Item selected at the first stage is not replaced before the next selection

Chapter 4

Box-and-whiskers plot: Original and full name for the **boxplot**

Boxplot: A graph for quantitative data (usually data from a continuous variable) which divides the range of values of the data into intervals with a quarter of the data in each. The simplest form of boxplot presents the five-number summary with the box going from the lower to the upper quartile, the median marked in the box, and the lines from the edges of the box extending to the minimum and maximum. A better version has the lines (the **whiskers**) extending to the last data value within a certain distance from the box, with smaller and larger observations marked by a star or an asterisk (*). The usual value for this distance is 1.5 times the interquartile range – the length of the box. Some boxplots mark observations that are more than 3 times this distance from the box edges by another symbol such as an *.

Extrapolation: Refers to ways of approximating values between given or known values

First quartile: Another name for **lower quartile**

Five-number summary: The set of numbers consisting of the median, the upper and lower quartiles, and the minimum and maximum

Hinges: Edges of the box – the lower and upper quartiles

Interquartile range: The difference (upper quartile) – (lower quartile)

Lower quartile: This has a quarter of the observations less than it; usually calculated as the median of the observations less than the median of the data

Quartiles: These divide a dataset (including the data median) into four groups with equal numbers of observations in each

Theoretical or population median: A special quantity for a general situation or population. A random observation from the general situation or population is equally likely to be greater or less than the theoretical or population median

Third quartile: Another name for **upper quartile**

Upper quartile: This has a quarter of the observations greater than it; usually calculated as the median of the observations greater than the median of the data

Whiskers: Lines extending from the edges of the box

Chapter 5

Pairs of observations: Data on two variables that are observed for each subject; each pair of observations belongs to one subject

Scatterplot: A plot used to explore relationships between two quantitative variables. It plots pairs of observations with the first value in a pair being the value on the *x*-axis, and the second of the pair being the value on the *y*-axis

Time series plot: Scatterplot against time in which we are looking for trends, patterns and variation over time. The data on the *y*-axis can be individual observations or averages or medians or percentages; can often plot more than one series on the same plot

Trends, patterns and variation over time: Indications that the data may be increasing or decreasing over time, or that there might be a pattern that occurs again, or that variation might be changing over time

***x*-axis:** The horizontal axis of a scatterplot

***y*-axis:** The vertical axis of a scatterplot

ISBN 978-1-107-65599-7

Answers

Chapter 1

Pre-test

1 Identify issues or questions; identify what can be measured or observed; identify variables and their types; identify subjects; design recording data sheet or spreadsheet, including agreement on and naming of categories of any categorical variables; plan how to collect randomly representative data; carry out a pilot

2 a, b
Amount of nutrient (continuous); location (categorical); day of week (categorical); week (categorical).

c Water sample

3 a Experiment

b Time to burn, sunscreen, gender, age, skin type

c Definition of time to burn; precise amounts of sunscreen on each arm; extent of rubbing in; same time before going out into sun.

4 a Observational study

b Number of passengers getting off (count); time of day (categorical); day of week (categorical); week (categorical).

5 a Survey **b** Bus passengers travelling to city

c Stratified

6 a

Leaf unit = 10

0	000011
0	23
0	445555
0	6
0	9
1	00
1	3
1	
1	
1	8

b Data mean = 55.15, data median = 48, data range = 189 – 1 = 188.

c With reference to the stem-and-leaf plot in answer (a), two: 0 and 50. In original, 20: all 20 values occur once each.

Exercise 1A

1 a Amount of sway in each direction before noise and at time of pushing button. Also record age group and gender.

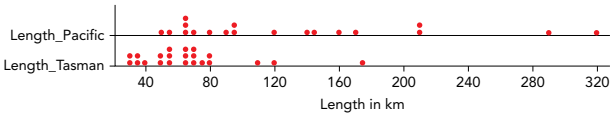
b Calculate difference in sway (after – before) for each direction.

c There are two differences for each subject. Dotplots and histograms on the same scale can be used to plot each of these differences split by age group and gender. Back-to-back stem-and-leaf plots can be used for each of these differences split by gender either within each age group or with age groups combined.

2 a Take pulse rates before and immediately after exercise; make sure the 'after' measurements are taken at exactly same amount of time after exercise finished. Calculate differences (after – before) to use to compare the two groups.

b Pulse rates are measured same way, e.g. over same time period for all subjects and before and after. Exactly the same exercise for the same length of time and some way of controlling intensity of exercise. 'After' measurements taken at exactly same amount of time after exercise finished.

3 a



b

Leaf unit = 10

Pacific		Tasman
	0	3333
54	0	445555
7666	0	6666777
9998	0	88
	1	0
32	1	2
4	1	
66	1	7
	1	
00	2	
	2	
	2	
	2	
8	2	
	3	
2	3	

ISBN 978-1-107-65599-7

c There are more longer rivers flowing into the Pacific Ocean. Probably because the largest mountains tend to be closer to coast on Tasman side; very much so in the South Island.

4 a

Leaf unit = 10.0

Alt rock		Indie
	1	3
5	1	
	1	
999888	1	889999
1100	2	00000111
32	2	22222333333
555	2	455555
7	2	77
9	2	8
	3	
2	3	3
	3	
6	3	
	3	
0	4	
	4	
	4	
	4	
98	4	

b There are 36 Indie songs. From stem-and-leaf, 18th and 19th songs are the 3rd and 4th values of 220 s. Looking at original data, these are both 226 s, so the median of the lengths of Indie songs is 226 s.

c $\frac{11}{23} = 26\%$

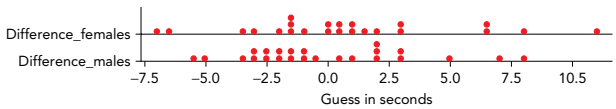
Enrichment

5 a Order of guessing – that is, whether reading first or second

b Passage being read; stopwatch

c The differences between guesses for each person: guess when reading – guess when not reading.

d Dotplots and back-to-back stem-and-leaf plots for (guess when reading) – (guess when not reading).



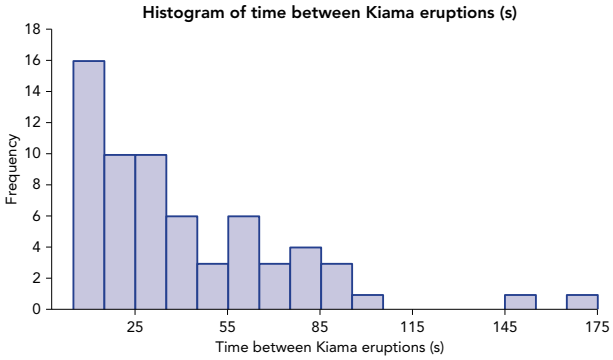
Leaf unit = 1.0

Differences_males		Differences_females
	-0	66
45	-0	
222233	-0	322
000111	-0	11100
1110	0	000011
332	0	223
5	0	
7	0	66
8	0	8
	1	0

e The differences are more variable for the females than the males; males have more negative differences than females. A negative difference is when the guess when reading is less than the guess when not reading.

Exercise 1B

1



a This one has maximum frequency in first bin, which is 5–15 s. One in Example 2 has maximum frequency in second bin, which is 10–20 s.

b $\frac{10}{64} = 0.15625$ or 15.625%

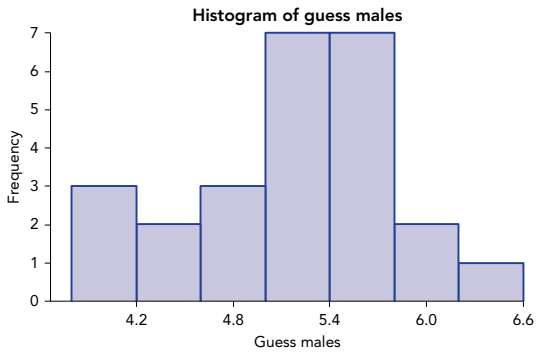
c No

d $\frac{17}{64} = 0.2656$ or 26.56%

e For part b, $\frac{12}{64} = 0.1875$ or 18.75%. For part d, $\frac{17}{64}$ as above. For part b answer is not the same because there are two values of 25 which are put in the next bin in histogram. For part d, there are no values of 20.

ISBN 978-1-107-65599-7

2 a



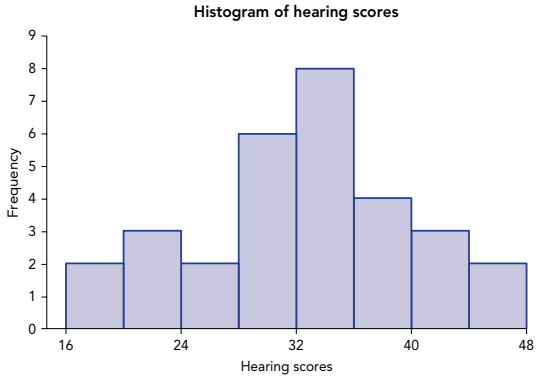
b

c The second one gives the impression of more smaller observations, but apart from that they're not very different.

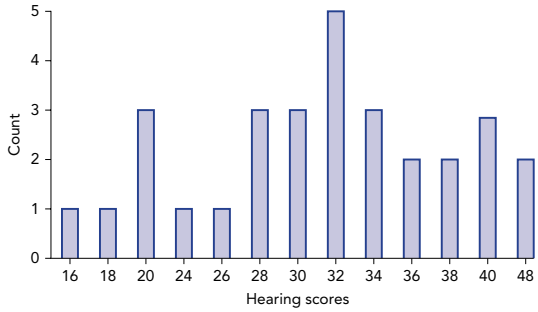
d $\frac{12}{24} = 0.5$. It will be the same because they both have bins starting at 4.2 and 5.4.

e $\frac{13}{24} = 0.542$. There is one value of 5.40, which is in the next bin in the histogram.

3 a Starting point below is 16, and bin size is 4.



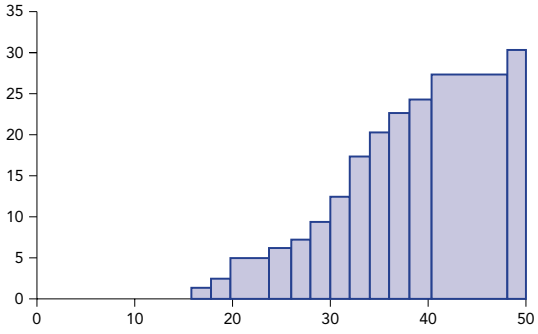
b Count



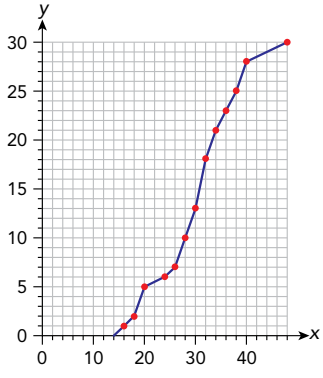
d Because they are count data, which take integer values, so a bar chart is suitable. But because they are quantitative and take a number of different values, a histogram is also suitable.

e This is personal preference; the histogram gives a smoother picture but the bar chart has all the original information.

4 a Cumulative histogram



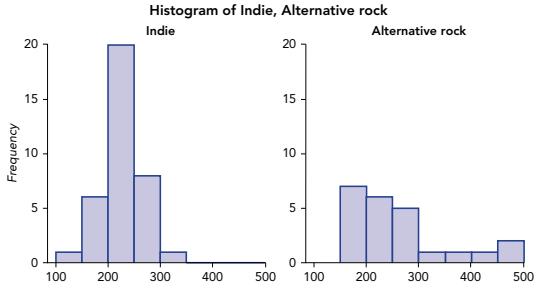
b Cumulative frequency plot



ISBN 978-1-107-65599-7

Enrichment

5 a



b Personal preference. An advantage of the histograms compared with the stem-and-leaf plots is their smoothness and they emphasise how the Alternative rock lengths tend to be clumped together at the beginning of their range. A disadvantage of the histograms compared with the stem-and-leaf plots is the loss of detail.

c The scale on the *y*-axis would change. The heights of the boxes would be the current heights/*n*, where *n* is number of observations, so the relative heights between the two histograms would change because there are a different total number of observations in each. The shape of the histograms would stay the same.

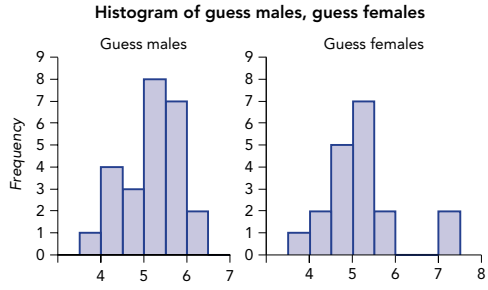
d The heights would need to be frequency/(number of observations × bin width). Because then the area of each rectangle would be frequency/number of observations, and the sum of the areas would be 1.

e There are 23 Alternative rock songs, and 5 in length from 300 to 500. So we need the area of the last rectangle to be $\frac{5}{23}$. So the height of that rectangle needs to be $\frac{5}{(23 \times 200)}$. The heights of the other rectangles will be as in part d – $\frac{\text{frequency}}{(23 \times 50)}$.

f There are some long Alternative rock songs but most are of similar lengths to Indie songs.

Exercise 1C

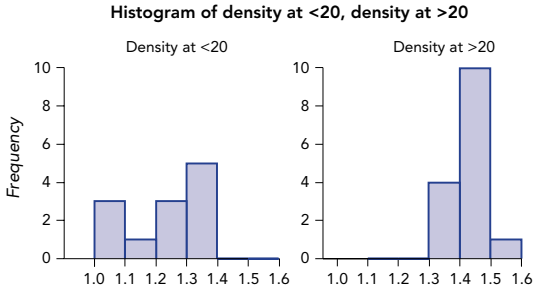
1 a



b Both male and female guesses vary around 5 m, with a few of both genders guessing at 4 m or below and up to 6 m with two females guessing more than 7 m. More males overestimated 5 m than underestimated it. The stem-and-leaf plot shows that a number of females slightly underestimated and overestimated, with few being very close.

c Personal preference but possibly stem-and-leaf in this case because of the small numbers of observations and the groupings of the female guesses which are lost in the histograms.

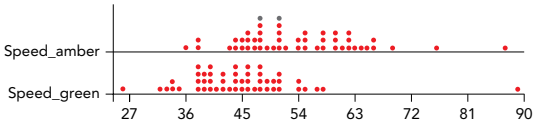
2 a



b The densities <20 km are more variable than those >20, with a number of densities either close to 1 or close to 1.3, while the densities at >20 tend to be more clumped around 1.4.

c Very few observations so probably the stem-and-leaf although the histograms do not lose much essential information in this case.

3 a A dotplot of the speeds on the same scale is below.



b Almost all the approaches to green lights are below, and mostly well below, the speed limit except for one extraordinary speed of almost 90 km/h. A number of approaches to amber lights were above the speed limit.

c The data plots do indicate that approach speeds to amber lights tend to be greater than to green lights, although most of the approach speeds to amber lights are below the speed limit.

4 a The internet is the least popular way of making complaints, with most months having fewer than 60 made this way. Complaints by phone are the most popular way, but there's a lot of variation in the

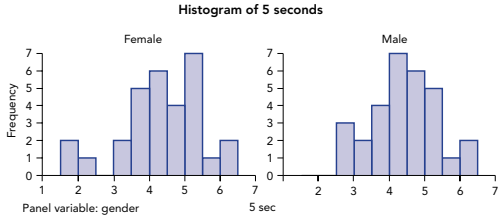
ISBN 978-1-107-65599-7

number over months. Email complaints are fairly regularly between 50 and 80 a month but there are some months with many email complaints, possibly similar in number to phone complaints.

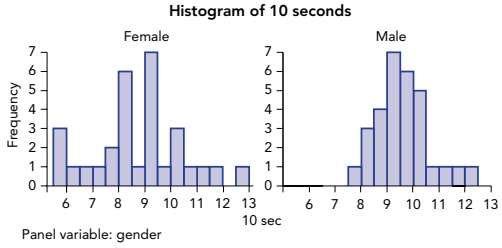
b Probably the dotplots as the behaviour is quite clear in this case in the dotplots and less so in the histograms.

c The rows correspond to months and each column gives the number of complaints in each month by the different ways. So the above observations are not in different groups – they are related to each other by months.

5 a i



ii

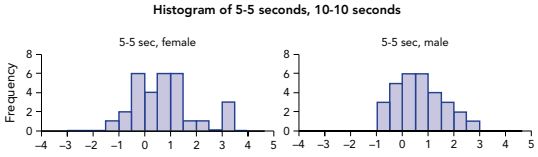


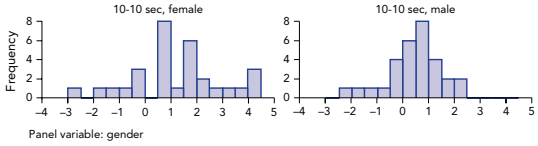
b They tended to underestimate both time intervals.

c Females were a little more variable in guessing 5 s, but with fewer underestimates. Females had much more variation than males in guessing 10 s.

d The variation in the females' guesses was much greater than the males for 10 s compared with 5 s.

e Can subtract guesses from the targets (5 s and 10 s) and plot these as below. These plots capture all the points above – the underestimation, the comparison of spreads and the comparison between males and females for each target and as they moved from guessing 5 s to guessing 10 s.





Chapter summary

Multiple-choice questions

1 C **2** B **3** D **4** C

5 C **6** A **7** D

Short-answer questions

1 Lengths of NZ rivers; lengths of songs in top 100; if used by investigators other than Cavendish, the Cavendish data on the density of Earth; complaints to Sydney airport.

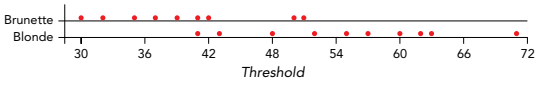
2 a As well as the usual care in taking the measurements and measuring the dose, the times between squeezes and between the dose and the post-dose squeeze would need to be carefully monitored.

b Yes, because it is a count variable and we can choose to collect the values in intervals if we wish.

c The number of repetitions until 50% of each person's maximum initial strength was reached for those who had the vitamin C and those who didn't.

d Use the same scale on the *x*-axis (and the *y*-axis for histograms).

3 a



b

Leaf unit = 1.0

Brunette		Blonde
20	3	
975	3	
21	4	13
	4	8
10	5	2
	5	57
	6	023
	6	
	7	1

ISBN 978-1-107-65599-7

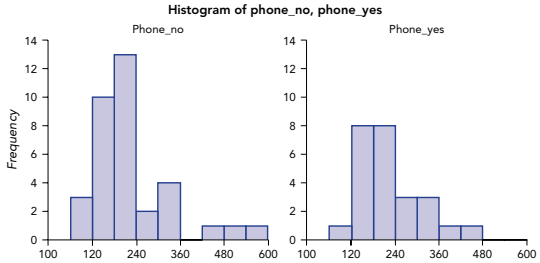
c Provided the subjects were randomly chosen, the plots indicate that the pain threshold is generally higher for blondes than for brunettes.

4 a

Leaf unit = 10

School holidays		No school holidays
	0	8
432220	1	00344
8888776	1	55566678889
43100	2	0011123334
85	2	6
321	3	0133
7	3	
3	4	2
	4	
	5	1
	5	5

b



c There's not much difference between months with school holidays and months without. The months with most complaints do not have school holidays but this may just be due to variation as there are more months without than with school holidays.

d Either are fine in this case.

5 a To be able to compare how the actual weight varies from the stated weight across the different brands

b Almost all packets have at least the stated weight. The 19 g packets are the most variable and the 20 g packets the least. The 20 g packets are also the most generous in general in having slightly more than the stated weight.

Extended-response question

6 a The bread showed mould before the expiry date.

b Very few thick slices so it's hard to comment on them. For the thin slices, being in sunlight tended to create mould sooner but there were a few extreme values.

c The variation out of sunlight seems greater than in sun for both thick and thin. This seems to be the main difference between sun and no sun. Generally the thick bread seems to show mould a little sooner than the thin bread but there's not much difference.

d Very few thick slices

e Brand A observations seem to be mostly less than 1.5 days, while brand B vary around 2.5 days for thick and mostly greater than 2.5 for thin.

f Same scales

g Can only compare two groups at a time in one plot using back-to-back stem-and-leaf plots.

Chapter 2

Pre-test

1 a No information on units and accuracy. It also looks like the units being used changed during the data recording.

b Different number of decimal places quoted so we don't know the accuracy or if it's the same across the observations. For example, is 160 = 160.00?

c The amount of protein has to be for a certain weight or size of cereal portion and this information is not given.

2 a The difference between actual and scheduled in one column and the name of the airport in the other. Each row contains the observations for a departure of a plane.

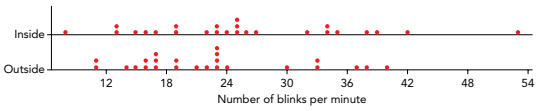
b The actual time could be in one column, the scheduled in another and the airport name in the third. Or the differences between actual and scheduled for each airport could be in a different column.

3 a Length is a continuous variable. Pie charts are not used for data from continuous variables.

b Dotplot, stem-and-leaf plot, histogram

ISBN 978-1-107-65599-7

4 a

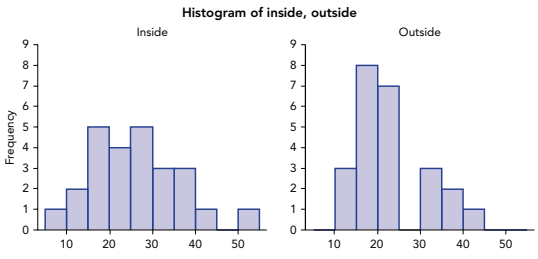


b

Leaf unit = 1.0

Outside		Inside
	0	8
411	1	33
99777665	1	56799
4333321	2	2334
	2	55567
330	3	244
87	3	589
0	4	2
	4	
	5	3

c



d

Variable	N	Mean	Median	Range
Inside	25	25.88	25.00	45.00
Outside	24	22.58	21.50	29.00

Exercise 2A

1 Yes. They are skewed to the right and strongly skewed.

2 The data on densities at <20 are somewhat skewed to the left but not strongly. The data on densities >20 are reasonably symmetric.

3 a Yes

b Skewed to the right

c No, because the skewness is quite strong.

4 a Yes

b Both are skewed to the right.

c The lengths of rivers flowing into the Pacific tend to be more skewed to the right than the lengths of those flowing into the Tasman.

5 Although the histogram doesn't look it, the stem-and-leaf plot indicates that the male guesses are reasonably symmetric. The female guesses are asymmetric but it is difficult to describe them as skewed to the left or right. These are only small groups of data.

Enrichment

6 a Both groups are skewed to the left.

b Group 1 (Mon–Wed) is more skewed to the left than group 2 (Thurs–Sun)

c Possibly because the groups involve different days and different service stations

Exercise 2B

1 a Because of the small number of observations, the stem-and-leaf plots are better to look at. For the females, there is a group above 5 m and another group below 5 m, and the data do appear bimodal.

b No, both the stem-and-leaf plot and the histogram indicate unimodality.

c There does not appear to be any reason for bimodality – it's a small sample so we shouldn't read too much into it.

2 a Yes

b Because the two datasets are so differently centred. It looks like one clump would be about 1.30 to 1.34 and the other about 1.40 to 1.44.

3 a The first and possibly the third might mislead into thinking the data are bimodal. The third or fourth might mislead into thinking multimodal but they are obviously very bumpy.

b 48 is not a large dataset. When 48 are spread over 9 bins, the number in each bin will not be large so different starting points will make a difference to the appearance. For more than 10 bins, the data are spread too thinly and histogram appearance will be very much dependant on choice of stating point.

4 a There is too much overlap of the datasets and their intervals as most observations are the same or very close to each other.

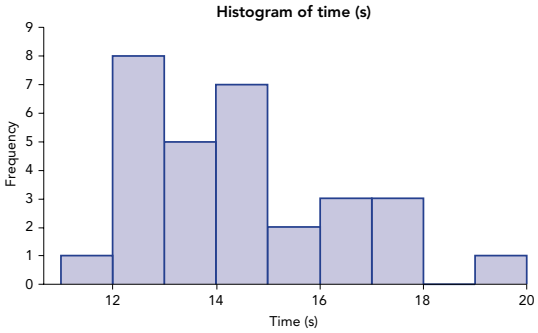
b Yes, and skewed to the right because they are both skewed to the right and their intervals with most observations have similar values.

ISBN 978-1-107-65599-7

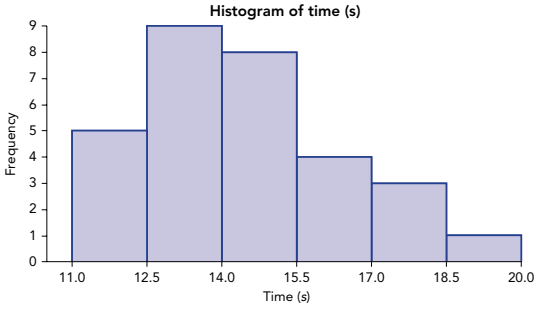
5 There are obviously cheaper days and more expensive days, but the prices need to move between highs and lows and will probably not change much in-between highs and lows. So each of these two intervals may be giving the general location of the prices between highs and lows.

Enrichment

6 a i Can't tell with only 30 observations – the two bins 12–13 and 14–15 may just be due to histogram choice or sampling variation



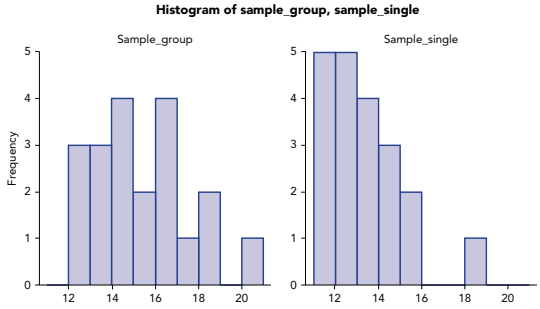
ii Definitely not bimodal



b There is no suggestion of bimodality.

c The plot suggests there may be two groups – that is, there may be bimodality – with one group centred between 12 and 13, and the other between 15 and 16.

d



The singles tend to walk faster and there are many times between 11 s and 14 s. The data are skewed to the right. The times of those walking in groups are more variable, and many are fairly evenly spread between 12 s and 17 s. This is why when the two groups are combined, the combined histogram suggests two groups but not strongly. The separate histograms clearly show differences between the groups, in spread and shape in particular.

Exercise 2C

1 a

Variable	N	Mean	Median	Range
cappuccinos	30	4.13	4.20	1.20

The data are skewed to the left, and the mean is less than the median. The half of the prices that are more than \$4.20 are all between \$4.20 and \$4.70. The half of the prices that are less than \$4.20 are between \$3.50 and \$4.20.

b The data look quite skewed to the left in the histogram, but the mean is only 7 cents less than the median. Would probably expect a greater difference from the histogram appearance. However 30 observations is small, so a different histogram of the data might look different.

2 a The range of time differences is quite large – 9 min 48 s. Half the people have their watch time at least 15 s faster than the actual time. The average amount people's watches are fast is 25.9 s, which is quite a bit more than the median.

b The histogram is unimodal with the most commonly occurring interval in this histogram centred on 0 – that is, on watches set correctly – but the median is in fact 15 s. Watches vary from being about 300 s slow to about 300 s fast. The histogram is perhaps slightly skewed to the right, and the mean is about 11 s more than the median.

ISBN 978-1-107-65599-7

c This histogram is only slightly skewed to the right but the mean is 11 s greater than the median. Because there are 101 observations, it would be interesting to see a dotplot or a stem-and-leaf plot to see more about the shape.

3 a The watches of females are more variable, ranging from about 300 s slow to 250 s fast. The watches of males range from about 200 s slow to 300 s fast, but there are more differences closer to 0 for the males. Both datasets are skewed to the right, but the data for females are more skewed, and their mean is 14.6 s greater than their median. The mean for the males is only 4.6 s greater than the median.

b Not much as the two groups are too close together in location, so the overall dataset doesn't suggest two groups. The features of the overall dataset are in-between the features of the two datasets without being very different.

c Not overall or in the male and female data. All three datasets have many observations not too far from 0. Even if different histograms would give some differences in appearance, there's no suggestion in any of these three that bimodality would appear.

4 a Count variables. So the medians will be whole numbers or halfway between whole numbers. The ranges will be whole numbers. However, the means could have decimal places, as we see here.

b The phone complaints are spread on either side of about 200 per month. They are slightly skewed to the right, with the mean slightly more than the median. Most months have the number of email complaints less than 80, and half the months have fewer than 75 complaints. But there is great variation in the number of complaints for months with more than 80 complaints, ranging up to almost 400, so that the average number is about 115 per month. There are even fewer internet complaints, with most being below about 40 per month, and half the months have fewer than 27 complaints per month. However, the months with more than 27 are quite variable, so that the average number is about 47 per month.

c There are many more phone complaints per month than either email or internet, and they are reasonably spread on either side of about 200 per month. The numbers of internet and email complaints are not just many fewer on average; most months have only small numbers – fewer or much fewer than 100. However, there are some months with large numbers so that the averages of 115 for email and 47 for internet are a little misleading.

Enrichment

5 a So that we can compare nutritional information across cereals.

b Because we are comparing different substances – we don't expect to get the same amounts of these different substances in 100 g cereal.

c Probably to the second decimal place for protein because the median looks to be halfway between 9.29 and 9.30. To the first decimal place for carbo and d. fibre because their medians end in 0.2 and 0.3 respectively. To the first decimal place for iron, because the range ends in 0.7 and the median in 0.0.

d Half of the cereals have protein per 100 g below about 9.3. Those above this have protein amounts spread out up to about 24 g per 100 g, but the average amount is only a little more than the median at about 10. Half of the cereals have carbohydrate amounts more than about 74, and the average is about 73, but there are a number with much lower carbohydrates, and there may be two groups of cereals. Half the cereals have d. fibre less than 8.3 and the average is only a little more at about 8.4, but there are a few with much higher d. fibre. The cereals that do give amounts of iron are in two groups – one group has levels about 6.5 and the other has levels about 10; the average and median do not give much information about iron levels.

e For each variable there are cereals that are quite different from others. There are a few cereals with much higher protein than others; there are a few cereals with much lower carbohydrate than others; there are a couple of cereals with much higher dietary fibre than others; and those that report iron content mostly fall into two groups – low iron or high iron.

f Iron and possibly carbohydrate. Some cereals may be especially designed to focus on some substances rather than others.

g We cannot see anything about how these variables relate to each other. For example, what do cereals with high protein tend to have for the other variables?

ISBN 978-1-107-65599-7

Chapter summary

Multiple-choice questions

1 D **2** C **3** B **4** C

5 C **6** D **7** D

Short-answer questions

1 a Yes, because the two categories with the most responses are separated.

b Most respondents strongly disagreed, but the next most chosen category was agree. These two categories had many more responses than disagree and neutral, with very few choosing strongly agree. So the feeling was quite split between those who thought it was a waste of money and those who didn't.

c It does in this case because it summarises the situation – the response was bimodal.

2 a Asymmetric. Difficult to describe as skewed to left or right because there are a few measurements spread out on left of 5.25, but the ones on right of 5.25 look skewed to right

b Unimodal; there is no suggestion of two groups.

c

Variable	N	Mean	Median	Range
Density Earth	29	5.4197	5.46	1.79

The mean is slightly less than the median, mainly because of the single observation that is much smaller than the rest.

3 a Would probably describe it as skewed to the left as the left-hand side is slightly more spread out

b There may be a group of scores less than the main group but there's not enough of a clump to call the data bimodal. Would probably be reluctant to call it anything definite.

c

Variable	N	Mean	Median	Range
Hearing scores	30	31.47	32.00	32.00

There is a wide range of scores out of 50, ranging from 16 to 48. The average and median are very close to each, at about 32, which also looks close to the centre of the histogram.

4 a For the Indie songs, close to symmetric; allowing for choice of histogram, there's not much difference between right- and left-hand sides. For the Alternative rock songs, skewed to the right because the lengths are spread out to the right from the first interval.

b Unimodal for Indie as data clustered around centre. Probably unimodal for Alternative rock with mode at left-hand end; there are a couple of extra-long songs, but only a couple.

c

Variable	N	Mean	Median	Range
Alt rock	23	259.8	220.0	347.0
Indie	36	227.06	226.00	202.00

The average length of the Alternative rock songs is almost 33 s longer than the average Indie song, but the medians are closer. In fact, more than half the Indie songs are in the top half of the Alternative rock songs. The lengths of the Alternative rock songs are more variable than those of the Indie songs.

5 a The 19 g and 21 g data look skewed to the right because the data look more spread out on the right. The 20 g data look skewed to the left because they look more spread out on the left.

b All three are unimodal as there is no suggestion of clumps of data other than one main clump.

c The average is more than the median for the 21 g data as expected with skew to the right, but the average for the 19 g data is less than the median. This is surprising looking at the histogram, but it must be because there are so few observations greater than the group 1.5–2.5 g. The 19 g data also is the most variable. It is also surprising that the average for the 20 g data is greater than the median, but these data are the least variable so the histogram may be slightly misleading. The 19 g are the most variable and the 20 g the least variable. The 19 g data also has more underweight packets, followed by the 21 g, with the 20 g data the best in meeting or exceeding the stated weight.

6 a Not one with equal bins

b Because the data have been collected into very wide categories of unequal interval lengths. There is not enough information to calculate these summary statistics.

c We could possibly approximate the median but it would be a very rough approximation.

Extended-response questions

7 a For brand A, not enough thick bread to describe. For thin bread, skewed to the right for those left in the sun as more spread out to the right. For

ISBN 978-1-107-65599-7

those not in the sun, asymmetric but difficult to describe as skewed to left or to right. For brand B, for thin bread, skewed to right for both yes and no to sun as more spread out on right. For thick bread, asymmetric but too bumpy to be able to be described as skewed to the right or left.

b For brand A, not enough thick bread to describe. For thin bread, unimodal although for those left in sun, there is a group of small values and a couple of large values, but not enough to call it bimodal. For brand B, the observations are more variable than suggesting groups. The data are bumpy but there don't appear to be definite clumps of data.

c Not enough data for brand A thick bread. Brand A thin bread in the sun is most variable, while brand B thin bread in the sun is least variable. Brand B bread lasts longer than brand A for both thick and thin, whether in sun or not. Brand B thin bread lasts longest whether in the sun or not, with average and median of 4.5 days out of sun, and average of 3.45 and median of 3 days in the sun. Being in the sun does tend to speed up formation of mould. For Brand B, the averages are all greater than the medians. This suggests skew to the right – hard to see in bumpy histograms. Perhaps, surprisingly, brand A thin bread has averages less than the medians as those histograms look a bit skewed to the right.

8 a For the lengths, they are a bit difficult to describe – possibly skewed to the right. For the angles, the advanced and intermediate are close to symmetric, but the beginner graph is skewed to the left.

b For the lengths, there appears to be bimodality for the advanced and intermediate, but the beginners look unimodal. For the angles, the advanced and intermediate are bimodal; the beginners are more unimodal.

c For the lengths, the top group are clearly boundaries, so there's a clump of data for the non-boundaries and then there are the boundaries. For the angles, the two apparent groups are angles close to 0 and 360 degrees, so these refer to hitting the ball in front – the bimodality is simply because putting the data on a line from 0 to 360 unwraps the wheel – a wagon wheel would give a better picture.

d The beginners do not hit as many balls, and the average and median lengths are quite a bit less than for intermediates and advanced. The advanced and intermediate groups have the same median length, but the advanced longer average is probably because they hit more boundaries.

e The advanced group hit to both sides of the wicket, while the intermediates hit more towards the on side if right-handed (the side where the batsman's legs are), and the beginners hit most towards the on side if right-handed. The summary statistics are of little assistance because the angles refer to a circle not to a line. We also need to know if a batsman is right- or left-handed.

Chapter 3

Pre-test

1 a 5% **b** 0.95 **c** $\frac{1}{6}$

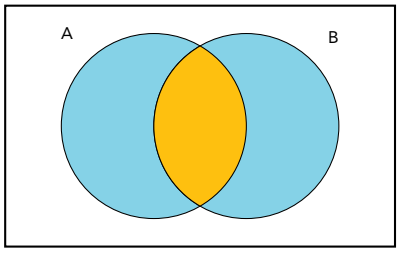
2 $\frac{3}{50}$

3 'don't win any prize'

4 $\frac{2}{38} = \frac{1}{19}$

5 a 'A and B' is shaded orange

b 'A or B' exclusive or is shaded blue



Exercise 3A

1 a {all 5 come, 4 come, 3 come, 2 come, 1 comes, none come}

b {ppp, ppf, pfp, fpp, ffp, fpf, pff, fff} p = pass, f = fail for the three subjects in order

c {0, 1, 2, ..., 12} number of bottles 'corked'

2 a unlikely **b** unlikely

c as likely to happen as not **d** unlikely

e certain **f** likely

3 a $\frac{5}{80} = \frac{1}{16}$ **b** $\frac{4}{79}$

ISBN 978-1-107-65599-7

4 **a** 0.5 **b** 0.3 **c** 0.2
d 0.7 **e** 0.6 **f** 0.4

5 **a** $\frac{16}{103}$ **b** $\frac{66}{103}$ (inclusive or)
c $\frac{37}{103}$

Enrichment

6 **a** **i** $\frac{71}{592}$ **ii** $\frac{94}{592}$ **iii** $\frac{387}{592}$

b For blond-haired students the probability is $\frac{94}{127}$ = 0.740, for brown-haired students the probability is $\frac{84}{268}$ = 0.313, so blond-haired students are more likely to have blue eyes. In each case, we use only the particular group of students which we are investigating (and the probabilities are called 'conditional'.

Exercise 3B

1 **a** In the table, AB represents Abracadabra first and Basil Boy second. Of course, the empty cells on the diagonal can't occur (a horse can't come in first and second in the same race).

	Abr	Bas	Clo	Def	Euc
Abracadabra		AB	AC	AD	AE
Basil Boy	BA		BC	BD	BE
Close Call	CA	CB		CD	CE
Defiant	DA	DB	DC		DE
Eucalyptus	EA	EB	EC	ED	

b 20, could be equally likely if the gambler is just choosing horses randomly (but not if the gambler knows something about their 'form').

c Unlikely, as some horses will be stronger and others will be weaker.

2 **a** $\frac{9}{60}$ **b** $\frac{13}{60}$
c $\frac{23}{60}$ (inclusive or – they could be studying both)

3 **a** $\frac{60}{95}$

b $\frac{61}{95}$. They may be clever students (children of academics at Stanford?), or it may be that their parents are just rating them as better than average.

c 'Deferred gratification' means waiting until later to get a reward. It seems that of those children who could wait more had above average competence 10 years later ($\frac{25}{35}$ = 71%) compared to those who couldn't wait ($\frac{36}{60}$ = 60%). Ability to defer gratification at age 4 indicates a greater degree of competence at age 14.

4 A total of 12 has probability $\frac{1}{36}$ as it can only be made in one way, 6 on first die and 6 on second die; a total of 11 can be made as 6+5 or 5+6 and so has probability $\frac{2}{36}$. Leibniz was wrong.

5 **a** The outcomes from the two dice are organised so that the same totals are all in the same column, then the number of outcomes in the column shows the probability of that total.

b $\frac{3}{36} = \frac{1}{12}$, total of 4 has the same probability

Enrichment

6 **a** 0.078

b 0.229, 'exclusive or' since a person can't die of two separate causes.

c The three causes listed are the leading causes of death for both males and females, though there are differences in other causes (e.g. males have higher rates of suicide than females – more than 3 times as high). Separating people into age groups will definitely affect the results as these conditions are likely to be suffered by people who are ageing.

Exercise 3C

1 **a** **i** $\frac{2}{30}$ = 0.067 **ii** $\frac{2}{34}$ = 0.059

b In each case we use the 'condition' to determine the reference group, the 30 smokers in part a, and the 34 high exercisers in part b.

2 **a** $\frac{28}{250}$ = 0.112 **b** $\frac{28}{110}$ = 0.255 **c** $\frac{28}{45}$ = 0.622

3 **a** P(7,11,2,3 or 12) = $\frac{12}{36} = \frac{1}{3}$

b P(5 given 4,5,6,8,9, or 10) = $\frac{4}{24} = \frac{1}{6}$

c P(5 given 5 or 7) = $\frac{4}{10} = \frac{2}{5}$

4 **a** $\frac{212}{885}$ = 0.240

b $\frac{203}{325}$ = 0.625, $\frac{118}{285}$ = 0.414, $\frac{178}{706}$ = 0.110

c $\frac{528}{1490}$ = 0.354

d Probability of survival gets lower with 'lower' class (crew between second and third class); sex/age – British idea of 'women and children first' into the lifeboats meant that these groups had higher chance of survival.

ISBN 978-1-107-65599-7

5 **a** $\frac{36}{60} = 0.6$ **b** $\frac{25}{35} = 0.714$

c Being able to wait for your reward at age 4 is associated with higher coping and mental competence at age 14.

d Those children who were able to delay their gratification may have been the ones who were cleverer and more resilient people, which showed by being rated more competent at age 14.

Enrichment

6 **a** 0.64. This seems quite low but actually the number is correct – 36% of newborn babies didn't survive to age 6.

b $\frac{25}{64} = 0.391$

c For a 36 year old the chance is $\frac{1}{16} = 0.0625$; for a 16 year old the chance is $\frac{1}{40} = 0.025$, which is two and a half times as likely ($\frac{0.0625}{0.025} = 2.5$).

d Fewer than 5 newborn babies out of 1000 die before the age of 6, so the probability of survival is greater than 0.995.

Exercise 3D

1 **a** Since the coin and the die have nothing to do with each other, it seems reasonable that an event involving the coin and an event involving the die are independent.

b If the two people are randomly selected then their views will be independent, so an event involving the first person and one involving the second person will be independent. But if you were interviewing (say) both members of a married couple then their views would be likely to be dependent.

2 If the first ball is replaced and the balls are thoroughly mixed before the second selection then any event involving the first selection will be independent of any event involving the second selection. But if the first ball is **not** replaced, the composition of the bag will be different depending on which ball was selected first, so the events will be dependent.

3 **a** $0.15 \times 0.15 = 0.0225$

b This event is the complement of not finding oil at either location, so its probability will be $1 - 0.85 \times 0.85 = 0.2775$.

4 The diagram shows that P(A and B) = 0 since the circles don't overlap; as long as P(A) > 0 and P(B) >0, as the diagram implies, the events cannot be independent. Intuitively, if we know that A occurs then we know that P(B) = 0, and vice versa.

5 **a** P(pet) = 0.6, P(survived) = 0.75

b P(pet and survived) = 0.45 (= 0.6 × 0.75, so the events are independent)

c P(survived given pet) = $\frac{45}{60} = 0.75$, P(survived given no pet) = $\frac{30}{40} = 0.75$

d If the events A and B are independent then the conditional probability of A given B is the same as the (unconditional) probability of A.

Enrichment

6 **a** The two events are complementary.

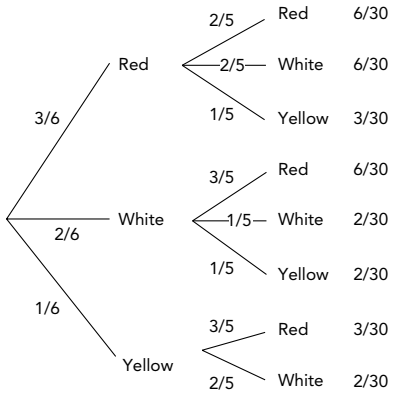
b P(all 6 O-rings hold) = 0.9776 = 0.870, calculate the probability by multiplying together the probabilities of the six independent events (extending the definition of independence to more events).

c This is the complement of the event in part b, so the probability is 1 – 0.870 = 0.130, and a 13% chance of problems is quite high for this situation!

d The probability of failure for any single O-ring may be greater than 0.023 in low temperature, but if the temperature affected all the O-rings then they would no longer be independent.

Exercise 3E

1 **a**



ISBN 978-1-107-65599-7

b Probabilities do sum to 1.

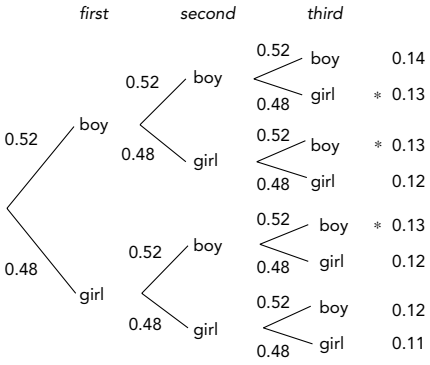
c P(yellow selected) = P(RY, WY, YR, YW) = $\frac{10}{30} = \frac{1}{3}$

d P(both balls same colour) = P(RR, WW) = $\frac{8}{30} = \frac{4}{15}$

e P(both balls different colour) = $1 - \frac{4}{15} = \frac{11}{15}$, the complement of the event in part d

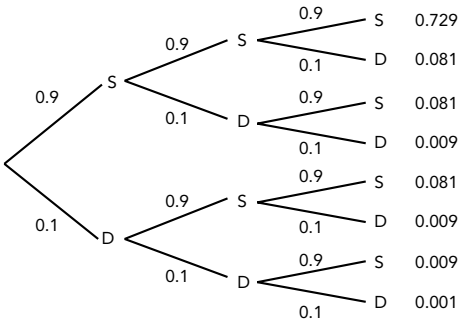
2 For the family of three children, we would say that the selection is with replacement, not because the babies are returned, but because boy and girl are both possible at each stage, and with the same probabilities at each stage.

3



P(exactly two of the three children are boys) = 0.13 + 0.13 + 0.13 = 0.39

4 **a**



b Probabilities are shown on branches of diagram.

c They do sum to 1.

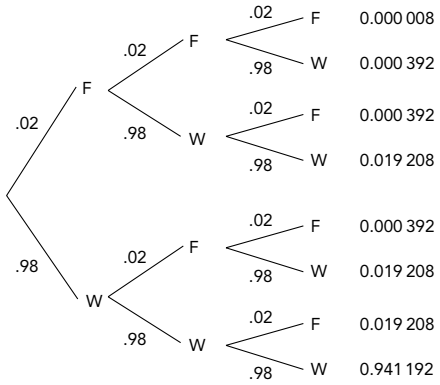
5 **a** P(none die) = P(SSS) = 0.729

b P(exactly one dies) = 3 × 0.081 = 0.243

c P(at least one dies) = 1 – P(none die) = 0.271

Enrichment

6 **a**



b Probabilities are shown on branches of diagram.

c P(3F) = 0.000 008, P(2F) = 0.001 176, P(1F) = 0.057 624, P(0F) = 0.941 192

d The probability of finding 2 faulty components (or more) is very small, around $\frac{1}{10}$ of 1%, so if you did find 2 faulty components you might conclude that the machine was not working at its usual standard, but rather, it was producing more than 2% faulty components.

e The probability of finding at one (or more) faulty component is around 6%, not common, but not too small to be suspicious. However, you might keep a close eye on the machine, and maybe carry out another test sooner rather than later.

Chapter summary

Multiple-choice questions

1 B **2** A **3** B **4** C **5** D

6 C **7** A **8** D **9** B **10** C

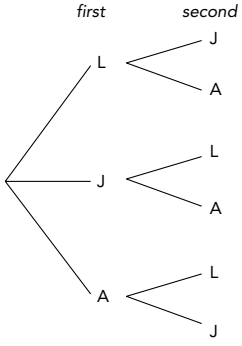
Short-answer questions

1 The experiment consists of tossing three coins, and we can list the sample space as {hhh, hht, hth, thh, tth, tht, htt, ttt} or equivalently show it on a three-stage tree diagram. Assuming that the coins are fair and tossed properly, all 8 outcomes are equally likely. They each buy their own coffees if hhh or ttt come up, which happens with probability 0.25.

ISBN 978-1-107-65599-7

2 **a** $\frac{21}{778} = 0.027$ **b** $\frac{268}{778} = 0.344$

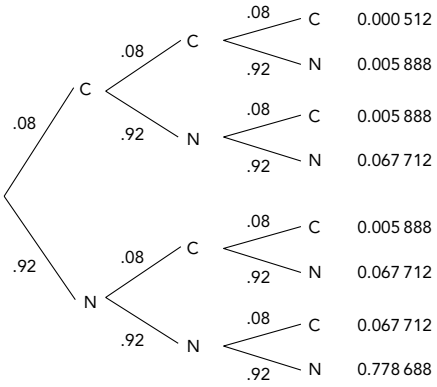
c $\frac{21}{96} = 0.219$ **d** $\frac{21}{193} = 0.109$



3 **a** Tree diagram, sampling without replacement, all branches equally likely.

b $\frac{4}{6} = \frac{2}{3}$ **c** $\frac{2}{6} = \frac{1}{3}$

4 **a** Tree diagram and probabilities (C = colour-blind, N = normal vision)



b P(none colour-blind) = P(NNN) = 0.778 688

c P(majority colour-blind) = P(2 or 3 colour-blind) = $3 \times 0.005\,888 + 0.000\,512 = 0.018\,176$

d P(female colour-blind) = P(she has colour-blindness gene on both X-chromosomes) = 0.08×0.08 (since they are independent) = 0.0064, which is just over one-half of a per cent.

5 **a** P(series system works) = P(all four components work) = 0.975^4 (by independence) = 0.904

b P(parallel system works) = P(at least one component works) = 1 – P(all four components fail) = $1 - 0.025^4$ (by independence) = 1 (actually, 0.999 999 6)

c The parallel system is much more likely to be working than the series system – so when you buy Christmas tree lights, don't buy the cheap ones that have the bulbs connected in series, but find the more expensive ones where they are connected in parallel!

Note that you could show the sample space in a four-stage tree diagram, but since you are only interested in these two events, it is easier to work directly.

Extended-response questions

1 **a** You can add any of the 6 rasa and then add any of the 5 other rasa, which gives 30 possibilities. But now you have counted each pair twice (e.g. salty first and hot second, hot first and salty second) even though it makes no difference which you put in first. So you have to divide the number of possibilities by 2, giving 15 combinations. You could show these in a two-stage tree diagram, or you could just list them as {bit-sou, bit-sal, bit-ast, bit-swe, bit-hot, sou-sal, sou-ast, sou-swe, sou-hot, sal-ast, sal-swe, sal-hot, ast-swe, ast-hot, swe-hot}.

b We can continue using the same approach: 6 possibilities for the first rasa, 5 for the second, and then 4 for the third, but since the order is not important for the dish, we have to divide by the 6 different orders for adding three tastes (for example, bit-sal-hot, bit-hot-sal, sal-bit-hot, sal-hot-bit, hot-sal-bit, hot-bit-sal all result in the same flavour). So the total number of possibilities is $\frac{(6 \times 5 \times 4)}{6} = 20$. Again, we could show these on a three-stage tree diagram or in a list.

c Picking four rasa to add to the curry can be done in exactly the same number of ways as picking two rasa to leave out of the curry!

2 **a** So that the effect of propranolol could be compared to something – the effect of a dummy pill.

b So that some patients didn't receive an extra psychological benefit just from being given a pill.

c

	Died	Survived	Totals
Regular	31	1006	1037
Sporadic	4	53	57
Totals	35	1059	1094

d 0.03, 0.07

e It seems that people who take their pills regularly do better than people who don't take them regularly, and this is true for placebo as well as

ISBN 978-1-107-65599-7

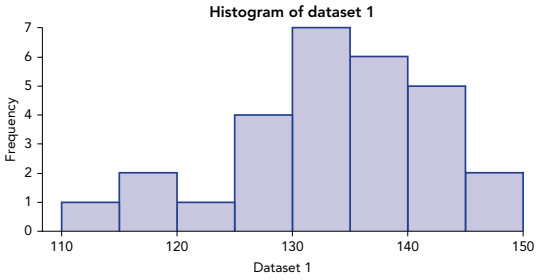
active pills. Maybe the patients who are sickest are less likely to remember to take their pill, or maybe there is some personality connection and patients who are more regular with their pills are also more focused on other aspects of their treatment.

3 There are many possible questions, and you have lots of examples in the chapter!

Chapter 4

Pre-test

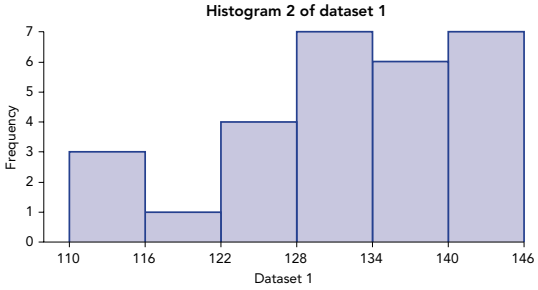
1 a



i Skewed to the left

ii Unimodal. (With only 28 observations, the 'bump' between 115 and 120 must be considered as just due to data variation.)

b



i Skewed to the left

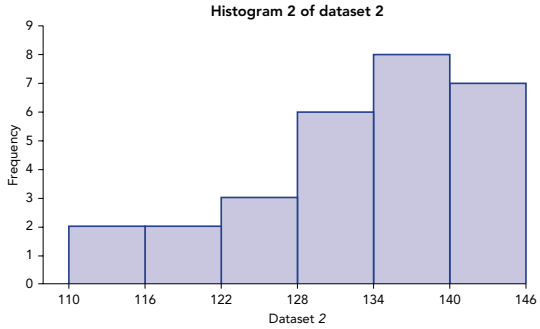
ii It looks as though there might be two groups – one between 110 and 116 and the other between 122 and 146, so it might be considered bimodal.

iii No for part **i**; yes for part **ii**. Using different bins for only 28 observations can give a different picture.

c Mean = 131.07 s, range = 35 s. The median is halfway between 14th and 15th observations, which are both 130.0 s.

d The mean is slightly greater than the median although the data look skewed to the left.

2 b



i Skewed to the left

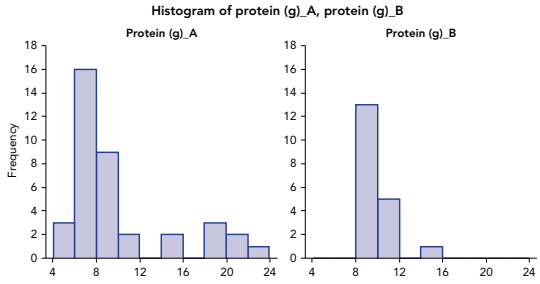
ii Unimodal

c Mean = 132.43 s, range = 35 s. The median is halfway between 14th and 15th observations, which are both 134.00 s.

d The data have a mean of 132.34 and a median of 134, and are skewed to the left. The data range over 35 s, from 110 to 145 s. The data appear to belong to one group – that is, there are no indications of bimodality.

e By rounding down to the nearest 5 s. For example, 116 is changed to 115 and 144 changed to 140. This might happen if the data were available only as a stem-and-leaf plot.

3 a



b

Variable	Mean	Median	Range
Protein (g)_A	10.142	7.95	17.7
Protein (g)_B	10.099	9.60	6.5

c Brand A has a greater range of amounts of protein (per 100 g of cereal) than brand B, but this may just be due to brand A having more types of cereal and so greater variety. The amounts for brand A range from 5.4 to 23.1, while for brand B, from 8.4 to 11.9. For both brands, the most common values of amounts of protein per 100 g are some

ISBN 978-1-107-65599-7

where about 8: 6–10 for brand A and 8–10 for brand B. Although both brands clearly have some types of cereals with emphasis on more protein, the data could not be described as bimodal. The average amounts are about 10 for both, while the median for brand B is greater than for brand A. The data for both brands are skewed to the right, due to the cereals with more protein per 100 g.

Exercise 4A

1 a The median is the 10th observation which is 5.1 m. The lower quartile is the median of the lower group of 9 observations, so is the 5th observation which is 4.7 m. The upper quartile is the median of the upper group of 9 observations so is the 5th observation from the top, which is 5.4 m. The interquartile range is therefore 0.7 m.

b The median of the males is 5.2 m, the interquartile range is 0.85 m and the minimum and maximum are 3.9 m and 6.4 m. The minimum and maximum of the female data are 3.8 m and 7.3 m (using the stem-and-leaf plots). So the overall range of the female guesses is greater than the males but the central half of the observations are slightly less variable. The median guesses are quite close and both are greater than 5 m.

2 a Indie: From question 4b of Exercise 1A, the median length of the Indie songs from the stem-and-leaf plot is 220 s. There are 18 song lengths shorter than this, so the lower quartile for the Indie songs is halfway between the 9th and the 10th observations, which are both 200 s in the stem-and-leaf plot. The upper quartile is halfway between the 9th and the 10th observations from the top, which are 240 and 250 s in the plot. So the upper quartile of the Indie song lengths is 245 s.

Alternative rock: There are 23 observations, so the median is the 12th observation which is 220 s. There are 11 observations in the lower group so the lower quartile is the 6th observation which is 190 s. The upper quartile is the 6th observation from the top, which is 290 s.

b Indie: Using the original data, the median and quartiles are 226 s, 201.5 s, 249 s.

Alternative rock: Using the original data, the median and quartiles are 220 s, 194 s and 298 s.

So although the difference can be up to 10 s, only two of these values have a difference of more than 5 s between the calculation from the original and that of the stem-and-leaf plot. Most of the values using the stem-and-leaf plot are lower than those obtained from the original data because of rounding down, so comparisons are less affected.

c The median lengths are the same (or similar), but the Alternative rock lengths have a smaller lower quartile and a greater upper quartile so the spread of the central 50% of values is considerably greater for the Alternative rock than for the Indie. The Alternative rock lengths range from 152 s to 499 s, while the Indie lengths range less – from 130 to 332 s.

3 a Using the stem-and-leaf plots, the median and quartiles of the differences for the males are: 0 s, –2 s, 2.5 s. For the females, the median and quartiles of the differences are 0, –1 s, 2 s.

b The differences for the males range from –5 to 8 s, while the differences for the females range from –6 s to 10 s. So the medians are the same for males and females (both 0). The central 50% are less spread out for the females than the males, but the overall range is greater for the females than the males.

4 a For samples of size 8, $n = 8$. So the third quartile is the $\frac{27}{4} = 6\frac{3}{4}$th from the bottom. That is, it is $\frac{3}{4}$ of the way between the 6th and the 7th. This is $\frac{1}{4}$ of the way down from the 7th observation which is $2\frac{1}{4}$ from the top – that is $\frac{(n+1)}{4}$th from the top.

For $n = 9$, the third quartile is the $\frac{30}{4} = 7\frac{1}{2}$th from the bottom. This is halfway between the 7th and 8th which is $2\frac{1}{2}$th from the top – that is, $\frac{(9+1)}{4}$th from the top.

For $n = 10$, the third quartile is the $\frac{33}{4} = 8\frac{1}{4}$th from the bottom. This is $\frac{1}{4}$ of the way from the 8th to the 9th which is $2\frac{3}{4}$th from the top – that is, $\frac{(10+1)}{4}$th from the top.

For $n = 11$, the third quartile is the $\frac{36}{4} = 9$th from the bottom. This is 3rd from the top – that is, $\frac{(11+1)}{4}$th from the top.

b There are 64 observations so the first quartile calculated this way is the $\frac{(64+1)}{4}$th $= 16\frac{1}{4}$th from the bottom. This is $\frac{1}{4}$ of the way from the 16th to the 17th observations, which are stated in Example 1 as being 15 s and 15 s. So the first quartile calculated this way is 14.25 s. The third quartile is the $\frac{(64+1)}{4}$th $= 16\frac{1}{4}$th from the top, so $\frac{1}{4}$ of the way from the 16th to the 17th from the top. But as stated in Example 1, these are both 60 s, so the third quartile calculated this way is still 60 s.

ISBN 978-1-107-65599-7

Enrichment

5 a There are 29 observations so there are 14 observations in each half. So the lower quartile is halfway between the 7th and 8th observations which are 5.29 and 5.3, so the lower quartile is 5.295. The upper quartile is halfway between the 7th and 8th observations from the top (so the 22nd and 23rd from the bottom), which are 5.61 and 5.62, so the upper quartile is 5.615.

b The first quartile is the $\frac{(29+1)}{4}$th = $7\frac{1}{2}$th observation from the bottom, so is the same as in part a. The third quartile is the $\frac{(29+1)}{4}$th observation from the top so is the same as in part a.

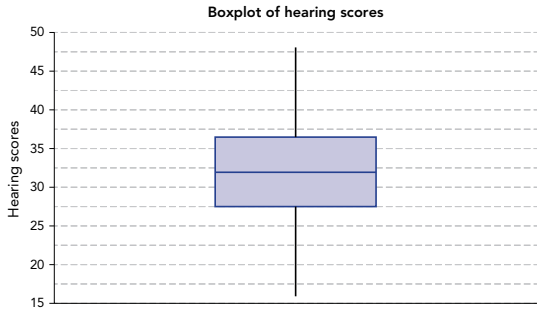
c No difference because the sample size + 1 is 30 so $\frac{30}{4}$ = 7.5 which is same as taking median of lower group. (Note that in general, if the sample size, n, is an odd number then $\frac{(n+1)}{4}$th gives the same as taking the median of the lower group). For a sample of size 20, $\frac{(n+1)}{4}$th = $\frac{21}{4}$ = $5\frac{1}{4}$. The lower group below the median has 10 observations in it so the median of this group is halfway between the 5th and the 6th observations. So to get a big difference between the two ways of calculating the quartiles, we need a big gap between the 5th and the 6th observations from the bottom. For example, the data ordered from the smallest could start as 1, 2, 3, 4, 5, 11, … The $5\frac{1}{4}$ method would give 6.5, while the usual method would give 8.

d i Without the smallest observation (4.07), the mean is 5.4679. With 28 observations, the median is halfway between the 14th and 15th observations, which are 5.46 and 5.47. So the median is 5.465.

ii The lower quartile is the median of the lower 14 observations so is halfway between the 7th and 8th, which are 5.30 and 5.34. So the lower quartile is 5.32. The upper quartile is halfway between the 7th and 8th observations from the top, which are 5.61 and 5.62. So the upper quartile is 5.615.

Exercise 4B

1 a



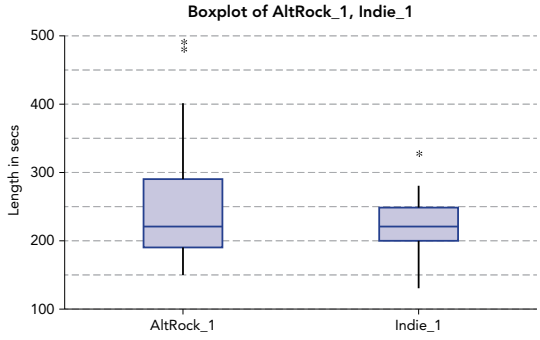
b From the boxplot, the median score is approximately 32, and the quartiles are 27.5 and approximately 36, so the central 50% of scores lie between 27.5 and 36. The minimum and maximum scores are, respectively, 16 and 48. The data appear to be symmetric.

c The histogram looks a little skewed to the left. The interval with the most observations in that histogram is the interval 32–36.

2 a The data in the stem-and-leaf plots is the original data rounded down by dropping the digit in the units place. These data are called AltRock_1 and Indie_1 in the boxplot below. Note that the five-number summaries of question 2a of Exercise 4A are:

AltRock: 150, 190, 220, 290, 490

Indie: 130, 200, 220, 245, 330

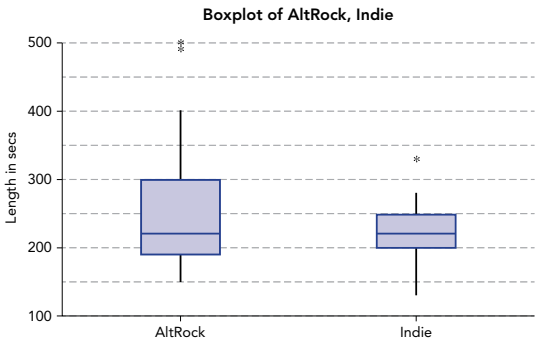


For the original data, the five-number summaries are found in question 2b of Exercise 4A and are:

AltRock: 152, 194, 220, 298, 499

Indie: 130, 201.5, 226, 249, 332

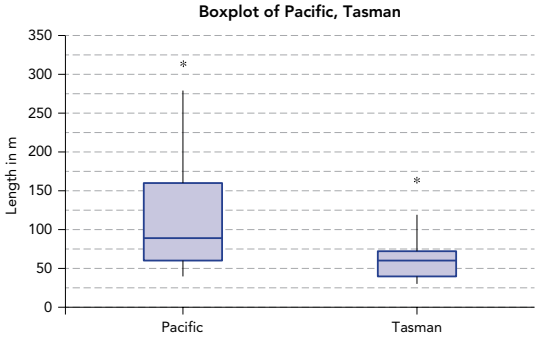
ISBN 978-1-107-65599-7



b Note that there is very little difference in the appearances of the boxplots between using the original data and the data with the third digit (in the unit place) dropped. The median lengths are similar, but the spread of the AltRock lengths is greater both in the central 50% of data and overall (the range is greater). The AltRock lengths are skewed to the right. The Indie lengths are closer to symmetric.

c The stem-and-leaf plots indicate that the 'middle' length is similar and that the AltRock lengths are skewed to the right. However, it looks as though the 'central' Indie lengths are more spread than the 'central' AltRock lengths – this is because there are more Indie songs so it's not easy to judge this from the stem-and-leaf plot.

3 a The stem-and-leaf plots for these data have leaf unit = 10 m, so the data in these plots are rounded down from the original data by dropping the units digit. The boxplots of these data are shown below.

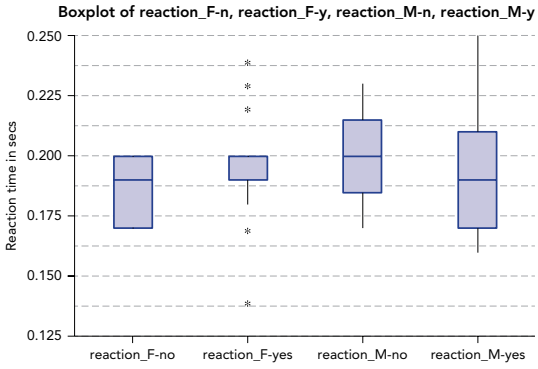


b The rivers flowing into the Pacific are generally longer and have greater variation in their lengths than those flowing into the Tasman; the median, interquartile range and overall range are all greater for the Pacific rivers than the Tasman. The lengths of the rivers flowing into the Pacific are skewed to the right, while those flowing into the Tasman are asymmetric.

c The 'story' given by the boxplots is very similar to those given by the stem-and-leaf and dotplots.

Enrichment

4 a



b The median and the lower quartile are both = 0.19.

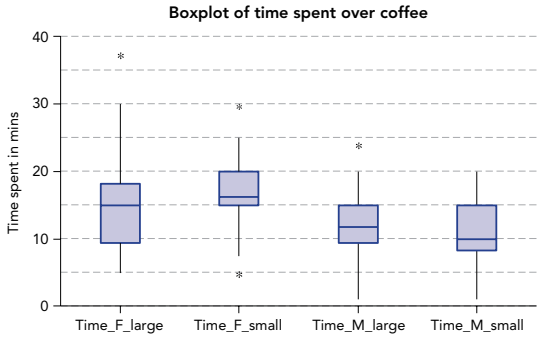
c The medians are about equal. The spread of the central 50% of data is greater for the females who did not have coffee, but the overall range for the females who did is greater. Because the data are rounded to 2 decimal places, there are a lot of repeated values. Also the sample sizes are fairly small – 11 and 17 – so with the repeated values, there are few different values. This is why the boxplots look strange.

d The median reaction time for the males who did not have coffee is less than for those who did. The spread is about the same for both groups, but the overall range for those who did have coffee is greater. The data for those who did have coffee is skewed to the right, while the data for those who did not have coffee is close to symmetric.

e The dotplots and histograms are of the original data – not rounded – so give more detail. The 'story' of the data is reasonably similar to the boxplots. The dotplots do not tend to give as clear a picture of the data, and the histograms, although better than the dotplots, are still a little 'bumpy' in representing the data.

ISBN 978-1-107-65599-7

5 a



b The times for females for large coffees have a median only just less than for small coffees, but are much more variable overall and in the central 50% of the data. Both of these datasets are asymmetric but it is not clear whether they can be said to be skewed to the right or left. Generally, males spend less time than females, with the median time for small coffees for males less than for large coffees. The variation for males is similar for large and small coffees, and similar to that for females with small coffees. The times for males are closer to symmetric than for females.

c The accuracy is to half a minute. Either it was difficult to observe people or they were asked and had to estimate their times.

Exercise 4C

1 a The median lengths for the intermediate and advanced are similar, but the lengths for the intermediate are more variable. The lengths for the beginners are much less than for the intermediate and advanced groups. The lengths for the beginners are skewed to the left, but for the intermediate and advanced groups, have a long left whisker and no right whisker, but the spread from the median to the upper quartile/maximum is greater than from the median to the lower quartile.

b The average length is less than the median for the beginners, about the same for the intermediate, and greater than the median for the advanced group.

c For the intermediate and the advanced groups, there is a group of lengths between 90 and 100 m – most likely boundaries (or near boundaries). So the upper quartiles for these two datasets are the same as their maximum – 100 m. This is why the averages are greater than the medians despite the long left whiskers – because there are so many observations at or near the maximum value.

d Because an angle of 0 or close to 0 is similar to an angle of 360 or close to it, and the angles tend to cluster close to 0 or 360. The data tend to be bimodal and boxplots are not good at representing bimodal data.

2 a The scores for lists 3 and 4 are very similar with approximately the same mean and median scores and the same quartiles. The only difference is very small – the maximum scores for list 4 are slightly greater than for list 3. These two lists tend to have lower scores than for list 2 but list 2 scores vary the most. List 1 scores tended to be the highest and also the least variable, although the range (= maximum – minimum) is about the same as the other lists.

b It would be of interest to investigate the relationships between people's scores over the four lists. For example, did the high scorers for list 1 also score highly for the other lists?

3 a The prices of all types of properties tend to be skewed to the right, so the average prices tend to be greater than the median prices, although, surprisingly not for the units, despite a group of higher unit prices. As would be expected, houses tend to be more expensive, although some of the apartments are as expensive as houses. The order of the median prices is houses, apartments, townhouses, units. The apartments have the most variation in prices, both the central 50% and the overall range, while the townhouses have the least variation, apart from one expensive one. The variation of the majority of the houses and the units is similar.

b Because they are quite skewed to the right.

c Yes, because of the longer whisker to the higher prices and the cluster of prices at the end of this whisker. However, if you look closely, you can see that the lower half of the box – the distance from the lower quartile to the median – is longer than the upper half and this is balancing out the longer right whisker. So the data are not clearly skewed to the right or the left.

4 Both airlines had approximately the same median difference, but airline B's differences were more variable, both the central 50% of differences and overall. For both airlines, about 75% of flights arrived after their scheduled time, but 25% arrived after but less than about 7 or 8 minutes after schedule. And for both airlines, about 25% arrived early. For airline A, 75% arrived by 10 min after the scheduled arrival time, but for airline B, only 75% arrived by 20 min

ISBN 978-1-107-65599-7

after the scheduled arrival time. Only a few arrivals for airline A were more than 20 min late, and none were more than 40 min. For airline B, 25% were more than 20 min late and up to 40 min late, and two were more than an hour late.

Enrichment

5 a Going down the stairs, the median time for females was slightly longer than for males, but the times for females were more variable. Both of these datasets are skewed to the left, so there was more variation amongst the faster students, and some were very fast – both males and females. There was one female who went down much more slowly than the others. Going up the stairs, the median times were about the same for males and females – about 26 s which was about a second slower than the median for going up – but there was a bit more variation in the males' times. There was generally more variation in the times for going up than for coming down. There was a slight tendency for males to be slower than females going up, but the overall range was about the same.

b Going down, the median time was smaller for peak 2 than for peak 1 or off-peak which had almost equal median times. The variation was greatest for off-peak, then for peak 2, while peak 1 had the least variation for most students except for a number who went down very fast. There were also some in peak 2 who went down very fast. For going up, the variation was generally greater than for going down no matter what time of day. 50% of those in peak 2 were faster than most going down in any time period, and 50% of those in off-peak were faster than many going down, but the slower 50% for both these time periods had more variation. Peak 1 going up had the greatest median and also the greatest variation.

c For off-peak and peak 2, there were a lot of students who went up the stairs at about the same speed, and then the other half were slower and more variable. If there were a lot of students going up at the same time, there would tend to be a crowd and a lower limit to the speed, with a lot of people doing this speed.

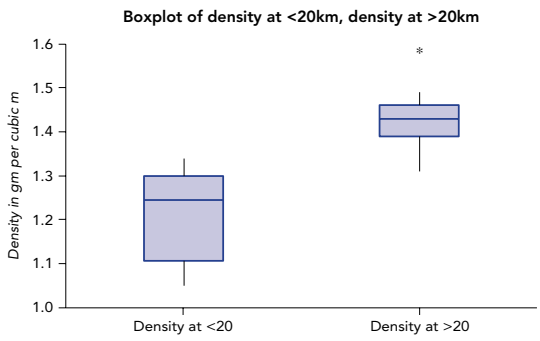
Chapter summary

Multiple-choice questions

1 B **2** D **3** C **4** D **5** C **6** D

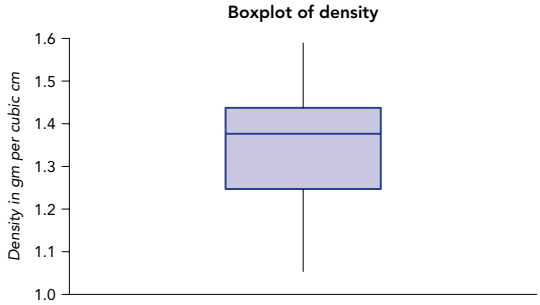
Short-answer questions

1 a



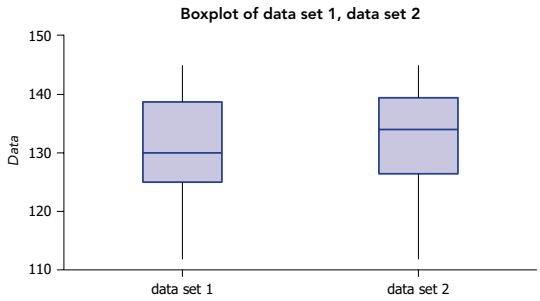
b The densities at >20 km are much greater and much less variable than the densities at <20 km. The median density at <20 km is about 1.25 g/cm^3 and at >20 km is about 1.43 g/cm^3. The densities at <20 km are skewed to the left – that is, the smaller densities are more variable – while the densities at >20 km are close to symmetric, but with one very large density.

c Can't see that there are two groups of densities – <20 km and >20 km – and that there are differences between them.



d Personal preference. But problem with all is that there are so few observations.

2 a

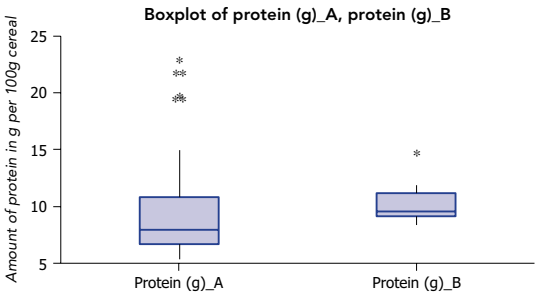


ISBN 978-1-107-65599-7

b The median of dataset 2 is greater than the median of dataset 1. Dataset 2 is skewed to the left; dataset 1 is asymmetric but can't be said to be skewed to the right or left.

c The asymmetry

3 a



b Brand B tends to have more protein with much less variation than brand A. Both brands have 75% of cereals with amounts of protein less than about 11 g per 100 g, but all of brand B's are at least about 8 g, whereas 50% of A's have less than about 7.5 g per 100 g. However, brand A has a number of cereals with very high protein – greater than 20 g per 100 g. In summary, brand B has consistently more protein than brand A, which has greater diversity, including some high protein cereals.

c The values of amounts of protein that different percentages of the cereals have more or less than. Also the contrasts between the brands are emphasised.

4 The speeds for peak 2 were faster and less variable than for the other time periods, although both males and females had groups that were slower. The medians for peak 1 were just a little greater than for peak 2, but there was much more variation in the times. There wasn't much difference between males and females in the two peak periods. Off-peak was the slowest but there was generally much more variation in the male speeds than the female speeds, except for a few females.

5 a What stands out is that there is very little variation in speeds over most people in all time periods, but in each time period there are some very fast speeds, and a few slow ones in off-peak and peak 2, especially females off-peak. In peak 1, the males and females had about the same speeds, while in peak 2 and off-peak, the females tended to be a bit slower. Peak 2 tended to be a bit faster than peak 1.

b The variation in speeds going up in peak 1 is greater than for the other two periods, while the variation in speeds going down is similar across the three time periods. The median times going down are all more than 25 s, but the medians for going up are less than 25 s for the two peak periods. No times are less than 20 s going up but there are some that are less than 20 s going down.

Extended-response question

6 a The golf balls bounced most, followed by the table tennis, bounce, rubber and then the tennis balls which bounced least. The table tennis balls had the most variation in the number of bounces, then the tennis balls, and then the rubber, while the variation for the golf and bounce balls was about the same and quite small. The data for the table tennis balls are skewed to the left (that is, there was more variation in the number of bounces for those with fewer bounces) but the data for the others are not far from symmetric.

b The golf balls were still the bounciest, but now the bounce and table tennis balls were about the same, and the tennis balls were bouncier than the rubber. The table tennis and the rubber balls had the most variation, followed by the golf balls, while the tennis and the bounce balls had the least. The data for the golf balls are skewed to the left, for the bounce skewed to the right, while the others are close to symmetric.

c The golf balls still had the greatest number of bounces, but are becoming more variable. The tennis balls were next, then the bounce, table tennis and rubber. The table tennis balls still had the most variation, followed by the golf balls, but the variation in the data for the tennis, rubber and bounce balls was very small. The data for the table tennis balls are very skewed to the left – the variation in the smaller number of bounces was large – but for the others, there is little skewness (or it is difficult to tell).

d The most striking aspect is that the golf balls stay as the bounciest, while the tennis balls move from least to second most bouncy, and the table tennis balls move from second most bouncy to second least bouncy. The table tennis balls stay as having the most variation. The variation for the tennis balls decreases, and increases for the golf balls. The variation for the rubber balls increases then decreases, and decreases for the bounce balls from the height of 1 m to 1.5 m.

ISBN 978-1-107-65599-7

Chapter 5

Pre-test

1 a Four years

b We need to know how the values in each table correspond to each other – that is, which weeks and years each of the values are for.

c Give the week and year with the data for each town underneath (or beside) each week and year.

2 a The data are linked or paired – for example, that 14 and 27 somehow belong together.

b What were the wire specimens? For example, were there 20 different wire specimens? Or 10 and each split into two, with one given to one lab and the second given to the other lab? Are the above data pairs of observations in some other way – for example, each lab carried out tests on the same days?

c The data are arranged in increasing order for both labs. This indicates that the data are probably not pairs of observations as it would be unlikely that this ordering would happen naturally with pairs of observations, especially as there are repeated values for lab 2.

3 Use the same service stations and collect data at the same time at each service station.

4 Different units, and different accuracies

5 Need careful description of the types of crimes so that the classification can be consistent from year to year. And need the size of the population in the region so the crime rate (number of crimes per head or per 1000 residents) can be investigated.

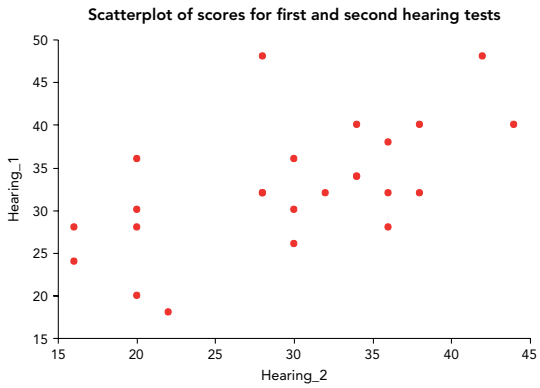
6 a No, the numbers are instead of names and have no numerical meaning.

b No, the 'distance' between strongly agree and agree is not the same as between agree and neutral.

c People's opinions on the question in part b are almost certainly dependent on their answers to the question in part a.

Exercise 5A

1 a



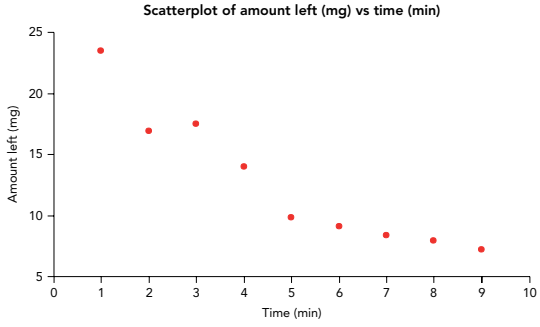
b No, unless test 2 was done after test 1 and it was of interest to see if people's scores went up or down.

c Yes. Although there is a lot of variation, those who scored reasonably on one test also tended to do so on the other.

d People's scores are linked but there is a lot of variation from one test to the other. Might say they are somewhat consistent or consistent to some extent.

e Amongst those who tend to score less – particularly those who score between 20 and 30.

2 a



b Because we want to see what is happening with the amount left when the experiment is run for different amounts of time.

c The amount left decreases, but decreases less and less as the time is increased.

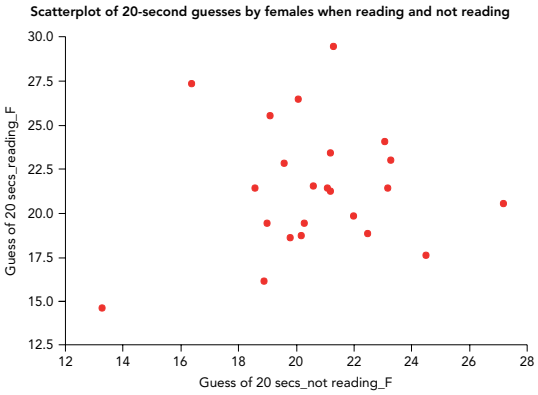
d The second one when the experiment was run for two minutes looks a bit strange and might need checking. However, this point is not far away from the trend, allowing for variation. Only more statistical analysis done at more advanced levels will

ISBN 978-1-107-65599-7

be able to tell us if there are any unusual points. No changes should be made to the data based on this plot.

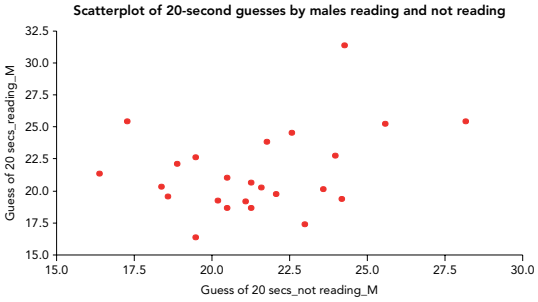
e The amount left might decrease slightly but not by much.

3 a



b No. There's a lot of variation and the plot doesn't suggest the guesses are related.

c



d Not really. Except for one person, the guesses when reading vary between about 16 and 25 no matter what the guess was when not reading.

e The two guesses most out by females when not reading were about 13.5 and 27.5 s. So the one most out was 27.5 s. The guess most out for a male when not reading was about 28.5 s. (Note that the exact values can be found from the original data as 27.2 s and 28.2 s).

f The scales are different because the extreme guesses are different, and the plots give the impression that the male guesses are more variable when not reading than when reading. But a closer look at the scales shows that we can't really use these plots to comment on how the males and females compare.

4 a It looks like people's ability to hold their breath increases quickly as they grow from childhood to adulthood, and then tends to gradually decrease, but with a lot of variability.

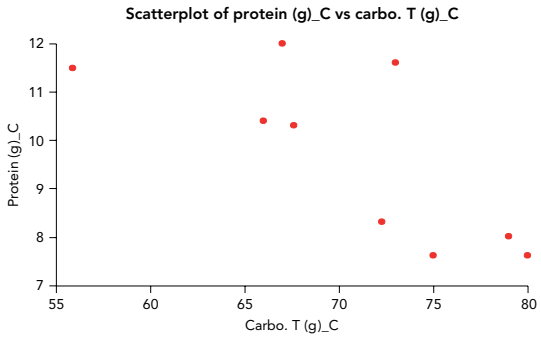
b Probably 40, but this is due to two large values. Possibly 20, but there are more people at 20, and the more people we have the more variation we are likely to get.

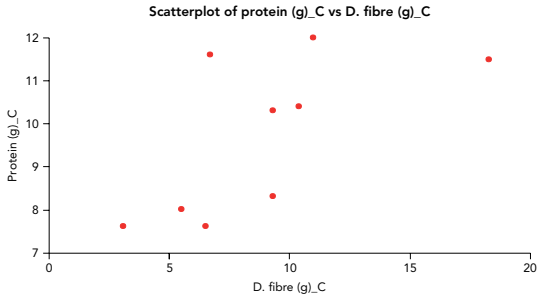
c They have very large values for how long they can hold their breath.

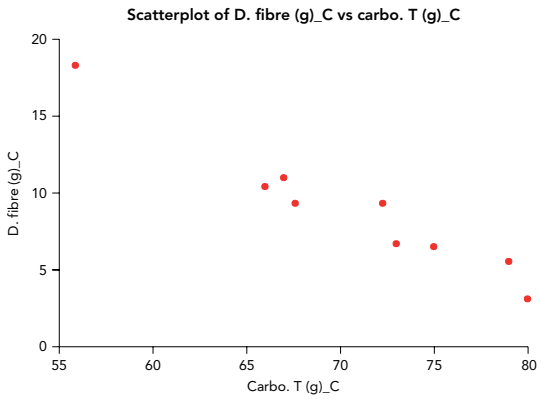
d Perhaps how much exercise they do, their general state of health, and for the adults, whether they smoke or not.

Enrichment

5 a







ISBN 978-1-107-65599-7

b The amount of protein tends to decrease as the amount of carbohydrate increases and tends to increase as the amount of dietary fibre increases. The amount of dietary fibre tends to decrease as the amount of carbohydrate increases.

c The amount of dietary fibre and the amount of carbohydrate.

d Difficult to say between the first two – the amount of protein seems to vary a lot with the amount of carbohydrate and with the amount of dietary fibre.

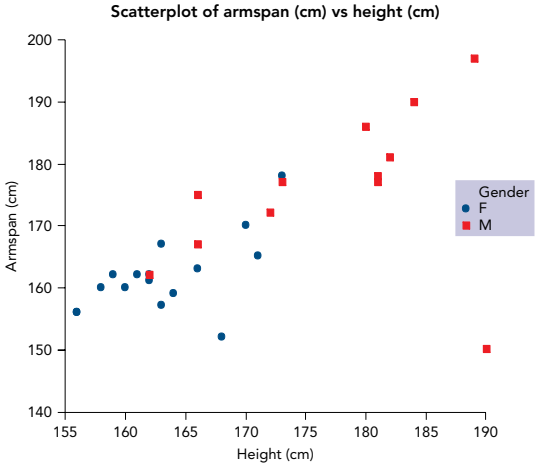
e No, because we're looking at relationships between the variables.

f There is a group of four cereals with high protein – probably featured as high protein. Apart from these, the amount of protein doesn't depend much on the amount of carbohydrate for many of the cereals until we get to those cereals with the highest amounts of carbohydrate and then the amount of protein decreases.

Exercise 5B

1 a An armspan of 10 cm is clearly wrong. An armspan of 100 cm for a height of 157 cm is almost certainly wrong too (arms are far too short for height). As it's impossible to know how to correct these, they will be omitted.

b



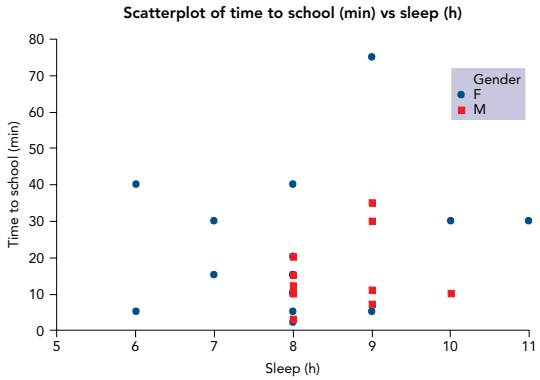
c Armspan, as we usually think in terms of how other body measurements depend on height – probably because height is one of the easiest measurements to take. Also many medical charts are given in terms of dependence on height.

d The observation with armspan = 150 cm and height = 190 cm. You might have already spotted this one and omitted it from your graph.

e Clearly armspan increases as height increases and the relationship looks like a straight line but with variation around it. The females in this group tended to be shorter and have smaller armspans than the males. The females also had quite a lot of variation and the relationship is perhaps not as strong as for the males.

2 a 16 hours sleep seems wrong. And yes, this response is the one who said 10 cm for armspan, so this supports a decision to omit this observation. 75 min to school is long but possible, and the other parts of this response seem reasonable values for those other variables.

b



c There is not a clear decision in this case; it will depend on individual preference.

d 9 hours sleep and 75 min to get to school doesn't leave much time for anything outside school, but this is still possible. So no other points can really be described as wrong. 11 hours sleep is a lot – not probable but still possible.

e There doesn't seem to be any relationship between hours of sleep and time to school. The females' sleep hours are more variable than the males'. Difficult because sleep is recorded to the nearest hour and because there are not many observations.

3 a The variation in shoulder width increases, and increases a lot, as the foot length increases.

b Although there is a lot of variation, the shoulder widths for males tend to increase with foot length for foot lengths up to 28 cm. But the shoulder widths for females do not tend to increase – just become more variable.

ISBN 978-1-107-65599-7

c There's a male observation with a large foot length but a very small shoulder width compared with everyone else – so low that it seems to be wrong.

4 a Yes because it looks like the relationship is fairly straight, so the ratio could be fairly constant.

b It looks like the waist measurement increases faster with hips for females than it does for males.

c There is variation in waist measurement across all the hip measurements for females, but there is more variation for the males for hip measurements around 90 cm than for other values of hip measurements.

5 a For many of the months, the plot is suggesting that for more callbacks there are fewer phone complaints. However, there are months in which the numbers of phone complaints are much the same but there are varying numbers of callbacks. There is one month with a very large number of both.

b There doesn't seem to be any particular differences between months with and without school holidays.

Enrichment

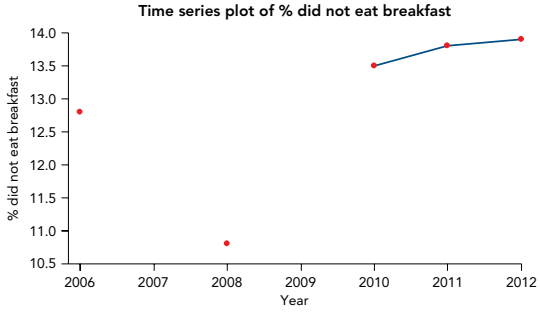
6 a They must be summary statistics for the country, so are most likely average life expectancy and average income per person. Note that the income data are inflation-adjusted.

b The shape in 1950 is slightly S-shaped – the increase in (average) life expectancy as income (average) increases is slow at first but then is faster. In 2007, the increase apart from Africa is fairly steady. The African data are very variable and for many African countries, there is a different relationship to the rest.

c Asian countries have moved up significantly with respect to both average income and average life expectancy. Indian subcontinent countries have also moved up but not as much as Asian. There is less variation amongst the Americas countries – probably a number of them, including South American countries, have increased both average incomes and life expectancy. The north African countries have increased both, but mostly the life expectancy. However the sub-Saharan African countries have become highly variable, with some increasing life expectancy much more than others, and some increasing average income without much increase in average life expectancy. (Note that this is from 1950 to 2007 and does not into account any developments, including wars, since 2007.)

Exercise 5C

1 a



b Apart from 2008, there appears to be a slight trend to increasing percentages of students who do not eat breakfast. However, we must remember that CensusAtSchool data are collected from whole classes within schools that choose to participate, so these are not randomly sampled schools or students.

2 a Usually in or just after January, although there are years with significant rain in other months.

b During 1987–1996, the highest monthly rainfalls were larger (over 300 mm) but over the other months, the rainfall amounts were similar to 1978–1986. However, there were greater contrasts in monthly rainfall during 1987–1996 – during 1978–1986, there were more months of moderate amounts of rainfall.

c Probably 1987, 1993

d There was a very wet summer in 1992, followed by a very dry summer in 1993, and summers with moderate but increasing rainfall to 1996, in which year, an extraordinary amount of rain fell in May.

3 The winning times are tending to decrease over the years, and the variation is also tending to decrease, despite unusual years in 1993 (slower time) and 1999 (faster time).

Enrichment

4 a The numbers of phone calls tended to decrease over time up to about 2004 where they varied between about 100 and 200 a month until 2006, except for very high numbers in April 2004, and even

ISBN 978-1-107-65599-7

more so in March 2005. The numbers of callbacks slowly increased over time, again having peaks in those two months. The number of complaints by internet didn't change greatly, apart from much greater numbers in early 2004 and mid-2005. The numbers of email complaints changed little until 2004 when they rapidly increased and stayed higher, decreasing slightly towards the end of 2005.

b Email in particular, callback a little but with less increase. All had peaks during 2004 and 2005.

c The numbers of phone complaints were always more variable than the others, but all became more variable during 2004–06.

d April 2004 and March 2005 (phone, callback), January, March 2004 and August 2005 (internet), and January, May 2005 (email).

e During mid-2004 to mid-2005, as email was up, the internet numbers were down. Over the years as phone numbers went down, callbacks gradually increased but then became more similar.

Chapter summary

Multiple-choice questions

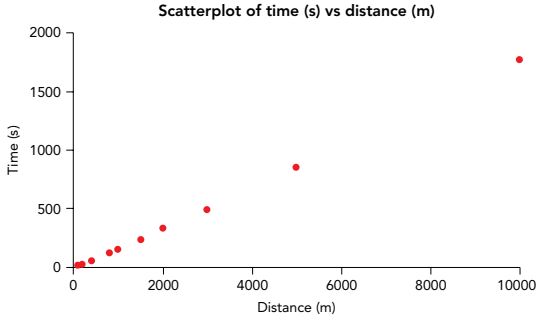
1 D **2** D **3** C **4** D **5** B **6** D

Short-answer questions

1 a Reaction times with the right hand because they are right-handed and we will tend to want to know how their left-handed reactions vary with their right-handed ones.

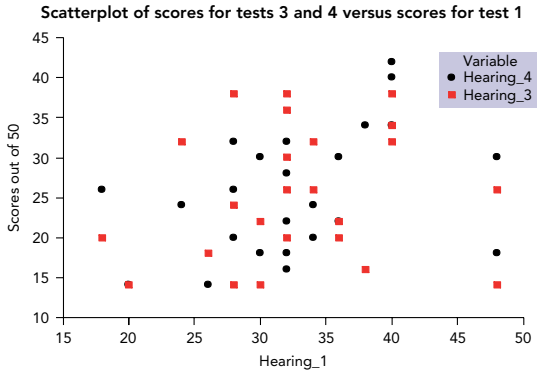
b By using four different symbols for the four different groups formed by the combination of gender and age group.

2 a



b Because that would mean that the speeds were constant.

3 a

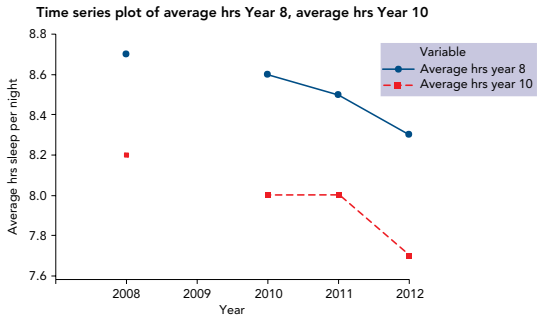


b The scores do not seem to be linked – there's no pattern of tending to increase or decrease.

c The scores on tests 1 and 2 seem to be more related to each other. The scores on test 2 are also closer to those on test 1, while the scores on tests 3 and 4 seem to be a little lower in general.

d The four points have scores of approximately 13 and 26 on test 3, and 17 and 30 on test 4. Note that it could be two people or four people. From the graph in question 1a of Exercise 5A, we see that they are two people who scored 27 and 42 for test 2. There are four people who scored very well on test 1.

4 a



b Yes, there seems to be a decrease.

c Skewed to the left as their means are less than their medians.

d Could use boxplots over school levels and years. Dotplots on the same scale would also be a possibility.

5 a Because the groups overlap a lot and it would be difficult to see any patterns.

b Yes, bigger/smaller guesses on one tend to have bigger/smaller guesses on the other.

ISBN 978-1-107-65599-7

c There is a lot of variation: from 20 cm to about 105 cm for the 50 cm guess, and from about 40 cm to 170 cm for the 100 cm guess. There's a fair amount of variation in people's 1 m guess for the same guess at 0.5 m, but much less than the overall variation for each guess.

d Not much. The females tended to underestimate 0.5 m a little more, and possibly there's a little more variation between their two guesses.

e Not much difference overall, and both have some extreme guesses.

6 a Apart from the north African countries, there is some suggestion that results are better for countries with higher average incomes. The relationship is not strong.

b There are five Asian countries with much higher average results, and there are north African countries with high average incomes but poor results.

c The test is on knowledge and procedures. The plot may indicate that some countries place more emphasis on this type of test than others do.

Extended-response question

7 a Bus arrivals at a selected bus stop on the route.

b Very variable, and quite jagged. Larger or smaller values are often followed by a jump in the opposite direction, whereas moderate deviations from 10 are often followed by moderate deviations.

c The holiday plot is more jagged but more regular. The term time is more jagged in the middle of the day but a bit more consistent at the beginning and the end.

d Larger or smaller values are often followed by a jump in the other direction. This happens more in the holiday plot.

e In the plot for the holidays larger times between buses tend to have smaller times between the next two buses. This supports what we could see in part d but there seems to be no relationship in the plot for term time.

ISBN 978-1-107-65599-7